Accident Analysis and Prevention 154 (2021) 106085

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Trajectory data based freeway high-risk events prediction and its influencing factors analyses

Rongjie Yu [a], Lei Han [a], Hui Zhang [b,*]

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 4800 Cao'an Road, 201804, Shanghai, China
[b] Intelligent Transportation Systems Research Center, Wuhan University of Technology, No.1178, Heping Road, Wuchang District, 430063, Wuhan, China

ARTICLE INFO

Keywords:
Freeway operation and management
Vehicle trajectory data
High-crash risk
Random-effects logistic regression model
Traffic safety pre-warning

ABSTRACT

The frequent crash occurrences have caused massive loss of lives and properties all over the world. In order to improve traffic safety, it is vital to understand the relationships between traffic operation conditions and crash risk, and further implement safety countermeasures. Emerging studies have conducted the crash risk analyses using discrete and aggregated traffic data (e.g., loop detector data, probe vehicle data), where crash events were selected as the prediction target. However, traditional traffic sensing data obtained at segment level cannot describe the detailed operation conditions for the vehicle platoons near crash locations. Thus, more microscopic and high-resolution traffic sensing data are needed. In addition, considering the random occurrence feature of crashes, high-risk events should be paid more attentions given their higher occurrence probability and consistent causations with crashes, which could proactively reduce crash likelihood. In this study, HighD Dataset from German highways was utilized for the empirical analyses. First, high-risk events were obtained using safety surrogate measures with Modified Time to Collision (MTTC) less than 2 s. Traffic operation characteristics within 5 s prior to event occurrence were extracted based on vehicle trajectory data. Then, a total of three different logistic regression models were established, which are standard logistic regression model, random-effects logistic regression (RELR) model, and random-parameter logistic regression (RPLR) model. Among which, the RPLR model was showed to have the best fitness and prediction accuracy. The results showed that the disturbed traffic flows in both longitudinal and lateral directions have positive impacts on high-risk events occurrence. Besides, too close following distance between vehicles would lead to high-risk events. Moreover, RPLR models could provide a high prediction accuracy of 97 % for 2 s ahead of the high-risk events. Finally, potential safety improvement countermeasures and future application scenarios were also discussed.

1. Introduction

Safety is the key to the transportation system worldwide. According to the World Health Organization (WHO), the number of fatalities caused by traffic crash remains unacceptably high, with an estimated number of 1.35 million deaths per year (World Health Organization, 2018). In order to improve traffic safety, proactively identify crash risk and further implement safety countermeasures are the critical steps, which have been widely adopted by National Highway Traffic Safety Administration (NHTSA, 2016) and Federal Highway Administration (FHWA, 2020). With the recent developments of traffic advanced sensing techniques, real-time traffic operational condition data have become more conveniently obtainable. And tremendous efforts have been investigated to develop real-time crash risk analyses (Oh et al., 2001; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015). The real-time crash risk analyses try to establish the relationships between crash probability and pre-crash traffic operational conditions (Hossain and Muromachi, 2012), which could be used to provide proactive warnings to trigger Active Traffic Management System (ATMS) and further reduce the crash potentials (Ahmed and Abdel-Aty, 2012). Besides, the real-time crash risk estimations also hold potentials for the connected and autonomous vehicle (CAV) application scenarios, but higher temporal-spatial resolution and precision of crash risk analyses are required (Katrakazas et al., 2015).

Within the existing crash risk analysis studies, researchers mainly employ spatially discrete and temporal aggregated traffic data (e.g., loop detector data, probe vehicle data, and automatic vehicle identification (AVI) data) to extract traffic characteristics. The most common

* Corresponding author.
E-mail addresses: yurongjie@tongji.edu.cn (R. Yu), 1651363@tongji.edu.cn (L. Han), zhanghuiits@whut.edu.cn (H. Zhang).

https://doi.org/10.1016/j.aap.2021.106085
Received 15 November 2020; Received in revised form 20 January 2021; Accepted 8 March 2021
Available online 24 March 2021

processing method is to extract traffic flow parameters (e.g., road volume, average speed) using the loop detectors upstream and downstream of the crash segment. And further adopt the descriptive statistics of traffic flow parameters to establish crash risk assessment models (Abdel-Aty et al., 2012; Kwak and Kho, 2016; Park and Haghani, 2016). However, given that crashes and traffic conflicts occur at certain points along the roads, it is more important to know the exact traffic variations near the crash locations rather than the aggregated traffic information at roadway segment level. In other words, more attentions should be paid to extract the microscopic traffic flow conditions and the vehicle platoon interactions just around the crash locations. Fortunately, with the development of the new generation of traffic sensing technology, advanced traffic flow collection methods (e.g., roadside lidar, drones) have emerged, and the application of connected and autonomous vehicles (CAV) in the future can also provide individual-level vehicle trajectories. The emerging full-covered and high-resolution trajectory data provides the possibility to extract microscopic traffic flow characteristics at crash locations, which have big potentials to significantly improve real-time crash risk analysis (Wang et al., 2019b). However, to the best of our knowledge, there are only a few studies using full-covered trajectory data for real-time traffic risk analysis.

Moreover, crash events were selected as the risk identification objects for the majority relevant studies (Roshandel et al., 2015). However, road crashes are rare events which is hard to collect enough crash samples in a short time. And it is difficult to know exactly the process preceding of crashes using solely crash data (Kuang and Qu, 2014). In addition, within the active traffic safety management framework, the main focus is to identify and analyze unsafe situations resulting in high-risk events (also named as near-crashes or safety critical events), as such “close calls” could foreshadow actual future crashes (Abbas et al., 2011; Kluger et al., 2016). High-risk events are regarded as any circumstance that requires a rapid evasive maneuver by the participant vehicle to avoid a collision (Dingus et al., 2006). There is a strong correlation between the frequency of high-risk events and crashes, and they have similar causal mechanisms (Guo et al., 2010). With the safety surrogate measure, high-risk event hold potentials in unveiling deeper understanding of crash mechanism. Therefore, it has been chosen as the main analysis object by studies that employed high-resolution traffic data (Wu and Jovanis, 2013; Wali et al., 2019). Thus, unsafe situations that identified as high-risk events were targeted along with their precursor characteristics in this study.

Referring to the abovementioned research gaps, this study aims at developing traffic operation risk prediction models with the emerging traffic sensing data (i.e., full-covered trajectory data) and further unveiling the influencing factors of high-risk events. The main contributions of this study can be summarized as follows:

(1) Utilized full-covered trajectory data rather than the traditional discrete aggregate data to extract traffic operation features.

(2) Developed risk identification and influencing factor analysis of high-risk events to further helping pre-warning in traffic management.

(3) Established high-precision risk identification and pre-warning models, and consider the data heterogeneity problem using a random-parameters logistic regression model.

The rest of the article is organized as follows. The second section reviews the previous studies that focused on high-risk event identification and their influencing factors analysis methods. Then, the next section provides detailed description of the data preparation procedures, which is followed by the methodology section. The fifth section presents the modeling results, and finally, summaries and discussions of the work are presented.

