Analysis and Prediction of Intersection Traffic Violations Using Automated Enforcement System Data

Yunxuan Li[a], Meng Li[a,*], Jinghui Yuan[b], Jian Lu[c], Mohamed Abdel-Aty[d]

[a] *Department of Civil Engineering, Tsinghua University, Beijing 100084, P.R. China*

[b] *National Transportation Research Center, Oak Ridge National Laboratory, Knoxville, Tennessee 37918*

[c] *School of Transportation, Southeast University, Nanjing, Jiangsu, 211189, China*

[d] *Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, Florida 32816-2450, United States*

Abstract

The automated enforcement system (AES) is an effective way of supplementing traditional traffic enforcement, and the traffic violation data from AES can also be effectively used for safety research. In this study, traffic violation data were used to analyze the influencing factors associated with traffic violations and to predict the probability of violations at intersections. The potential factors influencing violations include 24 independent factors related to time, space, traffic and weather. Results from a logistic model showed that the midday period, weekends, residential districts, collector roads, congested traffic conditions, high traffic flow, lower wind speed and low temperature would increase the probability of traffic violations. The probability of violations was predicted by the random forest algorithm, which was proven to be the best traffic violation prediction model among logistic regression, Gaussian naive Bayes, and support vector machine. Moreover, the proximity weighted synthetic oversampling technique (ProWSyn) method was applied to reduce the impact of the imbalance ratio (IR) and improve the model's prediction performance. The receiver operating characteristics (ROC) curves and Precision-Recall (PR) curves illustrated that the random forest algorithm using oversampling data had the best classifier prediction performance than undersampling data. The area under curve (AUC) and out-of-bag (OOB) error with IR=1 reached 0.914 and 0.0787, which showed the better performance of the random forest algorithm using ProWSyn in dealing with imbalanced traffic violation data.

***Keywords**:* Automated Enforcement System; Traffic Violation; Random Forest; Imbalance Ratio

1.Introduction

The committing of traffic violations is a worldwide road safety problem. Violations such as running red-light and speeding easily lead to crashes (Smith et al., 2000; Eccles et al., 2012; Porter et al., 2013; Jahangiri et al., 2016; Cheng et al., 2019). The traditional traffic law enforcement strategy is to increase the number of police officers and expand patrol areas. However, this strategy is unsafe, expensive, and requires large staffing resources for law enforcement agencies (Retting et al., 2004; Long et al., 2013; Goldenbeld et al., 2019). In recent years, an effective violation countermeasure to supplement traditional enforcement had become available, the automated enforcement system (AES) (Kerimov et al., 2017; Park et al., 2019; Li et al., 2020). Although some researchers and organizations have been very vocal in their opposition to AES (Fang et al., 2018; Balasubramanian et al., 2019), arguing that AES is unfair to drivers and unnecessary for improving road safety (Lund et al., 2009; Baratian-Ghorghi et al., 2016), scientific studies and examples show that traffic safety is greatly affected by the availability of AES in the street and road networks (Baratian-Ghorghi et al., 2016; Kerimov et al., 2017; Park et al., 2019). Since the obvious effects of AES on real traffic management, the system has been employed in Europe, the United States and China (Ahmed and Abdel-Aty 2015; Zhang et al., 2016).

The research advantage of AES is automatically detecting traffic violations 24 hours a day, accumulating a large number of historical violation records that enable researchers to conduct in-depth studies of traffic violation behavior. These violation data records combined with the time, space and environmental data to constitute a continuous data environment, which is helpful for researchers in analyzing and predicting the probability of traffic violations. Previous studies had shown that using the traffic violation data from AES could objectively and effectively evaluate traffic violation hotspots and coldspots from a spatiotemporal perspective (Li et al., 2020). The findings are useful for traffic managers and police officers in developing control strategies for different locations and types of hazards, such as at intersections. To advance research on the mechanisms of traffic violations, this study uses the logistic model to quantify the impact of various factors on the probability of violation occurrence at intersections. The second purpose of this study is to use the AES data to predict the probability of violations at intersections. Since quantity of non-violation data is more than violation data, an effective oversampling method is also used to improve the performance of the prediction model. Results from the violation factors analysis and prediction model would be beneficial for traffic management authorities to implement more effective countermeasures to prevent traffic violations, and further reduce the number of traffic crashes at the source.

