Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach

Junhua Wang[a], Tianyang Luo[a], Ting Fu[a,b,*]

[a] College of Transportation Engineering, Tongji University, 4800 Cao'an Highway, Shanghai, 201804, China
[b] Department of Civil & Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada

ARTICLE INFO

*Keywords:*
Urban expressway
Floating car trajectory
Traffic platoon
Crash propensity prediction
Binary logistic regression
Support vector machine

ABSTRACT

Predicting crash propensity helps study safety on urban expressways in order to implement countermeasures and make improvements. It also helps identify and prevent crashes before they happen. However, collecting real-time wide-coverage traffic information for crash prediction has been challenging. More importantly, previous studies have failed to consider the characteristics of the traffic platoon (vehicle group) that the crash vehicle belongs to before the crash occurs. This paper aims to model crash propensity based on traffic platoon characteristics collected by the floating car method, which provides a time-efficient and reliable solution to collecting traffic information. Crash and floating car data are collected from the Middle Ring Expressway in Shanghai, China. Both the binary logistic model and the support vector machine are applied. A data preparation method, involving crash data filtering, floating car data filtering and data matching on the road network, is introduced for the safety analysis purpose. Results suggest that the traffic platoon information collected from floating cars accompanied works reasonably in predicting crashes on expressways. The support vector machine, with an overall accuracy of 85%, outperformed the binary logistic model which had an overall accuracy of 60%. Results further suggest the application of floating car technologies and the support vector machine in real-time crash prediction. Despite this, the study also concludes the merits of the binary logistic model over the support vector machine model in explaining the impact of different factors that contribute to crash occurrences.

## 1. Introduction and Literature Reivew

### 1.1. Crash prediction in ITS industry

Traffic crashes on highways can lead to higher chances of serious injuries and fatalities with the increased operating speed (Brubacher et al., 2018; Yu et al., 2018), compared to urban streets. Huge loss of life and property has been recorded due to the high lethality of these crashes (Wang et al., 2019). For instance, 96% percent of road traffic fatalities happened on highways in the US in 2016 (US DOT, 2017). Transport Canada (2017) has reported that 57% of fatal crashes and 24% of injury crashes happened on rural locations which include highways and roads with the speed limit over 60 km/h in the year of 2015. Crashes on the expressway are likely to be more fatal with its highest operating speed among all types of highway. In China, crashes that happened on expressway locations caused 32% of direct property loss and over 9% of the total death roll nationwide (PSM China, 2017). Expressway crashes have remained as a major concern in field of road safety.

ITS applications have been widely implemented to improve traffic management on expressways, in terms of both efficiency and safety. Real-time crash prediction helps identify and prevent crashes before they happen. Therefore, it has become a hot topic in ITS industry, and alerting systems with real-time crash prediction are regarded as a promising solution to road safety issues in urban expressway environments. Previous studies have proved the efficiency of real-time crash prediction in improving safety on expressways or highways (Oh et al., 2005; Lee et al., 2002, 2003; Abdel-Aty et al., 2004; Abdel-Aty and Pande, 2005; Wu et al., 2018). The performance of real-time crash prediction relies greatly on, in addition to the performance of the prediction model applied, a time-efficient and reliable data collection method.

### 1.2. Overview of data sources for crash prediction

Collecting traffic data has been traditionally limited to sensors at fixed locations, such as inductive loops, infrared, ultrasonic and microwave sensors, and video cameras (Oh et al., 2005; Lee et al., 2002,

[*] Corresponding author at: College of Transportation Engineering, Tongji University, 4800 Cao'an Highway, Shanghai, 201804, China.
*E-mail addresses:* benwjh@163.com (J. Wang), lty-paul@qq.com (T. Luo), ting.fu.tj@gmail.com (T. Fu).

https://doi.org/10.1016/j.aap.2019.105320
Received 14 May 2019; Received in revised form 17 September 2019; Accepted 26 September 2019
Available online 04 October 2019

2003; Abdel-Aty et al., 2004; Abdel-Aty and Pande, 2005; Ahmed et al., 2012; Golob and Recker, 2003; Wang et al., 2015; Hassan and Abdel-Aty, 2013). Loops have been standards in many jurisdictions and have been widely used (Coifman, 2005; Fu et al., 2017). However, they are being gradually replaced by other techniques due to the high maintenance cost (Klein et al., 2006; Ahmed and Abdel-Aty, 2012). The accuracy of infrared, ultrasonic and microwave sensors has always been challenged, especially when they are used under high traffic volume conditions (Lesani et al., 2015). Several studies have used video cameras in data collection for road safety analysis (Fu et al., 2018; Guo et al., 2019; Wu et al., 2019; Xing et al., 2019). Video cameras are a good alternative, but video data are hard to obtain due to privacy and security concerns (Zimmer, 2005). Overall, these stationary sensors collect traffic information discretely and only from locations where they are installed but fail short in collecting data continuously over the road network due to their limited coverage.

Floating car trajectory has become a practical data source in recent decades because of its various merits including being low cost, having wide coverage and being easily accessible. Several studies have used floating car data in collecting traffic information for different purpose. Some looked at traffic conditions (Cathey and Dailey, 2002; Rahmani et al., 2014, 2015; Dewulf et al., 2015; Wang, 2012). For example, Rahmani investigated route travel time with the help of the floating car technology (Rahmani et al., 2014, 2015). Cathey and Dailey (2002) proposed the TriMet and ProbeView systems based on floating car data collected from public transit system to provide commuters real-time transit information. Some explored the application of the floating car technology for speed monitoring. For instance, Diependaele et al. (2016) used floating car information to monitor free flow speed in urban road environments. Vanlommel et al. (2015) applied the floating car technology to help section control (also called average speed control) and found that the system worked efficiently in reducing speed limit violations and speed differences between vehicles.

The floating car technology is a practical data collection method for real-time crash propensity prediction as being an installation-free method with sufficiently large coverage. Some studies have applied floating car data in road safety applications. Only a few of them have tried to use it in crash prediction. Sun and Sun (2015) introduced a dynamic Bayesian network (DBN) model in crash prediction using data from floating car and smart phone. The DBN model was able to achieve a prediction accuracy of 76.4%. However, research on using floating car data in crash prediction has still been quite limited, especially for expressway environments.

### 1.3. Overview of modeling approaches for crash prediction

Crash prediction mainly replies on modeling approaches. Modeling crashes or crash risk has been always an important part in the area of road safety as it helps identify issues and factors that lead to higher crash risks, provide support for treatment implementations, and prevent crashes from happening. Different safety models have been proposed within the past few decades. Recently, different advanced models have been proposed in order to achieve a better accuracy in crash prediction. These models includes advanced regression models derived from traditional methods (Oh et al., 2005; Lee et al., 2002; Golob and Recker, 2003; Abdel-Aty and Abdalla, 2004) or new prediction models based on state-of-the-art machine learning approaches (Abdel-Aty and Abdalla, 2004; Sun and Sun, 2016; Schlögl et al., 2019; Formosa et al., 2019).

Machine learning techniques, such as the support vector machine algorithm (SVM) and the fuzzy clustering models, provide a novel solution to investigating the traditional topic of safety using a new and potentially more accurate way (Pan et al., 2017). For instance, Chang (2005) compared the negative binomial regression and the artificial neural network in modeling the crash frequency on freeways, and found that the artificial neural network slightly outperformed the negative binomial regression. Yu and Abdel-Aty (2014) compared the fixed parameter logistic model, the SVM model, and the random parameter logit model in predicting predict crash injury severity on a mountainous freeway based on real-time traffic and weather data. Sun and sun (2016) proposed a new model based on clustering algorithm and SVM model get an accuracy of 78.0% in crash prediction. Theofilatos et al. (2019) compared different machine learning methods (including K-Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forest, SVM and Shallow Neural Network) and the Deep Feedforward Neural Network (DFNN) for real-time crash prediction, and found that the DFNN had the more balanced performance in terms of different performance metrics compared to other models. However, applications of using this type of technique in safety analysis and crash prediction have still been limited. In addition, despite the great improvements achieved in crash modeling, due to the limitations of data collection methods applied, many of these studies failed to provide a reliable prediction outputs (Oh et al., 2005; Lee et al., 2002).

Most previous studies have used traffic data collected at fixed locations where the crash occurs; however, the surrounding traffic environment (the group of vehicles), that the crash vehicle has been in before the crash happens, can also have a strong impact on the outcome of the crash. In traffic engineering, the group of vehicles that travel closely together is defined as a traffic platoon (Zabat et al., 1995). Therefore, it is important to consider characteristics of the traffic platoon which the crash vehicle belongs to before the crash occurs. However, most previous studies have failed to consider the impact of the traffic platoon on the outcome of crashes.