2. Background

2.1. High-risk events identification

The identification of high-risk events was mainly conducted from two approaches: (1) using only vehicle kinematic variables to identify evasive maneuvers and (2) incorporating roadway environment information to develop safety surrogate measures. As for the first approach, high-risk events were identified if the vehicle kinematic parameters exceed certain preset thresholds (e.g., lateral acceleration>0.7 g in Dingus et al. (2006), longitudinal acceleration>0.65 g in Lee et al. (2011)). This approach has been popular, given it only requires the ego vehicle’s motion information; meanwhile, it also caused a high false alarm rate. For the second approach, safety surrogate measures (e.g., Time to Collision (TTC), Deceleration Rate to Avoid a Crash (DRAC)) were used to identify high-risk events (Vogel, 2003; Ambros et al., 2014). Taking into account the motion information of vehicles and surrounding traffic participants, this approach could provide high identification accuracies and can further classify different risk levels of high-risk events.

The current safety surrogate measures could be mainly divided into temporal indicators (e.g., TTC, Post-Encroachment Time (PET)) and non-temporal indicators (e.g., DRAC, Jerks). Among which, TTC was the most popular measurement. However, assuming that consecutive vehicles will keep constant speeds until the collision occurs, TTC cannot be applied to potential conflict scenarios when the following car is slower than the preceding car. To overcome this, many studies had made some modifications on the basis of TTC and proposed new indicators. For instance, Ozbay et al. (2008) proposed a modified traffic to collision (MTTC), which could consider the vehicle velocities along with their relative accelerations. And it was proven that MTTC could cover all collision scenarios and was a very effective indicator to identify high-risk events (Yang, 2012). Therefore, MTTC was used to identify high-risk events in this study.

2.2. High-risk events analysis methods

The high-risk events analyses were mainly conducted to identify their occurrence precursors and the influencing factors, whereas the typical approach is to compare normal traffic operation conditions with high-risk events patterns. Different methods have been utilized and they can be divided into two main categories: generalized linear regression models (Yu et al., 2013; Wang et al., 2015; Shi and Abdel-Aty, 2015) and machine learning models (e.g., neural network in Abdel-Aty and Pande (2005); Abdel-Aty et al. (2008); support vector machine in Yu and Abdel-Aty (2013c); Sun et al. (2014), and Bayesian networks in Hossain and Muromachi (2012)). Among which, the majority machine learning models operate in a non-inferential approach and their results are difficult to interpret (Roshandel et al., 2015; Das et al., 2020), while the results from logistic regression (LR) models are easy to be interpreted. Different kinds of logistic regression models such as matched case-control LR models and Bayesian LR models have been employed and they had been proved to have good model fit and predictive performance (Ivan and Konduri, 2018). Therefore, LR models were utilized for high-risk events prediction and influencing factors analyses.

In addition, the unobserved heterogeneity has always been regarded as one of the most critical issues in traffic safety analyses (Mannering et al., 2016). It was concluded that if the heterogeneous effects were not being properly considered, biased coefficient estimations and erroneous predictions could be obtained (Yu and Abdel-Aty, 2013b; Mannering et al., 2016). To account for the heterogeneity across the high-risk events, the random-parameters logistic regression (RPLR) model has been widely utilized because their parameter estimations can vary across different levels, which is important for capturing unobserved heterogeneity. For instance, Yu and Abdel-Aty (2013b) employed random parameter logistic regression models to develop crash risk evaluation models, and better goodness-of-fit has been achieved. In addition, random-effects logistic regression (RELR) model is another approach to account for the unobserved heterogeneity. Xu et al. (2013) employed Bayesian random intercept logistic regression models to predict the crash risk under different weather conditions to account for the

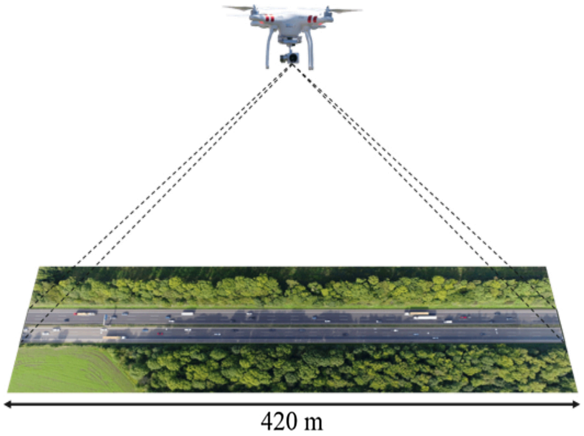


Fig. 1. Highway segment recorded by drone (Krajewski et al., 2018).

unobserved heterogeneity and better goodness-of-fit were achieved. Therefore, in this study, random-effects logistic regression model and random-parameters logistic regression model would be applied to analyze the heterogeneous influencing factors of high-risk events.

3. Data preparation

3.1. *HighD dataset introduction*

The HighD Dataset (short for "The Highway Drone Dataset"), which collects naturalistic vehicle trajectories on German highways using drone videography, was utilized as the empirical data (Krajewski et al., 2018). The drone measurements provide an average of 17 min recordings and cover about 420 m of highway segment, as shown in Fig. 1. The recording works took place at six different highways (split into 60 road segments) in Germany, which is 44,500 driven kilometers and 147 driven hours in total.

Each HighD trajectory dataset contains vehicle trajectories and their physical features. The trajectory of each vehicle is characterized by its position, velocity, and acceleration in longitudinal and lateral directions. The physical features include the width and length of the vehicles. Meanwhile, a global coordinate system was used to record the positions of lane markings and vehicle positions, shown in Fig. 2. Where the origin of the coordinate system is set at the top left corner of the road segment. The x values increase as it moves to the right, and the y values increase when vehicles move towards the bottom of the road.

Given the high coverage of different traffic status and low positioning error of the measured vehicles (generally less than ten centimeters), HighD dataset has been widely applied to traffic simulation modeling and driving behavior analysis (Kruber et al., 2019; Mahajan, 2019).

3.2. *High-risk events identification and characterization*

(a) **High-risk events identification**

As mentioned above, MTTC was selected as the identification indicator of high-risk events in this study, and 2 s was selected as the threshold for high-risk event determination according to a previous study (Meng and Qu, 2012). Fig. 3 shows a typical scenario for car following, and the MTTC of the following vehicle can be calculated as follows:

$$t_1 = \frac{-\Delta v - \sqrt{\Delta v^2 + 2\Delta a D}}{\Delta a} t_2 = \frac{-\Delta v + \sqrt{\Delta v^2 + 2\Delta a D}}{\Delta a}, \ if\ \Delta a \neq 0 \tag{1}$$

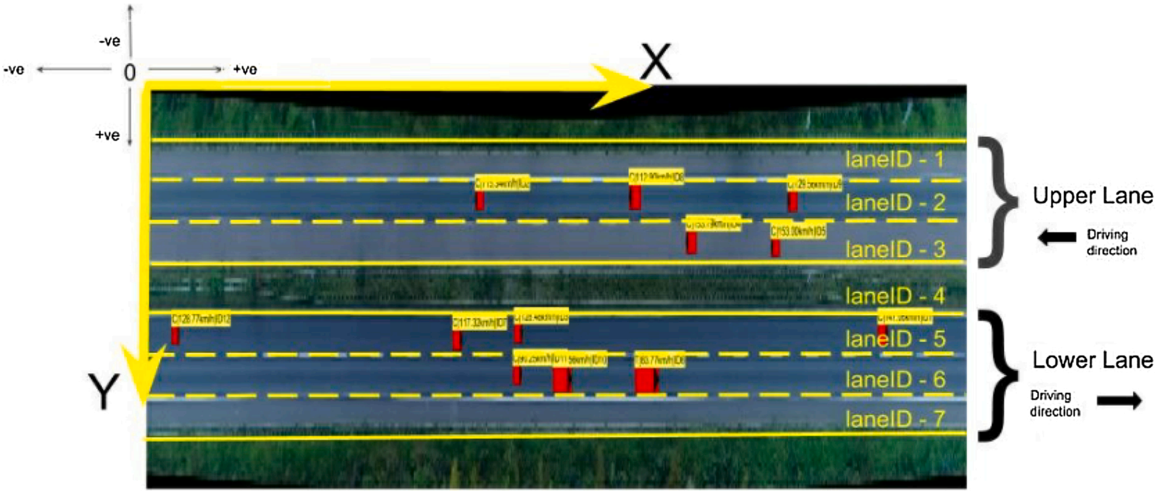


Fig. 2. Global coordinate system in HighD Dataset (Krajewski et al., 2018).