The remainder of this paper is organized as follows: Section 2 presents a literature review; Section 3 describes the data in this study; Section 4 analyzes the traffic violation influencing factors; Section 5 establishes and analyzes the traffic violation prediction model, and Section 6 contains the final discussion and concludes this study.

2.Literature review

2.1 Traffic violations influencing factors

Studies conducted all over the world support a strong relationship between traffic violations and crashes (Jovanović et al., 2011; Zhang et al., 2013; Zhang et al., 2018;). In general, driving behavior is considered to be an important factor affecting traffic violations. Most of the existing literature focuses on analyzing or modeling driver behaviors on some specific traffic violations, such as red-light running or drunk driving (Jahangiri et al., 2016; Stringer 2018; Fei et al., 2020). These studies considered that drivers committing traffic violations usually exhibit aggressive driving behavior and consequently had a higher risk of being involved in serious or fatal traffic crashes. For example, Reason (1998) defined violations as ‘‘deliberate deviations from those practices deemed necessary to maintain the safe operation of a potentially hazardous system”. He concluded that traffic violations could be dealt with by trying to change subjects’ motives, attitudes, beliefs and norms, and by improving the overall safety culture. The link between drivers’ aggressive behavior and their personality and predisposition was supported by findings that those who more frequently violate the red light rule have three times driving violations than those who respect the red light (Goldenbeld et al., 2019). In addition, some personal characteristics were responsible for the occurrence of traffic violations. Gender was found to be an important factor affecting the probability of various types of violations because of a gender difference in risk perception. On one hand, female drivers were more prone to small driving errors and slips, especially in situations requiring increased attention and perception (Özkan, 2006). On the other hand, male drivers tend to drive in a riskier manner, and also to speed and drive while intoxicated more often than females (Oppenheim et al., 2016). Similarly, some studies had shown that young drivers more frequently violate traffic rules and were involved in traffic crashes more than older drivers (Mårdh, 2016; Ayuso et al., 2020). The reason may be due to the inexperience of young drivers as well as their higher level of excitement-seeking, greater perceived likelihood of an accident, and lower aversion to risk-taking (Adanu et al., 2020). Traffic violations also resulted from drowsiness and habits of distraction such as talking to passengers and using cellphones, which usually reduced drivers’ concentration and sensitivity to traffic signs and signals (Regan et al., 2011; Lipovac et al., 2017; Sullivan et al., 2020).

The recent literature generally focused on these subjective personal factors, whereas some potential traffic factors, such as violation time, roadway conditions, traffic flow, and other prevailing environmental factors, were rarely discussed (Zhang et al., 2018; Li et al., 2020). However, traffic violations, just as traffic crashes, were the results of interacting with these factors. Most of the potential factors consisted of several independent variables and the interaction between two or more of them. For example, violations such as stopping beyond the stop line and unlawful cut-ins were more likely to occur at intersections in the morning and evening rush hours, especially in the commercial district (Li et al., 2020). The morning and afternoon sunlight may reduced the color visibility of signal lights, and lead to red-light running violations (FHWA, 2007). In addition, the land use (e.g., residential district, commercial district) and road types (e.g., ramps, local road) might cause certain types of traffic violations to appear periodically (Champahom et al., 2020). In China, for example, parking violations were significantly common after school, and generally occurred every working day. Moreover, factors such as weather and time of day also affected drivers' behavior to a certain extent (Theofilatos et al., 2014; Xing et al., 2019). Severe weather, bright sun, dust, and debris reduce visibility, distract drivers and affected drivers' ability to observe signs, signals, and other traffic control devices promptly. Rain and snow could also make roads slippery and increase braking distance, further affecting drivers' behavior at intersections.