### 1.4. Purpose and scope of the study

This paper aims to model traffic crashes using floating car trajectory data. This is intent to address the mentioned gaps in predicting traffic crashes with the help of an easily accessible data collection method. A study focusing on predicting vehicle crashes on the Middle Ring Expressway in Shanghai, China, is conducted. Crash data is originally from the Road Network Monitoring Center, the Shanghai Municipal Communications Commission (RNMC SMCC). The raw floating car database involves over 110,000 taxis throughout the city of Shanghai. To filter out the data of the related floating vehicles from the large database, and to match them with the crash data on the selected expressway, a data preparation approach is applied. The concept of traffic platoon is introduced in identifying floating cars accompanied (floating cars that are in the same traffic platoon with the crash vehicle at the occurrence of the crash). A methodology is proposed which predicts crash propensity based on the traffic platoon characteristics collected from floating cars accompanied. The binary logistic regression model and the SVM model are trained, validated and compared using the crash data and filtered floating car data. As shown in Fig. 1 which provides the flowchart of the study, this paper will 1) provide details about the data preparation work, 2) describe the methodology applied, and 3) discuss the results of the models. Finally, conclusions are provided regarding the keying findings, the main contribution, the limitations and future work of this study.

## 2. Data Preparation

### 2.1. Data source

To build the crash prediction model, the Shanghai Middle Ring Expressway was selected, as shown in Fig. 2. With a total length of 69.8 km, it is one of the four main ring expressways located in Shanghai. Data in the study include: 1) crash records; and 2) floating car data related to these crashes.

Historical crash data are from the RNMC SMCC. The crash data include records for the crashes occurred on the Shanghai Middle Ring Expressway during the two months from August 1 to September 30,

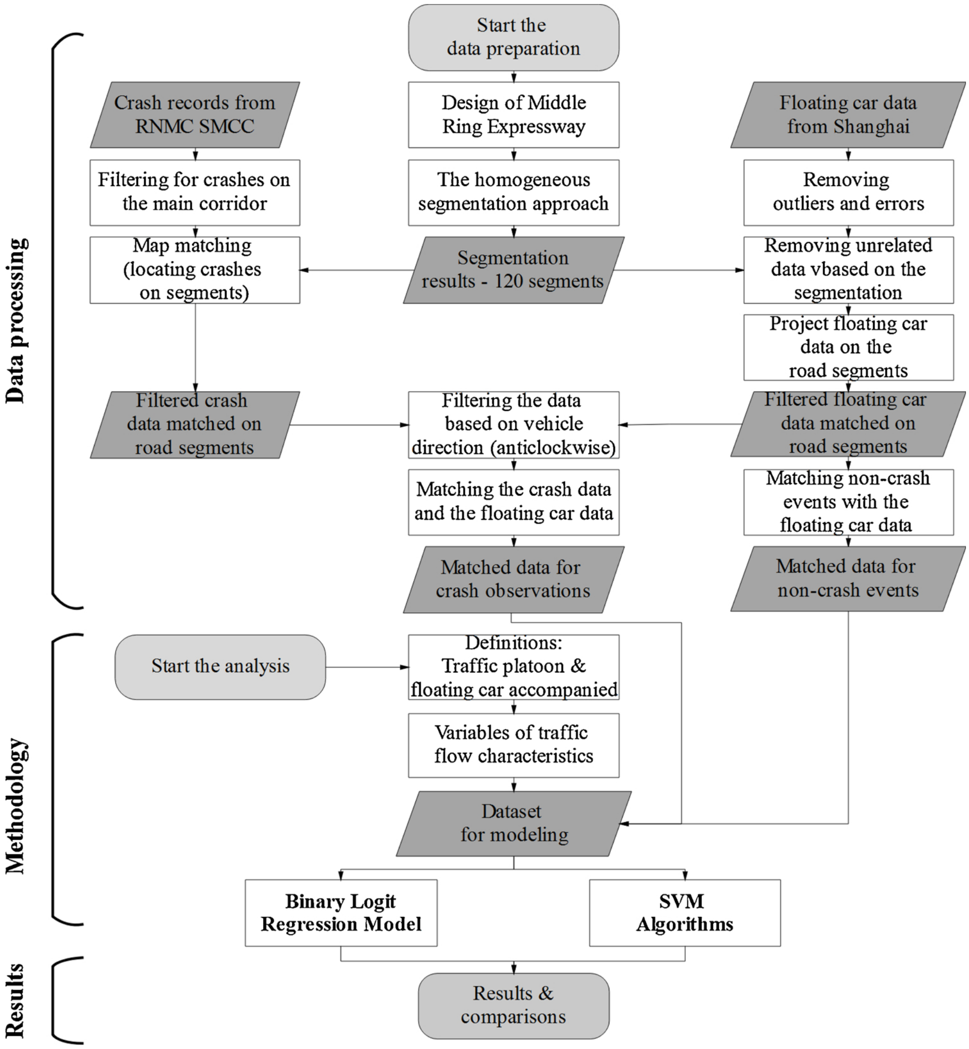

Data processing
Start the data preparation
Crash records from RNMC SMCC
Design of Middle Ring Expressway
Floating car data from Shanghai
Filtering for crashes on the main corridor
The homogeneous segmentation approach
Removing outliers and errors
Map matching (locating crashes on segments)
Segmentation results - 120 segments
Removing unrelated data vbased on the segmentation
Project floating car data on the road segments
Filtered crash data matched on road segments
Filtering the data based on vehicle direction (anticlockwise)
Filtered floating car data matched on road segments
Matching the crash data and the floating car data
Matching non-crash events with the floating car data
Matched data for crash observations
Matched data for non-crash events
Methodology
Start the analysis
Definitions: Traffic platoon & floating car accompanied
Variables of traffic flow characteristics
Dataset for modeling
Binary Logit Regression Model
SVM Algorithms
Results
Results & comparisons

**Fig. 1.** Flowchart of the study.

2015. A total number of 550 crashes are involved, with 488 of them observed on the corridor and the rest 62 on ramps. This study focuses on the traffic crashes that happened on the main corridor. In the original database, locations where these crashes happened, were recorded manually by the local police department in textual descriptions using road names and landmarks as references; therefore, location information needed to be converted into geographic coordinates.

The floating car database is originally sourced from over 110,000 taxis. The database includes location (in geographic coordinates), timestamp, instantaneous speed, heading direction and other

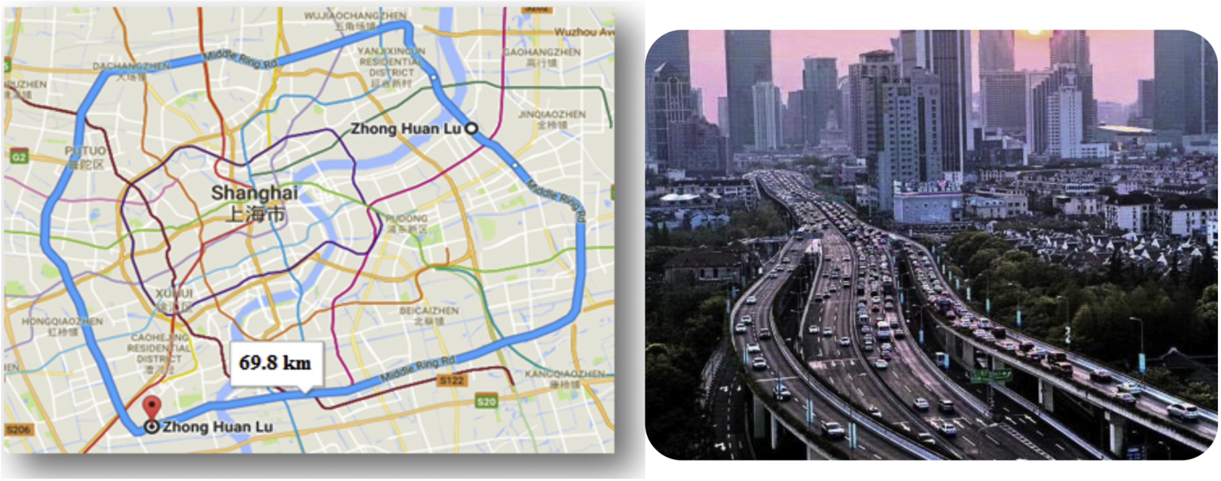


**Fig. 2.** Map of Shanghai Middle Ring Expressway.

**Table 1**
Information in the Floating Car Database.

| Field Name | Explanation | Symbol | Data Type | Data Example |
|---|---|---|---|---|
| **Data Number** | ID for each data record | ID | NUMBER(110) | 90221738 |
| **Trajectory ID** | ID of the car trajectory, using the unique ID of the GPS device | MTDID | NUMBER(100) | 113760 |
| **Timestamp** | Time of the record when it happens | GATHER_TIME | TIMESTAMP(6) | 2015-08-01 16:57:48 |
| **longitude** | Longitude, East (in decimal degrees) | LONGITUDE | NUMBER(106) | 121.397789 |
| **latitude** | Latitude, North (in decimal degrees) | LATITUDE | NUMBER(106) | 31.167996 |
| **Instantaneous Speed** | Instantaneous speed when it happens (km/h) | SPEED | NUMBER(6,2) | 63 |
| **Status** | If the taxi is occupied: *0 empty car, 1 heavy car, 2 task car, 3 other* | STATUS | NUMBER | 1 |
| **Heading Direction** | Moving direction of the car, numbered counterclockwise from true north as: *0 – [0° ~ 45°), 1 – [45° ~ 90°),…,7-[315° ~ 360°)* | DIRECTION | NUMBER | 3 |
| **Locate** | Whether the car is correctly positioned: *1 – successfully positioned, 0 – positioning failed* | ISLOCATE | NUMBER | 1 |
| **Data Reception Time** | The time when the server received the record | RECEIVED_TIME | TIMESTAMP(6) | 2015-08-01 16:57:52 |

information of floating vehicles, observed every 20 s, for all these taxis throughout the city of Shanghai during the previous years. A total number of 3,231,112 observations is included in the original database. Table 1 provides details of the database, including names of the parameters (field name, in the database), explanations of the parameters, their symbols in the database, the type of data, and the example of the data.