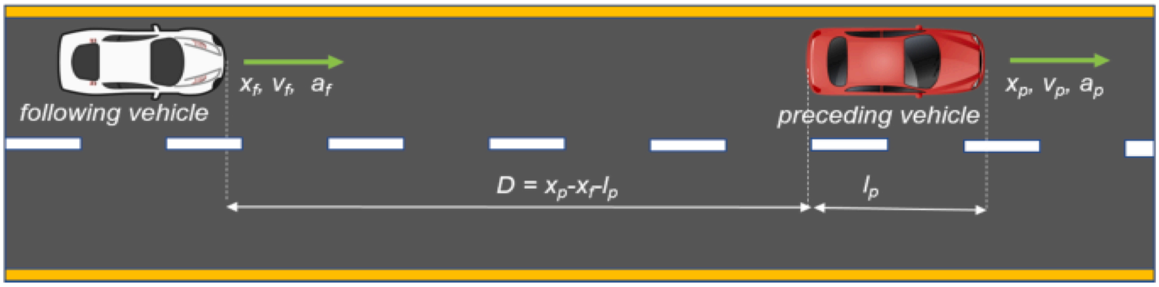


Fig. 3. Car following scenario.

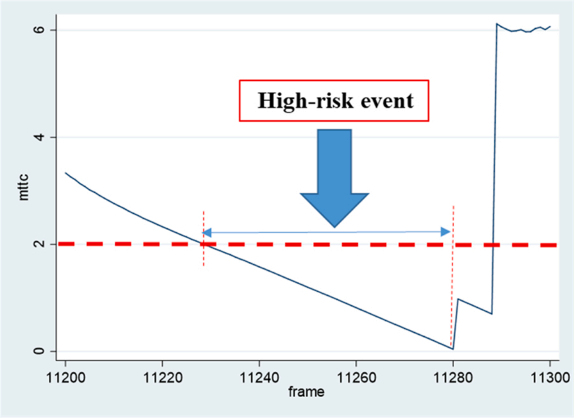


Fig. 4. High-risk events identification.

$$MTTC = \begin{cases} \min(t_1, t_2), & \text{if } t_1 > 0, t_2 > 0 \\ \max(t_1, t_2), & \text{if } t_1 {}^* t_2 > 0 \\ D/\Delta v, & \text{if } \Delta a = 0 \end{cases} \tag{2}$$

$$\Delta v = v_f - v_p, \quad \Delta a = a_f - a_p, \quad D = x_p - x_f - l_p \tag{3}$$

The MTTC value of vehicles in any frame in the trajectory data can be calculated through the above equations (1)-(3). When the MTTC is lower than 2 s, it can be recognized that the vehicle has encountered a high-risk event (Meng and Qu, 2012), as shown in Fig. 4.

(b) **Traffic flow characteristics extraction**

In order to analyze the high-risk events occurrence precursors and explore the influencing factors, traffic flow characteristics need to be extracted to quantitatively describe the traffic flow changes prior to high-risk events. Almost all of the current traffic safety studies use aggregated traffic operation variables (e.g., segment speed, segment volume) to characterize the precursory features of crashes. With trajectory data, the features could be quantified and extracted more precisely. However, it's still a difficult point to characterize the traffic flow state (including temporal and spatial characteristics) using the full-covered vehicle trajectory data.

Furthermore, a traffic flow characteristics extraction method based on trajectory data was proposed to explore the impact of traffic flow changes in seconds on the high-risk events. The extracted temporal-spatial range was set as shown in Fig. 5. And considering the temporal and spatial volatility of traffic flow in the longitudinal and lateral direction, 18 traffic flow characteristic variables were calculated for modeling.

(1) Temporal range

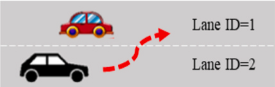
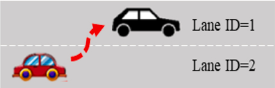
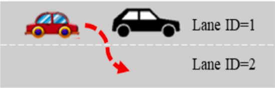
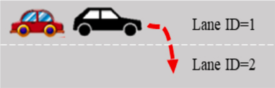
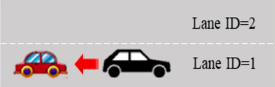
The moment when the MTTC of a vehicle reaches 2 s was set as the zero moment, and 5-second of raw traffic data before the moment are extracted. Then, the 5-second interval is processed into five time slices, which are named as time slice 1 to time slice 5 with time slice 1 refers to the 1-second time slice closest to zero moment.

(2) Spatial range

Since the HighD Dataset include three-lane and two-lane highway segments, in order to standardize the extraction of traffic flow characteristics, the data of two lanes (lane ID = 1, 2) are selected to analyze for each event, and the lane ID setting rules are as follows:

a Lane ID = 1: the lane where the high-risk event happened at the zero moment.
b Lane ID = 2: the lane was set according to the conflict situation of high-risk events, and the rules are shown in Table 1.

Table 1
The rules of lane ID setting.

Schematic diagram of lane ID setting	Types of high-risk event conflicts	Lane ID = 2 setting
	Conflict due to the insertion of the preceding car	The lane where the preceding car is located before lane changing
	Conflict due to the insertion of the following car	The lane where the following car is located before lane changing
	Conflict due to the lane changing of the following car	The lane where the following car is located after lane changing
	Conflict due to the lane changing of the preceding car	The lane where the preceding car is located after lane changing
	No car changes lanes when the gap is too close	The lane on the left side of lane with lane ID = 1

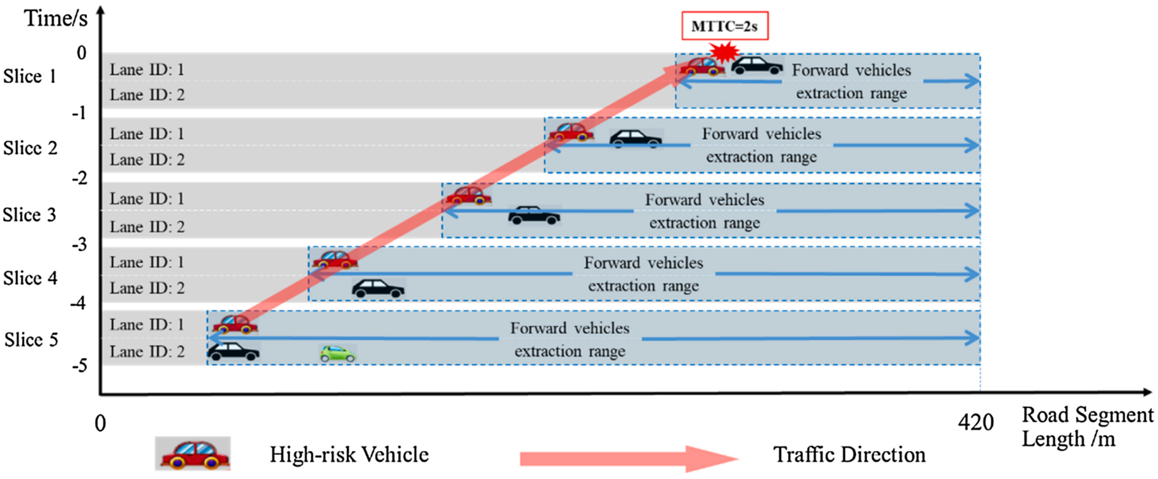


Fig. 5. The temporal-spatial extraction range.

Table 2
The 18 traffic flow characteristic variables for calculation.

Variable level	Variable	Definition
Velocity	MaxYV	Maximum lateral velocity of front vehicles
	aXV	Average longitudinal velocity of front vehicles
	SdXV	Standard deviation of longitudinal velocity of front vehicles
	CvXV	Coefficient of variation of longitudinal velocity of front vehicles
Acceleration	MaxYA	Maximum lateral acceleration of front vehicles
	aACC	Average longitudinal acceleration of front vehicles
	SdACC	Standard deviation of longitudinal acceleration of front vehicles
	CvACC	Coefficient of variation of longitudinal acceleration of front vehicles
	MaxACC	Maximum longitudinal acceleration of front vehicles
Deceleration	aDEC	Average longitudinal deceleration of front vehicles
	SdDEC	Standard deviation of longitudinal deceleration of front vehicles
	CvDEC	Coefficient of variation of longitudinal deceleration of front vehicles
	MaxDEC	Maximum longitudinal deceleration of front vehicles
Vehicle distance	MinD	Minimum distance between vehicles
Time-varying	VfXV	Time-varying stochastic volatility of longitudinal velocity of front vehicles
	VfACC	Time-varying stochastic volatility of longitudinal acceleration of front vehicles
	VfDEC	Time-varying stochastic volatility of longitudinal deceleration of front vehicles
Vehicle number	NuC	The number of front vehicles

Focus on investigating the impact of preceding traffic flow changes on vehicles before high-risk events, the information of vehicles in front of the high-risk vehicle up to the end of the road segment were extracted. So, the spatial extraction range is delimited in the area of two lanes, which will change as the high-risk vehicles travel within 5 s.

(3) Traffic flow characteristic variables extraction

Considering the spatial volatility of preceding traffic flow in the longitudinal and lateral direction. The average value, standard deviation, and other indicators are extracted from the velocity, acceleration, and deceleration level. As for time variability, the time-varying coefficients of longitudinal velocity and acceleration are calculated. In addition, the minimum distance between vehicles and the number of front vehicles are also collected. Eventually, 18 traffic flow characteristic variables were calculated for modeling, as shown in Table 2.

Taking into account the abovementioned temporal and spatial ranges, a total of 180 variables can be obtained. In order to standardize the naming of the large number of variables, a unified nomenclature method was proposed for the variables. The nomenclature includes two numeric characteristics and one letter, as shown in Fig. 6, where the first word represents 18 variables. The second number takes the value of 1, 2, 3, 4, or 5 referring to the five 1-second time slices. And the last number takes the value of 1 or 2 referring to the two lanes. Finally, a total of 180 (18 variables × 5 time slices × 2 lane sections) traffic flow characteristic variables were calculated.

3.3. Non-high-risk events identification and characterization

Besides, to compare the differences between traffic conditions in normal conditions and high-risk events, the non-high-risk events (the MTTC of vehicle never reaches 2 s) were also needed. Then, the case-control data structure was adopted, which was frequently utilized in the disaggregate crash occurrence studies (Yu and Abdel-Aty, 2013a; Chen et al., 2018).

Since the MTTC values of non-high-risk events are bigger than 2, the time when the vehicle reaches the minimum MTTC was set as the zero time. Then the time-space ranges are the same as the extraction range of the high-risk group, and the extraction of the non-high-risk events are below three steps:

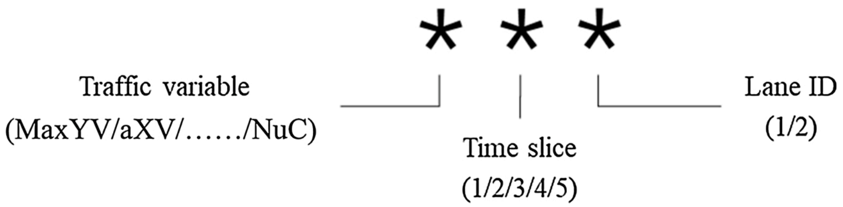


Fig. 6. Nomenclature of the traffic flow characteristic variables.

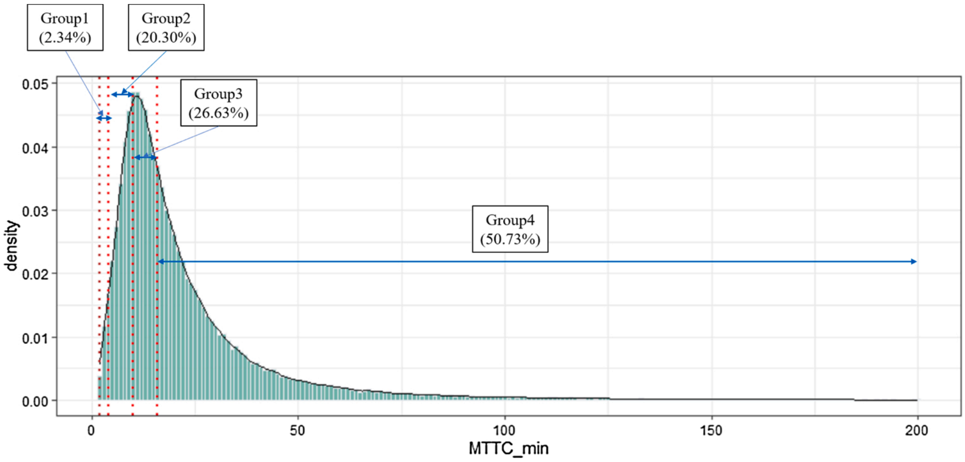


Fig. 7. Distribution of minimum MTTC of vehicles with MTTC always higher than 2 s.

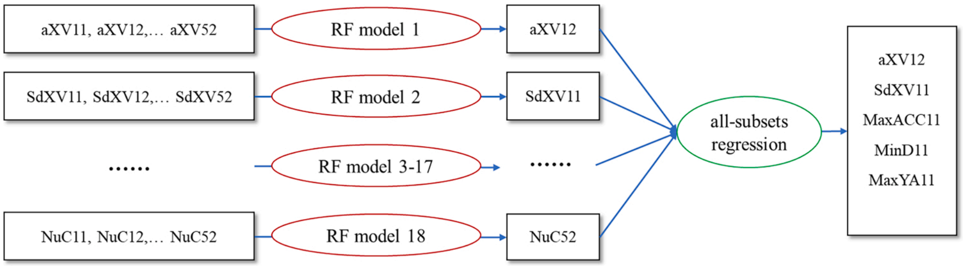


Fig. 8. Modeling variable selection.