### 2.2. Expressway segmentation

In order to match the crash data and the floating car data on the selected expressway, a segmentation process was conducted which divides the expressway into sub-segments. The homogeneous segmentation approach, as suggested in the Highway Safety Manual (AASHTO, 2010), was applied. The road section (the expressway) was divided into homogeneous sections according to 1) ramp locations, 2) number of lanes, 3) lane width, 4) road marking design, 5) speed limit, and 6) curvature. To obtain the related information, the geometric design of the expressway was referenced. However, according to (Cafiso et al., 2018), short lengths in segmentation lead to uncertain results in crash analysis. As recommended in (Ogle et al., 2011), 5 short segments that were less than 160 m were excluded in the study. Long segments may also “create a bias in the identification” of the crash locations for the safety analysis purpose (Cafiso et al., 2018). Therefore, they should be further divided if necessary (no over-length segment was observed in this study).

Fig. 3 presents the segmentation results. As shown in Fig. 3a, after segmentation, the expressway was divided into 120 segments. In order to make the procedure of matching floating car data easier, the segments were grouped into five different groups based on the five road sections determined by the pentagon shape of the expressway. This helps avoid unnecessary iteration steps between the GPS records and road segments in a different road section. Fig. 3**b** provides the distribution of the lengths of the segments. The lengths range from 200 m to 1000 m, which are reasonable according to (Cafiso et al., 2018). To simplify the data matching process, straight lines connecting the nodes of segmentation were used to represent the segments. Note that: offsets occurred when simply using straight lines representing the real-world arced segment, as described in Fig. 3**c**. Offsets from straight-line segmentation for different road sections has been calculated. In this study, the largest offset is 10 m, which is accounted in matching the floating car data as discussed later in this paper.

### 2.3. Crash data preparation

As a common issue when using historical crash data, the precision in identification of crash locations in the study has not always been reliable (Qin and Wellner, 2012). In this study, the crash locations from the police report were recorded as road sections, which were determined mainly according to locations of traffic ramps. Therefore, most of road sections in the crash location records share the same nodes with the road segments defined before in the paper. According to the number of road segments that the location information covers, records of crash location information can be classified into three types: 1) records that cover one road segment; 2) records that cover two road segments; and 3) records that cover more than two road segments. Crash records with a location that covers one or two road segments are investigated, while those with a location that covers over two road segments are considered imprecise and have thus been excluded (57 of them). To identify the locations of the crashes on the road segments, crashes with location records covering one segment were mapped in the middle of that segment. Those with the location records covering two road segments were mapped at the road node connecting the two segments.

### 2.4. Floating Car data preparation

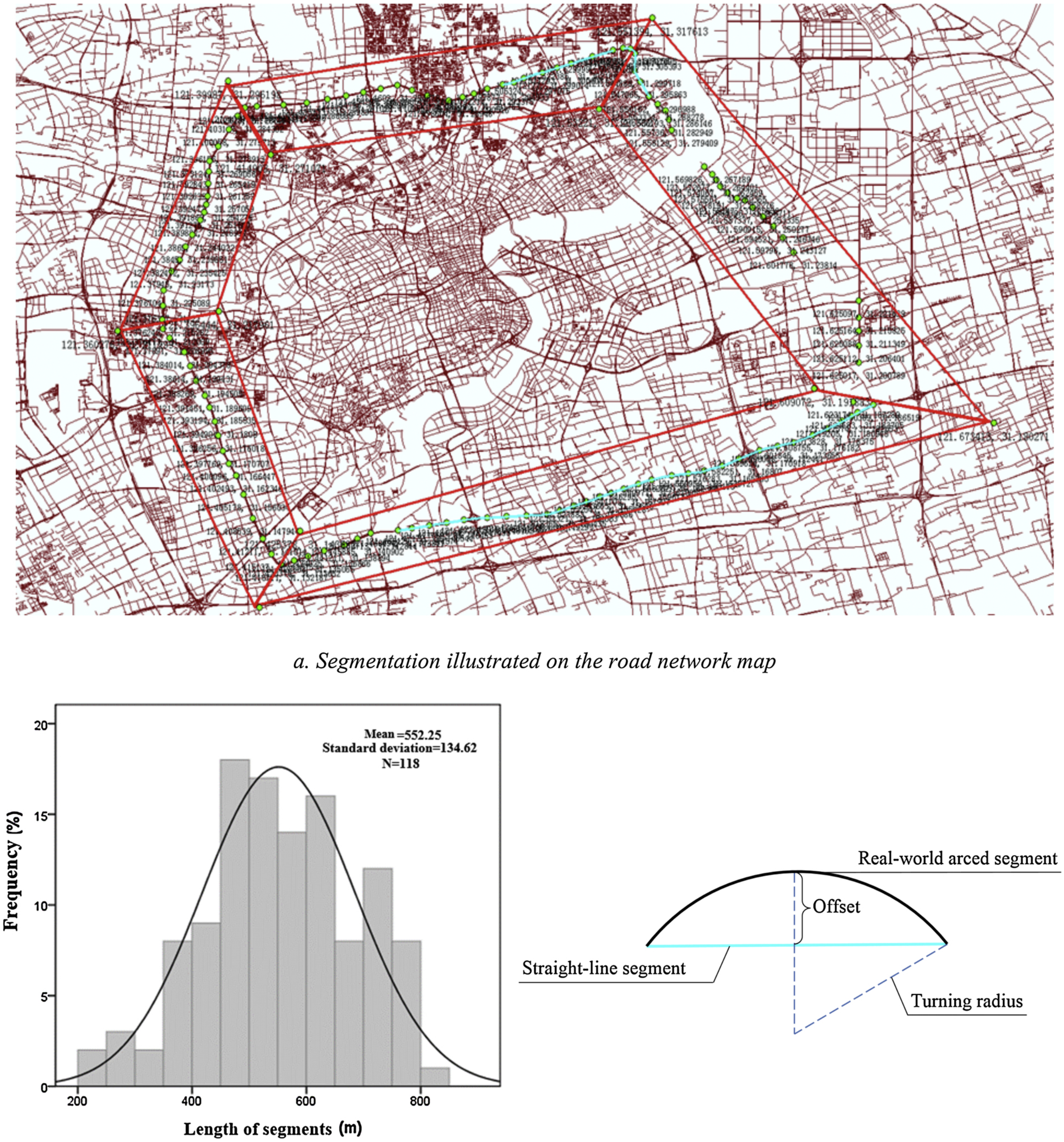
GPS data for the floating vehicles are not good enough to be used directly due to the errors of the technology (Zhang et al., 2011). Several issues existed for using the raw GPS floating car database: 1) different types of outliers (errors in the location or speed that are too large to fix) were recorded and should be removed; 2) errors were generated from hardware failure of the GPS device (the device failed and sent unusual repeated or chaotic data); and 3) as a general issue with the GPS technology, location information recorded by GPS devices in this study had an error of ± 10 m (i.e. the precision measured by the 95% quantile was < 10 m) which required a data matching procedure to locate the GPS observations on the road section studied.

To filter out these noises and errors and to prepare the data for the analysis, different steps were applied:

- ***Step 1*** **Remove outliers and hardware failure errors** Outliers were removed by applying a threshold of maximum speed (180 km/h). Errors generated from hardware failure were simply removed from the database.
- ***Step 2*** **Remove unrelated data from other roads** A large proportion of the data were collected from other roads in Shanghai and should be removed. To remove the unrelated data, a distance threshold ($R_{threshold}$) was introduced, which is referred as the maximum distance that an effective GPS record can have from the defined road segments. The error in the precision of the GPS technology, offsets from the segmentation procedure, and road width were considered. The maximum distance can be calculated as

$$R_{threshold} = D_{error} + \frac{1}{2}D_{width} + D_{offset} \tag{1}$$

where, $D_{error}$ is the error in precision, $D_{width}$ is the road width, $D_{offset}$ is the maximum offset from the segmentation. In this study, $D_{error} = 10\,m$, $D_{width} = 40\,m$, and $D_{offset} = 10\,m$, therefore $R_{threshold} = 40\,m$. Then, GPS



*a. Segmentation illustrated on the road network map*

*b. Distribution of length of segments*

*c. Offset by representing arced segments using straight lines*

**Fig. 3.** Segmentation of the Middle Ring Expressway.

records that fall within a distance of $R_{threshold}$ to the defined road segments are considered as being on the expressway section, as presented in Fig. 4. The GPS records which do not fall within that distance were removed.