Table 3
The 5 modeling variables.

variables	Description	Units	Mean	Std. Dev.
aXV12	Average longitudinal speed of front vehicles in lane two 0−1 s prior to high-risk events	m/s	28.05	7.02
SdXV11	Standard deviation of longitudinal velocity of front vehicles in lane one 0−1 s prior to high-risk events	m/s	2.09	1.60
MaxACC11	Maximum longitudinal acceleration of front vehicles in lane one 0−1 s prior to high-risk events	m/s^2	0.46	0.33
MinD11	Minimum distance between vehicles in lane one 0−1 s prior to high-risk events	m	29.54	31.18
MaxYA11	Maximum lateral acceleration of front vehicles in lane one 0−1 s prior to high-risk events	m/s^2	0.21	0.17

(1) Set the extraction range of non-high-risk events. In order to exclude the impact of the occurrence of high-risk events on the extraction of non-high-risk events, it is assumed that the subsequent impact of high-risk events on normal traffic flow is about 20 s, so the data of 20 s before and after the high-risk events is first removed.

(2) Set the total number of extractions for non-high-risk samples. According to previous research (Ahmed et al., 2012), when the ratio of the case group to the control group is 1:4, the established model has a better fit. Hence, the ratio of the number of high-risk events and non-high-risk events is set to 1:4.

(3) For ensuring that the extracted non-high-risk events can accurately reflect the overall vehicle MTTC distribution, the whole vehicles with MTTC always higher than 2 s is divided into 4 groups based on the minimum MTTC distribution (2∼200 s) as shown in Fig. 7, each group is randomly sampled by the proportion of vehicles.

3.4. Modeling variable selection

Through the abovementioned data process procedures, a total of 1280 events (256 high-risk events and 1024 non-high-risk events) were extracted. To enhance the modeling efficiency, a variable selection procedure was conducted, as shown in Fig. 8.

First, random forest (RF) model, which has been widely used for variable selection (Shi and Abdel-Aty, 2015; Yu et al., 2019), were employed to rank the variables' importance. Considering these variables are originally from 18 categories, random forest models were constructed for each variable type using *randomForest* R package (John Ehrlinger, 2016). The 18 variables that have the greatest impact on high-risk events were firstly obtained from each type of traffic flow characteristic variables. Then, an all-subset regression model was established for final modeling variable selection by using *regsubsets* R package (Thomas Lumley, 2010), and adjust R^2 was used as a criterion for comparing models with different subsets of variables. After checking for the correlation effects between variables, five influencing factors were identified for the further analyses, and their descriptive statistics are shown in Table 3.

4. Methodology

In order to analyze the probability of high-risk events and their heterogeneous influencing factors, three models were established: (1) standard LR model, (2) RELR model, and (3) RPLR model. The standard LR model with fixed coefficients was first estimated to explore the effects of traffic characteristics on the high-risk events, the RELR model was employed to account for unobserved heterogeneity caused by geometric characteristics, and the RPLR model was employed to capture unobserved heterogeneity across individual high-risk events.

Suppose the high-risk events have the outcomes y = 1(high-risk events) and y = 0(non-high-risk events) with respective probability *p* and 1-*p*. The models can be set up as follows:

$$y \sim Binomial(p) \tag{4}$$

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{X}\boldsymbol{B}_{t[i]} + \alpha_{j[i]} \tag{5}$$

where β_0 is the intercept and $\boldsymbol{X}$ is the vector of explanatory variables, $\boldsymbol{B}_{t[i]}$ is the vector of random coefficients for the explanatory variables,

$$\boldsymbol{B}_{t[i]} \sim N\left(\boldsymbol{M}_{\boldsymbol{B}}, \sum_{B}\right), \; for \; t = 1, \ldots, T, \tag{6}$$

where $\boldsymbol{M}_{\boldsymbol{B}}$ represents the mean of the distribution of the coefficients and $\sum_B$ is the covariance matrix representing the variation of the coefficients. *t* stands for the crash unit index (high-risk events observation and their matched non high-risk events cases). *j*[*i*] indexes the segment where observation *i* occurs and $\alpha_{j[i]}$ is the random effects variable defined in the model, which represents the segment-specific random effects in this study:

$$\alpha_{j[i]} \sim N\left(\boldsymbol{U}_{j\gamma}, \sigma_\alpha^2\right), \; for \; t = 1, \ldots, 6, \tag{7}$$

where $\boldsymbol{U}$ is the matrix of segment-level predictors, γ is the vector of coefficients for the segment-level regression, and σ_α is the standard deviation of the unexplained segment-level errors.

For the standard LR model, the $\boldsymbol{B}_{t[i]}$ and $\alpha_{j[i]}$ were fixed for all observations, for the RELR model, $\boldsymbol{B}_{t[i]}$ were a fixed matrix for all observations, and in the RPLR model, $\alpha_{j[i]}$ is not included. The models were estimated using NLOGIT 5 (Greene, 2012).

Besides, to compare the goodness-of-fit and prediction accuracies, AIC, sensitivity, false alarm rate, and AUC value are applied in this study. The AIC was used to measure the goodness of fit of statistical models, as shown in Equation 8. The smaller the value of AIC, the higher the model goodness-of-fit. Sensitivity and false alarm rate are estimated according to the confusion matrix (Table 4) and Equations 9–10. A fixed threshold for classification of 0.5 was adopted, which was widely adopted in the literatures (Yu et al., 2020; Jiang et al., 2020). High

Table 4
Confusion matrix.

	True High-risk events	True Non-high-risk events
Predicted High-risk events	True Positive (TP)	False Positive (FP)
Predicted Non-High-risk events	False Negative (FN)	True Negative (TF)

Table 5
Modeling results based on training data.

Variable	LR model	RELR model	RPLR model
Intercept	−2.410(0.691)**	−2.074(0.927)*	−1.951(0.520)**
aXV12	0.056(0.019)**	0.055(0.020)**	0.048(0.015)**
SdXV11	0.640(0.125)***	0.620(0.129)***	0.449(0.087)***
MaxACC11	1.774(0.480)***	1.755(0.480)***	1.458(0.354)***
MaxYA11	8.519(1.106)***	8.610(1.125)***	1.654(0.095)***
Std. Dev.			0.355(0.062)***
MinD11	−0.370(0.040)***	−0.362(0.042)***	−0.288(0.033)***
α(segment-specific random effects)	N/A	0.060(0.246)**	N/A
AIC	298.8	300.8	296.8
Sensitivity	0.89	0.89	0.98
False Alarm Rate	0.06	0.06	0.06
AUC	0.92	0.92	0.96

Note: Tabular values indicate parameter estimates (standard errors) for each model.

*** Statistically significant at α = 0.001.

** Statistically significant at α = 0.01.

* Statistically significant at α = 0.05.

Table 6
Modeling goodness of fits based on testing data.

	LR model	RELR model	RPLR model
Sensitivity	0.89	0.89	0.98
False Alarm Rate	0.06	0.06	0.06
AUC	0.92	0.92	0.96

sensitivity means the model can predict most of the high-risk events correctly, while a low false alarm rate indicates the model predicts most non-high-risk events correctly. AUC is defined as the area under the receiver operating characteristic (ROC) curve, which can evaluate and compare the prediction accuracy of the model. The value of AUC is between 0 and 1. The larger the AUC value, the better the model prediction is, and the stronger the classification ability.

$$AIC = 2k - 2ln(L) \tag{8}$$

where k is the number of model parameters, L is the model log-likelihood value.

$$True\ Positive\ Rate(Sensitivity) = TP\ /(TP + FN) \tag{9}$$

$$False\ Postive\ Rate(False\ Alarm\ Rate) = FP\ /(FP + TN) \tag{10}$$

5. Modeling results

5.1. Model results

To evaluate the model performance, 70 % of the 256 high-risk events and 1024 non-high-risk events are randomly selected for training, while the rest 30 % are used as the test set to assess the models. Table 5 presents the final modeling results based on training data and Table 6 presents the modeling goodness of fits based on testing data. Parameter estimates, standard errors, and goodness-of-fit for each model are listed. In addition, Table 7 provides the marginal effects for predictor variables by model formulation.

Table 7
Marginal effects for predictor variables by model formulation.

Variable	Marginal effects by model formulation		
	LR model	RELR model	RPLR model
aXV12	0.003	0.003	0.002
SdXV11	0.035	0.038	0.033
MaxACC11	0.097	0.105	0.107
MaxYA11	0.466	0.502	0.121
MinD11	−0.019	−0.021	−0.021

Note: Values indicate the change in probability of high-risk event for 1-unit change in the predictor.

From the results, it can be seen that the three models provided similar modeling outcomes. The five significant variables are all within the time slice of 0−1 second prior to high-risk events occurrence, and the same variables were identified and consistent estimated coefficient signs. Since the RPLR model provides the best model goodness-of-fit (lowest AIC and better AUC values), the marginal effect will be analyzed with this model, and we can conclude that:

(1) The aXV12 is significant with a positive sign, i.e., the average longitudinal speed of front vehicles in lane two 0−1 s prior to high-risk events is positively related to the high-risk events occurrence possibility. Its marginal effect is 0.002, which means that one unit increase of aXV12 would increase the high-risk events occurrence possibility by 0.2 %. This can be interpreted as that increase of velocity in the adjacent lane symbolizes that some risky driving behaviors in front of the vehicle, such as high-speed vehicle insertion or high-speed lane change, which will increase the sideswipe and rear-end crash risk.

(2) The SdXV11 is significant with a positive sign, that is to say, the standard deviation of the longitudinal velocity of front vehicles in lane one 0−1 s prior to high-risk events is positively related with the high-risk events. And its marginal effect is 0.033, meaning that one unit increase of SdXV11 would increase the high-risk events occurrence possibility by 3.3 %. This can be interpreted as that increase of variations of front vehicle's velocity will lead turbulent traffic, which will largely increase the probability of high-risk events.

(3) The MaxACC11 and MaxYA11 are positively significant, which means that maximum longitudinal and lateral acceleration of front vehicles in lane one 0−1 s prior to high-risk events are both positively related to the high-risk events. It can be seen from their marginal effects that one-unit increase of MaxACC11 and MaxYA11, the high-risk events occurrence possibility would increase by 10.7 % and 12.1 %. This can be interpreted as that increase of maximum lateral acceleration and longitudinal acceleration in front vehicles symbolizes that there are some hazardous driving situations such as sudden lane changes and rapid acceleration in the traffic ahead, which will increase the probability of rear-end crash.

(4) The MinD11 is significant with a negative sign, where the minimum distance between vehicles in lane one 0−1 s prior to high-risk events is negatively associated with the high-risk events occurrence possibility. Its marginal effect is -0.021, which means that one unit decrease of MinD11 would increase the high-risk events occurrence possibility by 2.1 %. This can be interpreted as that decrease of the minimum distance between vehicles in front vehicles will lead drivers to take actions such as emergency

Table 8
Variable selection comparison in three pre-warning models.

Velocity level	Original Model	Model -1 s	Model -2 s
Velocity	aXV12(+)		aXV51(-)
	SdXV11(+)		SdXV51(+)
Acceleration	MaxYA11(+)	MaxYA21(+)	MaxYA31(+)
	MaxACC11(+)		
		SdACC21(+)	
			CvACC31(+)
Vehicle distance	MinD11(-)	MinD21(-)	MinD31(-)
Time-varying		VfXV22(+)	

Note: Tabular values indicate variables (the sign of the parameter estimates) for each model.

Table 9
Comparison of model AUC values.

Model	LR model	RELR model	RPLR model
Original Model	0.92	0.92	0.96
Model -1 s	0.86	0.87	0.97
Model -2 s	0.76	0.76	0.97

braking or lane changes, which will largely increase the occurrence probability of high-risk events.

As for the model comparison, based on the goodness of fits of models, we can know that the AIC of the RELP model is higher than that of the LR model, which means that the model fitness is declined while the heterogeneity due to different road sections has been clearly observed. In addition, the AIC values of the RPLR model have a minimum value among the three models, which means that the model can effectively reflect the heterogeneity between individuals, and the RPLP model has the biggest sensitivity (0.98) and AUC value (0.96) among the three models, which means that the model can effectively improve the prediction of the model. The improvement of the goodness of fits is consistent with the conclusions of existing studies (e.g., Yu and Abdel-Aty (2014) (AUC: 0.69 to 0.82); Bakhshi and Ahmed. (2020) (AUC: 0.69 to 0.77))

5.2. Pre-warning model development

Moreover, considering the difference in the time of proactive warning information release, two new models were established by adjusting the time period of the original modeling variables:

- Model -1 s: using information from 2 to 5 time slices prior to high-risk events to obtain 1 s ahead risk warnings;
- Model -2 s: using information from 3 to 5 time slices prior to high-risk events to obtain 2 s ahead risk warnings.

Then, the same three modeling approaches were applied to the two new models. Table 8 shows the variable selection in three models. Table 9 presents the AUC value of the nine established models.

From the results, it can be seen that as the time segment of the model changes, the significant variables of the three models are different.

(1) As for the model -1 s, four significant variables are all within 1–2 seconds, which is the time slice closest to the high-risk events occurrence in the model. The MaxYA21 and MinD21 are still significant variables and have the same coefficient signs as the original model. The SdACC21 is significant with a positive sign. This can be interpreted as that increase of variations of front vehicle platoon's acceleration can also lead to a turbulent traffic condition and increase the crash risk. And the VfXV22 is positively significant. This can be interpreted as that rapid velocity change of vehicles will lead to increased volatility of traffic in adjacent lanes, which will cause dangerous behaviors such as the insertion of vehicles at different speeds and increase the sideswipe and rear-end crash risk.

(2) As for the model -2 s, three significant variables are within 2–3 s, while two significant variables are within 4−5 s. The MaxYA31 and MinD31 are still significant variables and have the same coefficient signs as the two models. The CvACC31 can also reflect the volatility of front traffic and is significant with a positive sign. The aXV51 and SdXV51 are significant in this model, which means that the crash risk can be affected by the traffic state before 4-5 s. This can be interpreted as that increase of variations of front vehicle platoon's velocity will lead to a turbulent traffic operation, which will increase the probability of high-risk events. It is interesting that aXV51 has a significant negative sign. This can be interpreted as that increase of front vehicle platoon's velocity will lead to a smooth traffic state, which will reduce the probability of high-risk events.

Comparing AUC values of models (Table 9), the following conclusions can be drawn: as the warning time releasing earlier, the AUC value of the logit models show a downward trend, which means that the prediction accuracy of the model decreases. In the case of the three models, the random parameter logit models have high AUC values (0.96, 0.97, 0.97) in three models, which means that the model can realize to early warn to drivers or connected vehicles and ensure a better prediction accuracy, thereby helping to improve the effectiveness of active early warning.