- ***Step 3*****Project the floating car data on the road segments** When matching the floating car data with the segmentation, the records were projected (or in another word, relocated) on the defined segments according to their GPS coordinates. As shown Fig. 4, records were projected on the perpendicular foot from their GPS position to the closest road segment. Locations in the database were updated accordingly, except for those falls in the part of the exterior angle of the node range, defined by the perpendicular from the node to the boundary lines (the light grey area in the figure). Records with their positions falling in the part of the exterior angle were projected on the nodes.
- ***Step 4*****Filter based on vehicle direction** Vehicles traveling on the inner and outer side of the ring road drive in a different direction, they hence they are in a different driving environment. This study focuses on crashes on the outer side of the middle ring expressway where vehicles move in the anticlockwise direction. The direction of the floating car can be determined by its heading direction and the direction of the road segment where vehicles on that segment move towards (anticlockwise in this case). The floating car was driving in an anticlockwise direction if the angle between the heading direction of the vehicle and the direction of the road segment fell in the

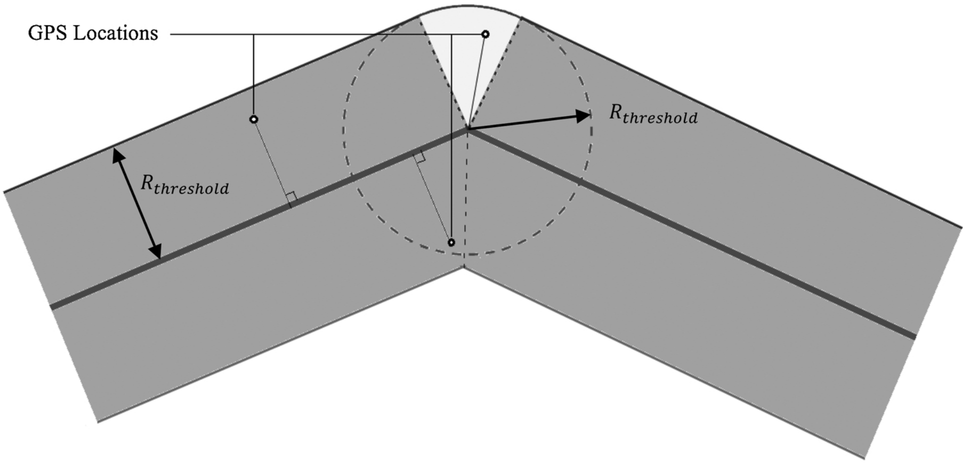


**Fig. 4.** Location projection on road segments.

range of [0°, 45°), whereas it was driving in a clockwise direction if the angle fell in the range of (135°, 180°]. Records with the vehicle moving in a clockwise direction were removed. A small portion of the records had the angle fallen within the range of (45° to 135°) and their direction were considered as unknown. They were most likely collected from the floating cars driving on the other road(s) at the grade separations along the expressway. These records were also removed from the database.

After these steps, the floating car data were filtered and matched on the defined segments for the safety analysis purpose in the study.

### *2.5. Matching the crash data and the floating Car data*

After floating car data being filtered and matched to the segmentation, the last step before conducting the safety analysis was to match the crash data and the related floating car trajectory data. As the study mainly focused on traffic moving anticlockwise on the outer side of the expressway, crash records were further filtered. Records for crashes that happened on the inner side of the expressway were removed.

The study used the concept of floating car accompanied (the floating car that occurred within 80 m upstream and 80 m downstream from the location of the crash when it occurred) which will be explained in the methodology part. Trajectories of floating cars accompanied were mainly used in predicting crashes in this study. According to the timestamps in the records, records of the crashes and their floating car accompanied are matched. Among these crashes, 118 had at least one floating car accompanied. Fig. 5 shows the histogram of the number of floating cars accompanied for the 118 crashes. Both crash and non-crash events are needed in crash prediction, where non-crash events are any traffic situations where no crash is observed (in a given period of time). 472 non-crash events were randomly selected in the same two-month observation period as the crash events. Data from the floating cars accompanied were also prepared for these non-crash events. The analysis was conducted using the data for the 118 crashes and the 472 non-crash events, and trajectory data of the floating car accompanied for these crashes and non-crash events.

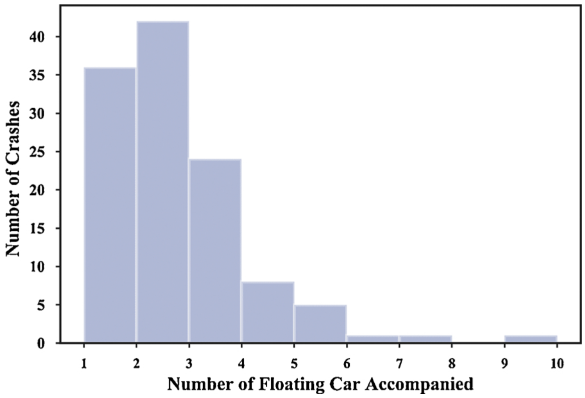


**Fig. 5.** Histogram of the number of floating car accompanied.

## 3. Methodology

The methodology section describes about: 1) the general concepts of traffic platoon, and floating car in platoons; 2) variables of traffic flow characteristics generated from floating car data; and 3) modeling methods used in the study.

### *3.1. Traffic platoon and floating Car in platoons*

#### *3.1.1. Traffic platoon and crashes*

A traffic platoon is a group of cars that travel together in approximately the same pattern (Li, 2017). The introduction of traffic platoon provides a new insight into explaining traffic crash outcomes on the corridor of the expressway. The traffic platoon is always in stable status where cars in the platoon keep their speeds similar, unless at least one of the cars within the platoon conducts an inconsistent maneuver, or an external car merges into the platoon – with such “turbulence” in the platoon, a crash can happen.

#### *3.1.2. Floating Car in platoon and floating Car Accompanied*

Floating cars equipped with location-based devices such as GPS devices are distributed throughout the road network, moving together with other vehicles in different traffic platoons. By collecting trajectory data from these cars, we are able to get the traffic information of different road platoons at different time and under different road environments.

In the case of a traffic event, a floating car accompanied is the floating car that travels within the traffic platoon where the event belongs to. By defining this, as presented in Fig. 6, floating cars accompanied are classified based on three different situations: 1) where it has always been in the traffic platoon till the occurrence of the event, as represented by car FC1; 2) where it has been in the platoon but exit before the event, as represented by car FC2; and 3) where the car just join the platoon at the occurrence of the event, as represented by car FC3. With the information from the floating cars accompanied, we are able to get the information on traffic characteristics of the traffic platoon where the event occurs or, in other words, we get traffic condition information of where crashes happen. In the study, floating cars

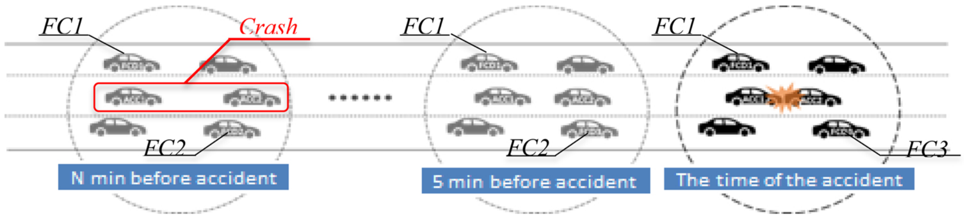


**Fig. 6.** Different situations of the floating vehicle accompanied.

accompanied were identified as those occurred within the crash area (or the area for the non-crash event), which is the area within 80 m upstream and 80 m downstream from where the crash happens.

### 3.1.3. Time range for data collection

Many previous studies have suggested that the traffic status in the period of 5–15 min. before the crash has a significant impact on the crash occurrence (Abdel-Aty and Pemmanaboina, 2006; Jung et al., 2010; Abdel-Aty et al., 2012; Pirdavani et al., 2015). This study divides the time range into two, the period during 5–10 min. before the crash, and the period during 10–15 min. before the crash. Traffic flow characteristics within the two periods will be extracted from the floating car trajectory data and their impact on crash are investigated.

## 3.2. Variables of traffic flow characteristics from floating Car data

At the occurrence of a crash, usually several floating cars accompanied are involved. With the trajectory data collected from the floating cars accompanied, traffic information of the platoon where the crash occurs can be extracted. Traffic information extracted from the trajectory data includes the speed of the platoon, the speed difference ratio, the average speed difference, the average lateral speed, the number of speed mutations, and the level of the speed mutation. Both these variables during the period 5–10 min. before the crash and during the period 10–15 min. before the crash are extracted.

### 3.2.1. Definitions of variables

*3.2.1.1. Average speed of the platoon (AS).* The speed of the platoon is defined as the estimated average speed of the traffic at the occurrence of the crash. It is calculated as the mean speed of floating cars in the same platoon, presented as:

$$S = \frac{\sum_{i=1}^{n} V_i}{n} \tag{2}$$

where, $V_i$ is the speed of the $i$-th floating car, $n$ is the number of floating cars companied. The average speed of the platoon is the mean speed of the platoon within a certain range of time, $AS = mean(S)$.

*3.2.1.2. Speed difference ratio (SDR).* Velocity discrepancy has been proved to be significantly related to crash risk (Ahmed and Abdel-Aty, 2012). SDR is used to describe vehicle speed discrepancy. It is the quotient of the standard deviation over the average speed (Ahmed and Abdel-Aty, 2012),

$$SDR_i = \frac{\sigma_i}{\bar{v}_i} \tag{3}$$

where, $\sigma_i$ is standard deviation of $i$-th floating car's speed within the crash area, $\bar{v}_i$ is the average speed of the $i$-th floating car.

*3.2.1.3. Average speed difference.* Sometimes SDR is not enough to present the speed feature of the floating car. For instance, Fig. 7 shows the speeds of two vehicles from the trajectory database. In the figure, the two vehicles had a same average speed (87 km/h) and a same standard deviation (11.2 km/h), and hence a same SDR (12.85%), while patterns of their speeds were evidently different (the speed of floating car A was more vibrated). Therefore, the average speed difference (ASD) is introduced in the study. ASD is the average of the speed change of the floating car, presented as:

$$ASD_i = \frac{\sum_{t=2}^{N} |v_{it} - v_{i(t-1)}|}{N-1} \tag{4}$$

where, $v_{it}$ is the instantaneous speed of the $i$-th floating car at its $t$-th record, $t \in N$, $N$ is the number of records for this floating car from the data.

*3.2.1.4. Average lateral speed (ALS).* Lane change maneuvers also affect the safety of cars in the platoon. The lateral speed is used to present the lane change behavior conducted by the floating cars accompanied. ALS is used to describe the frequency of lane change maneuvers made by each individual car in the platoon. The ALS can be calculated as:

$$ALS_i = \frac{\sum_{t=2}^{N} \left| \frac{D_{it} - D_{i(t-1)}}{\Delta t} \right|}{N-1} \tag{5}$$

where, $D_{it}$ is the distance from the car to the center line of the road section for the $i$-th car at its $t$-th record during the investigated time range.

*3.2.1.5. Number of speed mutations (NoSM).* Cases where the speed of the car changed sharply with the acceleration rate (absolute value) over $0.3\ \mathrm{m/s^2}$ are considered as speed mutations. This threshold is determined based on the $85^{\mathrm{th}}$ percentile rate from the filtered database, as presented in Table 2. Acceleration rates for the floating cars are calculated based on speed information extracted from the adjacent points of the car trajectory. The total number of speed mutations for the floating cars accompanied in a crash during the investigated time range can be then identified and investigated.

*3.2.1.6. Level of speed mutation (LoSM).* The level of speed mutation is represented by the average acceleration rate of the floating cars within the platoon during the investigated time range.

### 3.2.2. List of variables

Information of the variables within the two observation periods (the period from 5 to 10 min before the crash, and the period from 10 to 15 min before the crash) were extracted for the 118 crashes and 472 non-crash events. Details of the 12 variables are given in Table 3. Descriptive statistics of the variables for the crash and non-crash events are provided in Table 4.

## 3.3. Modeling methods

After data were prepared, binary logistic regression and SVM models were applied. This section briefly introduces the model applications in the study. For details of the models, one can refer to (Rodríguez, 2007; Ma and Guo, 2014).

### 3.3.1. Binary logistic regression model

The potential outcome of a crash can be evaluated as a binary problem. The binary logistic regression model is introduced to predict the relationship between the variables of the traffic condition, and the potential outcome of a crash (a non-crash event as 0, and a crash as 1). The model can be presented as:

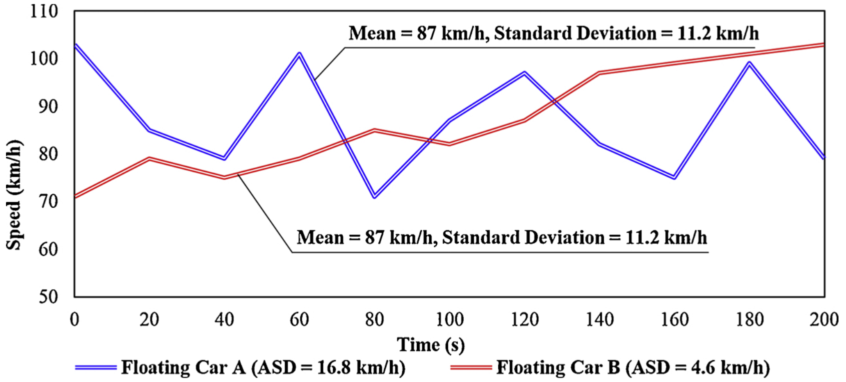


**Fig. 7.** Sample of different speed feature with a same SDR.

**Table 2**
Quotient Value of Acceleration Rates.

| Statistic | Mean | Variance | Quotient Value | | | | |
|---|---|---|---|---|---|---|---|
| | | | *25th* | *50th* | *75th* | *85th* | *95th* |
| Speed change rate ($m/s^2$) | 0.138 | 0.035 | 0.018 | 0.069 | 0.194 | 0.292 | 0.486 |

**Table 3**
Variables Involved.

| Symbol | Time Ranges | Explanation of the Variable |
|---|---|---|
| *AS_1* | 1st period: 10 ~ 15 min before accident | Average Speed |
| *SDR_1* | | Speed Difference Ratio |
| *ASD_1* | | Average Speed Difference |
| *ALS_1* | | Average Lateral Speed |
| *NoSM_1* | | Number of Speed Mutation |
| *LoSM_1* | | Level of Speed Mutation |
| *AS_2* | 2nd period: 5 ~ 10 min before accident | Average Speed |
| *SDR_2* | | Speed Difference Ratio |
| *ASD_2* | | Average Speed Difference |
| *ALS_2* | | Average Lateral Speed |
| *NoSM_2* | | Number of Speed Mutation |
| *LoSM_2* | | Level of Speed Mutation |

**Table 4**
Statistics of Variables.

| Variable | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| ***Crashes (Obs. = 118)*** | | | | |
| *AS_1 (km/h)* | 7.20 | 5.04 | 0.08 | 24.53 |
| *SDR_1* | 12.94 | 6.56 | 0.63 | 36.28 |
| *ASD_1 (km/h)* | 9.90 | 4.75 | 0.63 | 22.40 |
| *ALS_1 (km/h)* | 0.42 | 0.36 | 0.08 | 2.00 |
| *NoSM_1* | 1.47 | 1.80 | 0.00 | 10.00 |
| *LoSM_1* ($m/s^2$) | 0.67 | 0.79 | 0.00 | 3.84 |
| *AS_2 (km/h)* | 8.30 | 4.92 | 0.16 | 21.03 |
| *SDR_2* | 14.94 | 6.87 | 1.05 | 41.65 |
| *ASD_2 (km/h)* | 11.14 | 5.64 | 1.00 | 36.28 |
| *ALS_2 (km/h)* | 0.43 | 0.35 | 0.04 | 2.14 |
| *NoSM_2* | 1.25 | 1.35 | 0.00 | 5.00 |
| *LoSM_2* ($m/s^2$) | 0.68 | 0.77 | 0.00 | 3.90 |
| ***Non-crash Events (Obs. = 472)*** | | | | |
| *AS_1 (km/h)* | 7.64 | 5.49 | 0.01 | 34.04 |
| *SDR_1* | 12.82 | 6.41 | 0.41 | 41.23 |
| *ASD_1 (km/h)* | 9.46 | 4.64 | 0.95 | 34.47 |
| *ALS_1 (km/h)* | 0.48 | 0.44 | 0.04 | 3.28 |
| *NoSM_1* | 1.36 | 1.50 | 0.00 | 8.00 |
| *LoSM_1* ($m/s^2$) | 0.72 | 0.98 | 0.00 | 8.31 |
| *AS_2 (km/h)* | 6.81 | 5.34 | 0.04 | 34.04 |
| *SDR_2* | 11.77 | 5.67 | 0.67 | 26.73 |
| *ASD_2 (km/h)* | 8.69 | 4.46 | 0.67 | 24.14 |
| *ALS_2 (km/h)* | 0.43 | 0.33 | 0.06 | 2.18 |
| *NoSM_2* | 0.94 | 1.21 | 0.00 | 6.00 |
| *LoSM_2* ($m/s^2$) | 0.45 | 0.60 | 0.00 | 2.94 |

$$E(y) = f(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) \tag{6}$$

where $y$ is the crash outcome, $y \in \{0,1\}$, $E(y)$ is the probability of having a crash outcome of $y$, $\beta$ is the vector of the regression coefficient generated from modeling, $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$, $x$ is the vector is the variables, $x = (x_1, x_2, \ldots, x_n)$, $n$ is the number of variables. Note that, the paper investigated 12 variables as presented in Table 3, while those tested to be statistically insignificant in the modeling will be removed.

To simplify the model, crash and non-crash probability can be also described using the following equation:

$$P_{y=1} = \pi = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}} \tag{7}$$

$$P_{y=0} = 1 - \pi = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}} \tag{8}$$

where $\pi$ is the probability of resulting in a crash. Then, the probability function can be further simplified using $\pi$ :

$$P(y) = \pi^y (1 - \pi)^{1-y},\ y = 0,1 \tag{9}$$

In the binary logistic regression model, $\beta$, the vector of regression coefficient for the variables, is calculated using the maximum likelihood estimation which, in the case with $n$ samples, can be expressed as:

$$L = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{n6}) - \ln(1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{n6}})\} \tag{10}$$

To validate the impact of an independent variable on dependent variables, the Odds Ratio ($OR$) is used in the study:

$$OR_i = e^{\beta_i} \tag{11}$$

In statistics, $OR_i = 1$ indicates that the investigated variable does not affect the potential outcome of a crash; $OR_i > 1$ shows that the variable has a positive impact in increasing the chance of a crash; $OR_i < 1$, shows that the variable decreases the chance of a crash.