6. Summary and discussions

Proactively identify high-risk events and then adopt feasible active traffic management strategies is a key approach to improve freeway traffic safety. However, current studies mainly relied on discrete and segment-based traffic data, where the aggregated traffic flow information could only obtain crash-segment operation status rather than detail traffic characteristics near exact crash locations. While the emerging full-covered and high-resolution trajectory data provide the possibilities of extracting detailed vehicle platoon interaction characteristics. On the other hand, since crashes are rare and random events, high-risk events should be paid more attentions given their higher occurrence probability and consistent causations with crashes.

Therefore, in this study, a high-risk event prediction and influencing factors analysis method based on vehicle trajectory data has been investigated. First, high-risk events were obtained using safety surrogate measures with MTTC less than 2 s. And the traffic operation characteristics within 5 s prior to high-risk event occurrence were extracted based on vehicle trajectory data. Then, a total of three different logistic regression models were established, which are standard LR model, RELR model, and RPLR model.

From the modeling results, it can be seen that the significant variables are consistent among the three developed models, and they are all within the time slice of 0−1 s prior to high-risk events occurrence. The main findings are as follows:

Table 10
AUC values comparison of RPLR and SVM models.

Model	RPLR model	SVM model
Original Model	0.96	0.93
Model -1 s	0.97	0.92
Model -2 s	0.97	0.81

Table 11
Comparison of parameter estimates between models.

Variable	HV model	MV model	LV model	RPLR model in Table 5
Intercept	0.238(0.091)	−2.446 (1.286) **	−2.142 (1.485) **	−1.951 (0.520) ***
Axv12	–	0.073 (0.037) **	–	0.048 (0.015) ***
SdXV11	–	0.331 (0.133) **	1.187 (0.267) ***	0.449 (0.087) ***
MaxACC11	–	1.423 (0.511) ***	4.122 (1.422) ***	1.458 (0.354) ***
MaxYA11	2.089 (0.135) ***	1.630 (0.151) ***	1.695 (0.329) ***	1.654 (0.095) ***
Std. Dev.	0.421 (0.078) ***	0.371 (0.151) ***	0.089 (0.111) ***	0.355 (0.062) ***
MinD11	−0.400 (0.091) ***	−0.299 (0.050) ***	−0.419 (0.094) ***	−0.288 (0.033) ***

Note: Tabular values indicate parameter estimates (standard errors) for each model.

*** Statistically significant at $\alpha = 0.001$.

** Statistically significant at $\alpha = 0.01$.

- For speed-related factors, the speed variance of the front traffic flow and the average speed in adjacent lanes are showed to have positive impacts on high-risk events occurrence probabilities.
- For acceleration-related factors, the maximum acceleration in longitudinal and lateral directions of the front vehicle platoon hold positive influences on the occurrence of high-risk events.
- As for the spatial distance, the minimum distance between vehicles of the front vehicle platoon is showed to have substantial negative impacts on high-risk events occurrence.
- Besides, attempts have tried to construct a lane-change model and a non-lane-change model based on whether there was vehicle lane-changing behavior before high-risk events. However, the results showed that there were no significant differences for the identified influencing variables and their estimated coefficients.

On the basis of understanding the influencing factors of high-risk events, two additional models with pre-warning requirements (2 s or 3 s prior to event occurrence) were established to explore the feasibility of active traffic management. Among the pre-warning models, the RPLR models have the best model goodness-of-fit and prediction accuracy (AUC ≥ 0.96). In addition, the RPLR models were also compared with support vector machine (SVM) with radial basis models, and the results are shown in Table 10. RPLR models have better predictability in three different models, which means it is possible to predict the probability of high-risk events occurrence in 2 s based on the current traffic state.

To explore the impacts of road traffic volume on the model, three traffic volume levels were determined based on the trisection points of traffic volume distribution. And three more individual models have been established, which are high volume level (HV) model, medium volume level (MV) model, and low volume level (LV) model. Comparisons of parameter estimates between the three new models and the RPLR model are shown in Table 11. And it can be seen that MinD11 and MaxYA11 are significant in each model. In the HV model, Axv12, SdXV11, and MaxACC11 are not significant, which may mean that the high traffic volume reduces the longitudinal volatility between vehicles. Besides, there are limited differences between the MV model and the RPLR model. While in the LV model, Axv12 is not significant for high-risk events, which may indicate that the impact of the velocity of adjacent lanes on high-risk events is very small in low traffic volume environment.

In addition, the differentiations between this study and existing literatures using trajectory data are shown in Table 12. It can be seen that although different vehicle trajectory datasets were adopted, the majority studies analyzed crash risk influencing factors with traditional approaches or either focused on specific vehicle interaction scenarios. In this study, the analysis level was extended to the microscopic vehicle interactions perspective. Microscopic traffic operating characteristics were extracted based on the individual vehicle trajectory data and the relationship between these features and high-risk events was explored.

As for the perspective of formulating improvement measures, the results of models showed that the disturbed traffic flows and too close following distance between vehicles have positive impacts on high-risk events occurrence. Therefore, reducing the speed variance should be the main target, such as using variable speed limits. Moreover, to avoid the critical risky behaviors such as emergency brakings and sudden cutting-ins, relevant Cooperative Vehicle Infrastructure System (CVIS) could play a role. Furthermore, the real-time prediction results could support the connected and autonomous vehicles' (CAV) trajectory planning decisions.

Furthermore, as this is the first attempt to analyze high-risk events with vehicle trajectory data, there is still plenty room for future studies. First, this study employed case-control data structure with a fixed high-risk events and non-high-risk events ratio (1:4), while the imbalance of crash dataset in the real traffic environment was not considered. Second, sensitivity analyses of identifying the spatial extraction range for interactive vehicle platoons need to be further explored. In addition, the severity and the conflict type of high-risk events should also be considered.

Table 12
Comparison of this study and existing crash risk studies using trajectory data.

Authors	Trajectory data	Analysis level	Variable extraction
Wang et al. (2019a)	Sampled floating car trajectory data	Analysis crash propensity at traffic platoon perspective	Characteristics of the traffic platoon *(i.e., average speed of the platoon)*
Li et al. (2020)	CV emulated trajectory data	Predict crash probability of road segment	Segment-based aggregated traffic variables *(i.e., average speed of the segment)*
Chen et al. (2020)	NGSIM trajectory dataset	Explore crash risk during lane-changing process	Traffic variables between vehicles involving in lane-changing *(i.e., mean distance between lane-changing vehicle and leading vehicle)*
This study	HighD trajectory dataset	Predict high-risk events considering individual driving behavior changes and traffic flow volatility	Front traffic flow variables based on individual vehicle trajectory *(i.e., maximum lateral velocity of front vehicles)*

CRediT authorship contribution statement

Rongjie Yu: Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision. **Lei Han:** Investigation, Writing - original draft, Writing - original draft, Writing - review & editing, Visualization. **Hui Zhang:** Methodology, Writing - original draft, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

This study was supported by the National Key R&D Program of China, No. 2019YFB1600703 and the National Natural Science Foundation of China (NSFC), No. 71771174 and No. 52072289. Sponsored by "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission

References

Abbas, M., Higgs, B., Medina, A., Yang, C.D., 2011. Identification of warning signs in truck driving behavior before safety-critical events. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 558–563.

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36 (1), 97–108.

Abdel-Aty, M., Pande, A., Das, A., Knibbe, W., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transp. Res. Rec.: J. Transp. Res. Board 2083, 153–161.