### 3.3.2. SVM algorithms

The accuracy of using traditional statistical analysis methods greatly rely on the size of the data. However, in many situations, data can be limited. The issue of small sampling has been always a problem in using crash data. In statistics, great efforts have been spent to overcome this issue. Among the many solutions, the support vector machine has been promising (Guo et al., 2010). The standard SVM is a supervised learning algorithm that works well in modeling binary classification problems (Stanevski and Tsvetkov, 2005). This makes it promising in crash prediction which always has binary outputs (either a crash, or a non-crash event).

The main parts in using the SVM method include the selection of the kernel function and the optimization of the related kernel parameters (Chang and Lin, 2007). To build a crash propensity prediction model using the SVM, the study applied the most widely used LIBSVM toolbox (Chang and Lin, 2007). The LIBSVM is a powerful tool that can solve both classification problems (using C-SVC, υ-SVC) and regression problems (using ε-SVR and υ-SVR). It provides four commonly used kernel functions including the linear, the polynomial, the radial basis and the sigmoid functions. It also allows customization of the kernel function matrix by the users. Meanwhile, the LIBSVM provides three parameters optimization methods: Grid Research, Genetic Algorithm and Particle Swarm Optimization (Chang and Lin, 2007).

For the binary classification case of crash prediction, the C-SVC classification model in the LIBSVM toolbox is chosen in the study. The independent variables, which are the inputs of the SVM, can be presented as:

$$x_j \in R, \ (i = 1,\ 2, \ldots, m) \tag{12}$$

where $R$ is the list of independent variables for samples in the training test, $m$ is the total number of samples in the training set. The dependent variable, i.e. the outcome of a potential crash, can be then presented as:

$$y_j \in \{1, -1\}, \ \ (1,\ \text{a crash};\ -1,\ \text{no crash}) \tag{13}$$

To determine the optimal interface in the binary classification case, the C-SVC model is applied, described as:

$$\min_{\omega,\mathrm{b},\xi} \frac{1}{2}\omega^T\omega + C\sum_{j=1}^{n}\xi_j \tag{14}$$

$$\textit{subject to}\ y_j(\omega^T\varnothing(x_j) + b) \geq 1 - \xi_j;\ \xi_j \geq 0,\ j = (1{,}2,\ \cdots, m) \tag{15}$$

where $\varnothing(x_j)$ is the mapping function, $\omega$, $\xi_j$ and $b$ are the parameters generated from modeling, and $C$ represents the penalty variable. More specifically, $\omega$ is a vector representing the weights of the variables in the SVM.

The kernel function can be written in the form of the mapping:

$$\mathrm{K}(x_j,\ x_k) = \varnothing(x_j)^T\varnothing(x_k);\ j,\ k = (1{,}2,\ \cdots, m) \tag{16}$$

The radial basis function (RBF) is selected as the kernel function in the study. The decision function of C-SVC model can then be given as:

$$f(x) = sgn(\omega^T\varnothing(x_j) + b) = sgn(\sum_{j=1}^{n} y_j\alpha_j e^{-\gamma||x_j - x||^2} + b) \tag{17}$$

where, $x_j$, $y_j$, $\alpha_j$, $b$ are the optimal parameters of the model. The performance of the model in classification are determined by the peripheral parameters $C$ and $\gamma$. To find the best-fit (C, γ) pairs in order to train the C-SVC model for identifying crashes and non-crash events, three optimization algorithms from LIBSVM toolbox were applied, including the Grid Research Algorithm (Grid), the Particle Swarm Optimization Algorithm (PSO), and the Genetic Algorithm (GA) (Chang and Lin, 2007; Zhong and He, 2012).

This paper uses the three optimization methods provided by the LIBSVM toolbox to search the best (C, $\gamma$) combinations and construct the C-SVC classification model, and compares the best combination of parameters and model classification accuracy.

## 4. Results

In training and testing the models, 70% of the 118 crashes and 472 non-crash events are randomly selected for training (the binary logistic regression model) and cross-validation (the SVM), while the rest 30% are used as the test set to assess the modeling performance. Results of the models and their performance in prediction are discussed in this section.

**Table 5**
Outputs from LR Test.

| | | $\chi^2$ | df | Sig. |
|---|---|---|---|---|
| **Step 1** | **Step** | 45.758 | 12 | 0.000 |
| | **Block** | 45.758 | 12 | 0.000 |
| | **Model** | 45.758 | 12 | 0.000 |
| … | **…** | … | … | … |
| **Step 9** | **Step*** | −2.322 | 1 | 0.128 |
| | **Block** | 37.403 | 4 | 0.000 |
| | **Model** | 37.403 | 4 | 0.000 |

Note: * $\chi^2$ for Step ($\chi^2_{\_Step}$) is the difference in $\chi^2$ between the model from the current step, and the previous one. $\chi^2_{\_Step} < 0$ represents a better goodness-of-fit for the current model.

### 4.1. Binary logistic regression model

The likelihood ratio test (LR test) is conducted to build the binary logistic regression model with the best goodness of fit. The outputs from the LR test are provided in Table 5. In the LR test, all the variables are included in the model in the beginning (as the initial step). Then, the least related variable is excluded, and the model is retrained in each subsequent step. The retrained model and the model from the previous step are compared, based on the difference in $\chi^2$. The iteration ends when the model with lowest $\chi^2$ occurs. As given in Table 5, the model from Step 9 has the lowest $\chi^2$ value (37.403) in this study, and is therefore chosen to predict crashes.

Based on the LR test, the binary logistic regression model is built. The results from modeling are provided in Table 6. From the results, variables including *Speed Deference Ratio* (SDR_2) and *Average Speed Difference* (ASD_2) in the second period have a significant impact on the crash occurrence. Besides, OR values of SDR_2 and ASD_2 are greater than 1, which indicates that the chance of a crash increases with these two variables. *Number of Speed Mutations* (NoSM_2) and *Level of Speed Mutation* (LoSM_2) are also found to be significant. The OR value of LoSM is 5.804 which indicates that LoSM increases the chance of crashes. Surprisingly, NoSM is found to reduce the chance of crashes. This is probably due to its high correlation with LoSM – LoSM increases when a higher number of speed mutations are observed. However, as they describe two dimensions of the speed mutation, they are both kept in the prediction model.

After modeling, the model is validated using the test group. Since crash prediction is a binary classification problem, the confusion matrix is introduced. Metrics from the confusion matrix are used to validate the performance of the model in crash prediction. Results are presented in Table 7. From the results, the prediction performance of the model works reasonably, with a sensitivity of 57.1%, a specificity of 60.6% and an overall accuracy of around 60% for the test group. The value of area under the curve (AUC) (from 0 to 1, where 0 means totally fail, 1 stands for the perfect modeling outcome), which is used commonly in machine learning techniques for presenting the prediction accuracy of the trained model (Turner, 2013), is also applied. The trained model is able to reach an AUC of 0.624.

### 4.2. SVM algorithms

Three optimization algorithms including the Grid, PSO and GA algorithms from LIBSVM toolbox are applied to check the best-fit (C, γ) pairs. The validation process, called five-fold cross-validation, is conducted using the LIBSVM toolbox to adjust the (C, γ) parameter pairs in order to tune the performance of the training. In the five-fold cross-validation process, the dataset is divided into five groups, with 4 for training and the rest one for validation. The cross-validation accuracy (CVAccuracy) is used to check the level-of-fit of the trained model. The cross-validation process is conducted for each of the (C, γ) pairs with different optimization algorithms.

**Table 6**
Modeling Results of the Binary Logistic Regression Model.

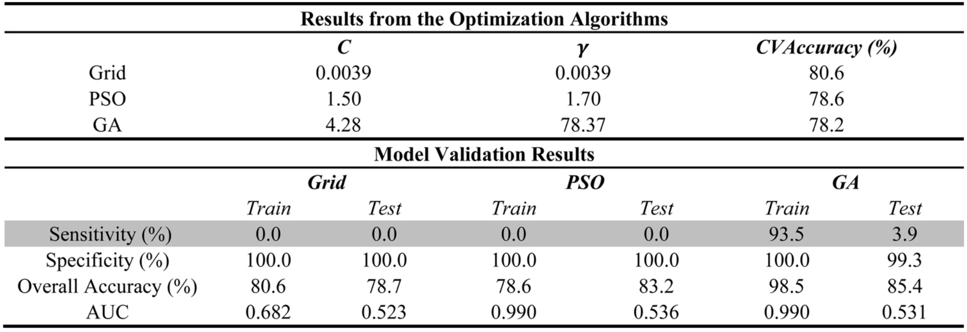
| Modeling Results | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Variable* | *B* | *S.E.* | *Wald* | *df* | *Sig.* | *OR* | *95% C.I. of OR* | |
| | | | | | | | *Lower limit* | *Upper limit* |
| SDR_2 | 0.056 | 0.024 | 5.363 | 1 | 0.021 | 1.057 | 1.009 | 1.108 |
| ASD_2 | 0.073 | 0.031 | 5.674 | 1 | 0.017 | 1.076 | 1.013 | 1.143 |
| NoSM_2 | −0.824 | 0.290 | 8.070 | 1 | 0.005 | 0.438 | 0.248 | 0.774 |
| LoSM_2 | 1.758 | 0.523 | 11.302 | 1 | 0.001 | 5.804 | 2.082 | 16.179 |
| constant | −2.922 | 0.355 | 67.666 | 1 | 0.000 | 0.054 | – | – |

Note: B is the mean of the regression coefficients, $\beta_0$ and $\beta$, S.E. is the standard deviation, df represents the degree of freedom, Wald is the results from Wald statistical test $(\frac{B}{S.E.})^2$, and Sig. is the significance from the Wald test.

**Table 7**
Results in Prediction of the Binary Logistic Regression Model.

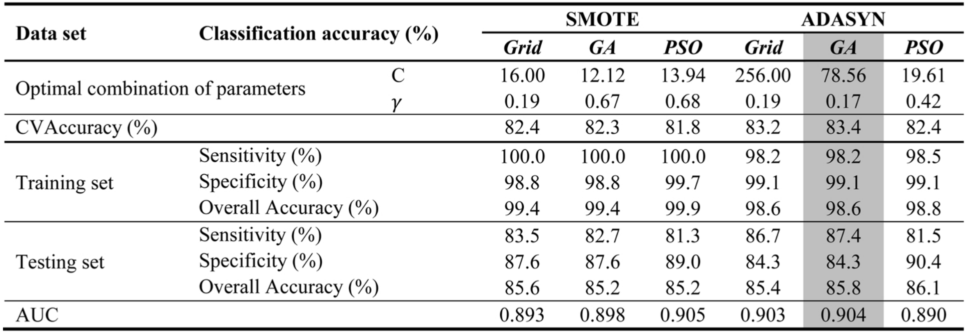
| | | Training Set | | | Testing Set | | |
|---|---|---|---|---|---|---|---|
| | | Predicted Condition | | | Predicted Condition | | |
| | | *Crash* | *Non-crash* | | *Crash* | *Non-crash* | |
| True Condition | *Crash* | 57 | 26 | ***Sensitivity = 68.7%*** | 20 | 15 | ***Sensitivity = 57.1%*** |
| | *Non-crash* | 130 | 200 | ***Specificity = 60.6%*** | 56 | 86 | ***Specificity = 60.6%*** |
| | | ***Overall Accuracy = 62.4%*** | | | ***Overall Accuracy = 59.9%*** | | |

**Table 8**
Results of Optimization Algorithm and Model Validation.



| **Results from the Optimization Algorithms** | | | | | | |
|---|---|---|---|---|---|---|
| | ***C*** | | ***γ*** | | ***CVAccuracy (%)*** | |
| Grid | 0.0039 | | 0.0039 | | 80.6 | |
| PSO | 1.50 | | 1.70 | | 78.6 | |
| GA | 4.28 | | 78.37 | | 78.2 | |
| **Model Validation Results** | | | | | | |
| | ***Grid*** | | ***PSO*** | | ***GA*** | |
| | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* |
| Sensitivity (%) | 0.0 | 0.0 | 0.0 | 0.0 | 93.5 | 3.9 |
| Specificity (%) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 99.3 |
| Overall Accuracy (%) | 80.6 | 78.7 | 78.6 | 83.2 | 98.5 | 85.4 |
| AUC | 0.682 | 0.523 | 0.990 | 0.536 | 0.990 | 0.531 |

**Table 9**
Modeling and Prediction Results with the Balanced Data.



| **Data set** | **Classification accuracy (%)** | | **SMOTE** | | | **ADASYN** | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***Grid*** | ***GA*** | ***PSO*** | ***Grid*** | ***GA*** | ***PSO*** |
| Optimal combination of parameters | | C | 16.00 | 12.12 | 13.94 | 256.00 | 78.56 | 19.61 |
| | | γ | 0.19 | 0.67 | 0.68 | 0.19 | 0.17 | 0.42 |
| CVAccuracy (%) | | | 82.4 | 82.3 | 81.8 | 83.2 | 83.4 | 82.4 |
| Training set | Sensitivity (%) | | 100.0 | 100.0 | 100.0 | 98.2 | 98.2 | 98.5 |
| | Specificity (%) | | 98.8 | 98.8 | 99.7 | 99.1 | 99.1 | 99.1 |
| | Overall Accuracy (%) | | 99.4 | 99.4 | 99.9 | 98.6 | 98.6 | 98.8 |
| Testing set | Sensitivity (%) | | 83.5 | 82.7 | 81.3 | 86.7 | 87.4 | 81.5 |
| | Specificity (%) | | 87.6 | 87.6 | 89.0 | 84.3 | 84.3 | 90.4 |
| | Overall Accuracy (%) | | 85.6 | 85.2 | 85.2 | 85.4 | 85.8 | 86.1 |
| AUC | | | 0.893 | 0.898 | 0.905 | 0.903 | 0.904 | 0.890 |

Results from the optimization algorithms are given in Table 8. Though the values for the best-fit (C, γ) pairs from different optimization algorithms vary greatly, CVAccuracy values from the validation process are high (around 80%) for all the pairs. However, oversampling issue exists due to the manual sampling of the crashes and non-crash events. This explains the low sensitivity values (highlighted in grey in Table 8) for different optimization algorithms.

To solve the issue of oversampling, two techniques are introduced, including the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2011) and the Adaptive Synthetic Sampling Approach

**Table 10**
Performance Comparison of Two Predictive Models.

| Performance Index | Binary logistic regression | SVM |
|---|---|---|
| Sensitivity (%) | 57.1 | 87.4 |
| Specificity (%) | 60.6 | 84.3 |
| Overall Accuracy (%) | 59.9 | 85.8 |
| AUC | 0.624 | 0.904 |

(ADASYN) (He et al., 2008). For details of the two techniques, one can refer to (Chawla et al., 2011; He et al., 2008). Results from the two different techniques are presented in Table 9. From the table, the performance of the two techniques are similar with their AUCs around 0.9 and the CVAccuracy values over 80%. This shows a good performance of using the SVM in predicting crashes. Among models generated from different combinations of data balancing techniques and optimization algorithms, the one with GA algorithm and the ADASYN approach (highlighted in grey in Table 9) is the best-fit model as it performs slightly better over the other ones with the highest CVAccuracy of 83.4% and the highest AUC of 0.904. Meanwhile, based on the confusion matrix, it reaches a sensitivity of 87.4 %, a specificity of 84.3% and an overall accuracy of 85.8%, which shows a quite balanced performance in predicting both crash and non-crash events.

### 4.3. Comparisons between models

As presented in Table 10, comparisons of the results from the two modeling methods are made. Four metrics are used for the comparison purpose including the sensitivity, the specificity, the overall accuracy, and the AUC. Key findings and discussions can be made between the applications of the two models:

- For crash prediction: All the performance metrics suggest that the SVM model performs observably better over the binary logistic regression model in predicting crashes.
- For variable explanation: Results from the binary logistic regression model presents the relationship between variables and the occurrence of a crash. The binary logistic regression helps identifying factors mainly contribute to crashes, on urban expressway in this paper. However, without explanatory variables, the SVM model failed to do so.

## 5. Conclusions

This paper proposes a methodology that uses floating car trajectory data and the machine learning approach to predict the occurrence of crashes on urban expressways. In the methodology, the traffic platoon is used in defining floating car accompanied. Trajectories of the floating car accompanied are then used to extract traffic information of the platoon where the crash vehicle belongs to at the crash occurrence. Different variables of the platoon are extracted from the floating car accompanied for prediction of crash propensity. Two modeling methods including a binary logistic regression model and a support vector machine model are introduced and compared to predict the occurrence of crashes. The methodology is tested based on the crash data and floating car data collected from Shanghai Middle Ring Expressway, one of the main rings in the city of Shanghai, China. Results from the two models are compared. Some key findings can be drawn:

- From results of the binary logistic regression model, traffic status of the period between 5 min. to 10 min. before the crash have a great impact on the occurrence crashes, while that of the period between 10 min. to 15 min. before the crash is not found to be significantly related to the occurrence of the crash.
- Among the variables, the impact of the variables related to short-term speed variance including the speed difference ratio, the average speed difference, and variables of speed mutations are found to be significant. This suggests stable driving speeds and speed changing maneuvers with reduced acceleration rate can help reduce crash risks.
- The outstanding performance of the SVM model supported by the indexes indicates that the SVM model works greatly for predicting crashes on urban expressways. Implementations of such models for monitoring real-time conditions are promising in helping to detect and prevent potential crashes from happening.
- Floating car trajectory data, with its advantage of being low cost and rich in information, and wide coverage, is proved to work as an excellent data source for collecting traffic information. It provides a tailor-made solution to collecting traffic data for safety analysis and crash prevention purposes.