Ahmed, M.M., Abdel-Aty, M.A., 2012. The viability of using automatic V ehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13 (2), 459–468.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transp. Res. Rec. 2280 (1), 60–67.

Ambros, R.T., Paukrt, J., Ambros, J., Turek, R., Paukrt, J., 2014. Road safety evaluation using traffic conflicts: pilot comparison of micro-simulation and observation-Jiří. International Conference on Traffic and Transport Engineering-Belgrade.

Bakhshi, A.K., Ahmed, M.M., 2020. Practical advantage of crossed random intercepts under Bayesian hierarchical modeling to tackle unobserved heterogeneity in clustering critical versus non-critical crashes. Accid. Anal. Prev. 149, 105855.

Chen, Z., Qin, X., Shaon, M.R.R., 2018. Modeling lane-change related crashes with lane-specific real-time traffic and weather data. J. Intell. Transp. Syst. 22 (4), 291–300.

Chen, Q., Gu, R., Huang, H., Lee, J., Zhai, X., Li, Y., 2020. Using vehicular trajectory data to explore risky factors and unobserved heterogeneity during lane-changing. Accid. Anal. Prev. 151, 105871.

Das, S., Dutta, A., Dey, K., Jalayer, M., Mudgal, A., 2020. Vehicle involvements in hydroplaning crashes: applying interpretable machine learning. Transp. Res. Interdiscip. Perspect. 6, 100176.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Bucher, C., 2006. The 100-Car Naturalistic Driving Study, Phase II-Results of the 100-Car Field Experiment. United States. Department of Transportation. National Highway Traffic Safety Administration.

Ehrlinger, John, 2016. The Randomforest Package.

FHWA, 2020. Highway Safety Improvement Program (HSIP) [Acesse Data: 10/18/2020]. https://safety.fhwa.dot.gov/hsip/hsip.cfm.

Greene, W.H., 2012. NLOGIT 5. Econometric Software, Inc., Plainview, New York.

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transp. Res. Rec. 2147 (1), 66–74.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.

Ivan, J.N., Konduri, K.C., 2018. Crash severity methods. Safe Mobility: Challenges, Methodology and Solutions. Emerald Publishing Limited.

Jiang, F., Yuen, K.K.R., Lee, E.W.M., 2020. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accid. Anal. Prev. 141, 105520.

Katrakazas, C., Quddus, M., Chen, W., Deka, L., 2015. Real-time motion planning methods for autonomous on-road driving: state-of-the-art and future research directions. Transp. Res. Part C: Emerg. Technol. 60, 416–442.

Kluger, R., Smith, B.L., Park, H., Dailey, D.J., 2016. Identification of safety-critical events using kinematic vehicle data and the discrete fourier transform. Accid. Anal. Prev. 96, 162–168.

Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highd dataset: a drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 2118–2125.

Kruber, F., Wurst, J., Chakraborty, S., Botsch, M., 2019. Highway traffic data: macroscopic, microscopic and criticality analysis for capturing relevant traffic scenarios and traffic modeling based on the highD data set. arXiv preprint arXiv, 1903.04249.

Kuang, Y., Qu, X., 2014. A review of crash surrogate events. Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management, pp. 2254–2264.

Kwak, H.C., Kho, S., 2016. Predicting crash risk and identifying crash precursors on korean expressways using loop detector data. Accid. Anal. Prev. 88, 9–19.

Lee, S.E., Simons-Morton, B.G., Klauer, S.E., Ouimet, M.C., Dingus, T.A., 2011. Naturalistic assessment of novice teenage crash experience. Accid. Anal. Prev. 43 (4), 1472–1479.

Li, P., Abdel-Aty, M., Cai, Q., Yuan, C., 2020. The application of novel connected vehicles emulated data on real-time crash potential prediction for arterials. Accid. Anal. Prev. 144.

Mahajan, V., 2019. Real-time Driving Intention Prediction and Crash Risk Estimation From Naturalistic Driving Data Using Machine Learning.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Meng, Q., Qu, X., 2012. Estimation of rear-end vehicle crash frequencies in urban road tunnels. Accid. Anal. Prev. 48, 254–263.

NHTSA, 2016. Large-scale Field Test of Forward Collision Alert and Lane Departure Warning Systems, Washington, DC. Obtained. from: https://www.nhtsa.gov/research-data/crash-avoidance/crash-warning-systems. [Acesse Data: 10/18/2020].

Oh, C., Oh, J., Ritchie, S., Chang, M., 2001. Real-time estimation of freeway accident likelihood. In: 80th Annual Meeting of the Transportation Research Board. Washington, DC.

Ozbay, K., Yang, H., Bartin, B., Mudigonda, S., 2008. Derivation and validation of new simulation-based surrogate safety measure. Transp. Res. Rec. 2083 (1), 105–113.

Park, H., Haghani, A., 2016. Real-time prediction of secondary incident occurrences using vehicle probe data. Transp. Res. Part C 70, 69–85.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–211.

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

Sun, J., Sun, J., Chen, P., 2014. Using support vector machine models for real-time crash risk prediction on urban expressways. In: 2014 TRB Annual Meeting. Washington D. C.

Thomas Lumley, 2010. The Leap Package.

Vogel, K., 2003. A comparison of headway and time to collision as safety indicators. Accid. Anal. Prev. 35 (3), 427–433.

Wali, B., Khattak, A.J., Karnowski, T., 2019. Exploring microscopic driving volatility in naturalistic driving environment prior to involvement in safety critical events—concept of event-based driving volatility. Accid. Anal. Prev. 132, 105277.

Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015. Real-time crash prediction for expressway weaving segments. Transp. Res. Part C Emerg. Technol. 61, 1–10.

Wang, J., Luo, T., Fu, T., 2019a. Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach. Accid. Anal. Prev. 133, 105320.

Wang, L., Abdel-Aty, M., Ma, W., Hu, J., Zhong, H., 2019b. Quasi-vehicle-trajectory-based real-time safety analysis for expressways. Transp. Res. Part C Emerg. Technol. 103, 30–38.

World Health Organization, 2018. Global Status Report on Road Safety 2018: Summary (No. WHO/NMH/NVI/18.20). World Health Organization.

Wu, K.F., Jovanis, P.P., 2013. Defining and screening crash surrogate events using naturalistic driving data. Accid. Anal. Prev. 61, 10–22.

Xu, C., Wang, W., Liu, P., 2013. Identifying crash-prone traffic conditions under different weather on freeways. J. Saf. Res. 46, 135–144.

Yang, H., 2012. Simulation-based Evaluation of Traffic Safety Performance Using Surrogate Safety Measures. Doctoral dissertation. Rutgers University-Graduate School-New Brunswick.

Yu, R., Abdel-Aty, M., 2013a. Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. Accid. Anal. Prev. 56 (5), 1–58.

Yu, R., Abdel-Aty, M., 2013b. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. Accid. Anal. Prev. 58, 97–105.

Yu, R., Abdel-Aty, M., 2013c. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accid. Anal. Prev. 50, 371–376.

Yu, R., Zheng, Y., Abdel-Aty, M., Gao, Z., 2019. Exploring crash mechanisms with microscopic traffic flow variables: a hybrid approach with latent class logit and path analysis models. Accid. Anal. Prev. 125, 70–78.

Yu, R., Wang, Y., Zou, Z., Wang, L., 2020. Convolutional neural networks with refined loss functions for the real-time crash risk analysis. Transp. Res. Part C Emerg. Technol. 119, 102740.