In general, information collected from floating car accompanied works reasonably in predicting crashes on expressways. Meanwhile, both of the two models work well in predicting crashes. The comparison results, the SVM model greatly outperformed the binary logistic regression model in predicting crashes. However, the SVM, as many machine learning algorithms do, falls short in providing explanations of the relationship between variables from the built environment and traffic, and the occurrence of a crash. The approach relying on trajectories of the floating car may not be able to predict the occurrence of crashes at locations such as freeway ramps or toll plazas where behavior adaptations (acceleration, deceleration and merging attempts, and traffic interruptions) frequently happen. For future work, the models will be tested and improved with data from a large number of expressway environments. Crash prediction at expressway ramps should be further explored using different methods. Other urban road environments should also be investigated. Real-time crash prediction using live floating car data will also be tested.

## Acknowledgements

Funding for this project was provided in part by the National Natural Science Foundation of China (71871161). The authors wish to thank the former student, Shuaiyi Sun from the transportation research group of at Tongji University for his help in collecting and preparing the data.

## References

AASHTO, 2010. Highway Safety Manual, s.l. American Association of State Highway and Transportation Officials.
Abdel-Aty, M., Abdalla, F., 2004. Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data. Transportation Res. Record (1897), 106–115.
Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Safety Res. 36 (1), 97–108.
Abdel-Aty, M., Pemmanaboina, R., 2006. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. Ieee Trans. Intell. Transp. Syst. 167–174.
Abdel-Aty, M., et al., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transportation Res. Record (1897), 88–95.
Abdel-Aty, M., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C Emerg. Technol. 288–298.
Ahmed, M.M., Abdel-Aty, M., 2012. The viability of using automatic vehicle 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. Transportation Res. Record (2280), 60–67.
Brubacher, J., et al., 2018. Road Safety Impact of Increased Rural Highway Speed Limits in British Columbia, Canada. Sustainability 10 (3555).
Cafiso, S., D'Agostino, C., Persaud, B., 2018. Investigating the influence of segmentation in estimating safety performance functions for roadway sections. J. Traffic Transp. Eng. 5 (2), 129–136.
Transport Canada, 2017. Canadian Motor Vehicle Traffic Collision Statistics 2015, s.l.: s.n.Canadian Motor Vehicle Traffic Collision Statistics 2015, s.l.: s.n.
Cathey, F.W., Dailey, D., 2002. Transit vehicles as traffic probe sensors. Transportation Res. Record (1804), 579–584.

Chang, L.-Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci. 43 (8), 541–557.

Chang, C.-C., Lin, C.-J., 2007. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 1–27 pp. 27.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, P.W., 2011. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16 (1), 321–357.

Coifman, B., 2005. Freeway detector assessment: aggregate data from remote traffic microwave sensor. Transportation Res. Record(1917) pp. 150-136.

Dewulf, B., et al., 2015. Examining commuting patterns using floating car data and circular statistics: exploring the use of new methods and visualizations to study travel times. J. Transp. Geogr. 48, 41–51.

Diependaele, K., Riguelle, F., Temmerman, P., 2016. Speed behavior indicators based on floating car data: results of a pilot study in Belgium. Transp. Res. Procedia 14, 2074–2082.

Formosa, N., Quddus, M., Ison, S., 2019. Predicting real-time traffic conflicts using deep learning. In: the 98th Annual Meeting of the Transportation Research Board. Washington DC.

Fu, T., et al., 2017. Automatic traffic data collection under varying lighting and temperature conditions in multimodal environments: thermal versus visible spectrum video-based systems. J. Adv. Transp. 2017, 1–17.

Fu, T., Miranda-Moreno, L., Saunier, N., 2018. A novel framework to evaluate pedestrian safety at non-signalized locations. Accid. Anal. Prev. 111, 23–33.

Golob, T.F., Recker, W.W., 2003. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng. 129 (4), 342–353.

Guo, Y., Graber, A., McBurney, R.N., Balasubramanian, R., 2010. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics 11 (447), 19p.

Guo, Y., et al., 2019. A comparison between simulated and field-measured conflicts for safety assessment of signalized intersections in Australia. Transp. Res. Part C Emerg. Technol. 101, 96–110.

Hassan, H.M., Abdel-Aty, M., 2013. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. J. Safety Res. 45, 29–36.

He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks. Hong Kong, China.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accid. Anal. Prev. 213–224.

Klein, L.A., Mills, M.K., Gibson, D.R., 2006. Traffic Detector Handbook: Volume II. s.l.:s.n.

Lee, C., Saccomanno, F., Lai, X., 2002. Analysis of crash precursors on instrumented freeways. Transportation Res. Record (1784), 1–8.

Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. Transportation Res. Record (1840), 67–77.

Lesani, A., Miranda-Moreno, L., Fu, T., Romancyshyn, T., 2015. Development and testing of an ultrasonic-based pedestrian counting system. In: the 94th Annual Meeting of the Transportation Research Board. Washington DC.

Li, B., 2017. Stochastic modeling for vehicle platoons (I): dynamic grouping behavior and online platoon recognition. Transp. Res. Part B Methodol. 95, 364–377.

Ma, Y., Guo, G., 2014. Support Vector Machines Applications. s.l.:s.n.Support Vector Machines Applications. s.l.:s.n.

Ogle, J.H., Alluri, P., Sarasua, W., 2011. MMUCC And MIRE: the Role of Segmentation in Safety Analysis. Transportation Research Board, Washington DC.

Oh, J.-S., Oh, C., Ritchie, S.G., 2005. Real-time estimation of freeway accident likelihood for safety enhancement. J. Transp. Eng. 131 (5), 358–363.

Pan, G., Fu, L., Thakali, L., 2017. Development of a global road safety performance function using deep neural networks. Int. J. Transp. Sci. Technol. 6 (3), 159–173.

Pirdavani, A., et al., 2015. Application of a rule-based approach in real-time crash risk prediction model development using loop detector data. Traffic Inj. Prev. 2015 (16), 786–791.

PSM China, 2017. Compilation of Road Traffic Accident Statistical Data of the People's Republic of China, 2016. Traffic Management Bureau of the Public Security Ministry of the People's Republic of China, Beijing, China.

Qin, X., Wellner, A., 2012. Segment Length Impact on Highway Safety Screening Analysis. Washington DC, s.n.. .

Rahmani, M., Jenelius, E., Koutsopoulos, H., 2014. Floating car and camera data fusion for non-parametric route travel time estimation. Procedia Comput. Sci. 37, 390–395.

Rahmani, M., Jenelius, E., Koutsopoulos, H., 2015. Non-parametric estimation of route travel time distributions from low-frequency floating car data. Transp. Res. Part C Emerg. Technol. 58 (Part B), 343–362.

Rodríguez, G., 2007. Lecture Notes on Generalized Linear Models. [Online] Available at:. http://data.princeton.edu/wws509/notes/.

Schlögl, M., Stütz, R., Laaha, G., Melcher, M., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Accid. Anal. Prev. 134–149.

Stanevski, N., Tsvetkov, D., 2005. Using support vector machine as a binary classifier. s.l. International Conference on Computer Systems and Technologies.

Sun, J., Sun, J., 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transp. Res. Part C Emerg. Technol. 54, 176–186.

Sun, J., Sun, J., 2016. Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model. Iet Intell. Transp. Syst. 10 (5), 331–337.

Theofilatos, A., Chen, C., Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Transportation Res. Record 1–10.

Turner, R.J., 2013. Area under the curve (AUC). Encyclopedia Behavioral Medicine 125–127.

US DOT, 2017. Transportation Fatalities by Mode, s.l.: Bureau of Transportation Statistics. US Department of Transportation.

Vanlommel, M., et al., 2015. An evaluation of section control based on floating car data. Transp. Res. Part C Emerg. Technol. 58 (Part C), 617–627.

Wang, Z., 2012. A comparison of floating car vs. Loop detector estimated freeway travel time delay. Int. J. Transp. Sci. Technol. 1 (2), 147–169.

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., Kong, Y., Fu, T., 2019. Expressway crash risk prediction using back propagation neural network: a brief investigation on safety resilience. Accid. Anal. Prev. 124, 180–192.

Wu, Y., et al., 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.

Wu, Y., et al., 2019. Developing a crash warning system for the bike lane area at intersections with connected vehicle technology. Transportation Res. Record 1–12.

Xing, L., et al., 2019. Examining traffic conflicts of up stream toll plaza area using vehicles' trajectory data. Accid. Anal. Prev. 125, 174–187.

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., Quddus, M., Wang, X., Yang, K., 2018. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. Accid. Anal. Prev. 304–310.

Zabat, M., Stabile, N., Frascaroli, S., Browand, F., 1995. The Aerodynamic Performance of Platoons: A Final Report. California Department of Transportation., Los Angeles, CA.

Zhang, Y., Zuo, X., Zhang, L., Chen, Z., 2011. Traffic congestion detection based on GPS floating-car data. Procedia Eng. 15, 5541–5546.

Zhong, S., He, Z., 2012. Application of particle swarm optimization algorithm based on classification strategies to grid task scheduling. J. Softw. 7 (1), 118–124.

Zimmer, M., 2005. Surveillance, privacy and the ethics of vehicle safety communication technologies. Ethics Inf. Technol. 7 (4), 201–210.