2.2 Traffic violations prediction analysis

Only a few studies established prediction models using traffic violation data (Amiruzzaman et al., 2018). Because traffic violations were similar to traffic crashes, some studies had used crash prediction methods to predict traffic violations. In crash prediction analysis, some statistical learning-based models such as Poisson Regression, Logistic Regression, and Bayesian Networks had been largely utilized. However, the disadvantage of these models was that the calculation process required a great deal of historical data, and the result was usually unsatisfying when treating features with a high number of categories. In recent years, studies using machine learning (ML)-based models such as support vector machines (SVM) (Arhin et al., 2020), artificial neural network (ANN) (Lee et al., 2018), random forest (RF) (Jiang et al., 2016) had often demonstrated satisfying results in crash prediction. The RF algorithm was one of the most successful general-purpose algorithms in modern times (Liaw et al., 2002; Dogru et al., 2018). It could reduce variance better than a single decision tree, and it was also robust regarding outliers and missing values. However, the RF algorithm still had some shortcomings: for example, it performed poorly for the classification of imbalanced data, failed to control the model during specific operations, and was sensitive to parameter adjustment and random data attempts (Ma et al., 2017).

Usually, there was an effective way to improve the RF algorithm: increase the accuracy of each classifier. Under the continuous data environment, the quantity of traffic operation data in the non-violation state was more than that in the violation state. For example, the monthly sample dataset of the normal state was usually 10-1,000 times that of the violation state. Although the number of traffic violations was much higher than the number of traffic crashes, the minority class (traffic violations) still had a higher classification error cost in the imbalanced data classification problem. There had been many attempts to solve imbalanced learning problems, such as various oversampling and undersampling methods (Jeong et al., 2018; Schlögl et al., 2019). In general, undersampling removed some majority class samples from an imbalanced data set with an aim to balance the distribution between the majority class and minority class samples, while oversampling does the opposite, that was generates synthetic minority class samples and added them to the dataset (Elassad et al., 2020). In comparing the oversampling and undersampling methods, oversampling was more useful than undersampling. Undersampling might lose the informative training instances from the majority class, especially if the number of instances was larger than the number of minority class instances. Oversampling had been shown to improve imbalanced data dramatically even for complex data sets (Yahaya et al., 2019, Yuan et al., 2019). The Synthetic Minority class Oversampling Technique (SMOTE) was one of the most influential data oversampling algorithms in machine learning and data mining (Chawla et al., 2002). However, there were some drawbacks to SMOTE. For example, the selection of a value for *k* was not informed by the nearest neighbor's selection. Consequently, it was impossible to completely reflect the distribution of original data because the artificial samples generated by the minor class samples at the edges might lead to problems such as repeatability and noisy, fuzzy boundaries between the positive and negative classes. Many extensions and alternatives had been proposed in recent years to improve SMOTE's performance under different scenarios (Kovács, 2019).

3. Data description

In comparison with the enforcement by on-site police officers, the AES is off-site and a more comprehensive enforcement method from the perspectives of violation types and service hours. Thus in recent years, the AES has been widely implemented in China, more in cities' central areas. In this study, the traffic violation data were derived from AES raw data, which were collected by the traffic administration of Wujiang from February to April in 2017, for a total of 85 days. During this period, 21,525 intersection violations were observed, such as violating traffic signs and markings, red-light running, speeding. As shown in Figure 1, 53 AESs are highly concentrated within the study area, which covering almost all signalized intersections except for those under construction. Table 1 shows the number of the different types of traffic violations detected by AES. The top three violation types, i.e. traffic sign or marking violation, stopping beyond the stop line and red-light running, account for 86% of all violations at intersections. Some of these violations led to crashes and/or traffic congestion. It should be noted that the traditional violation data are often taken from police reports, in which the recording time and locations are not quite accurate. The quality and availability of the violation data from AES, however, are more accurate and complete.

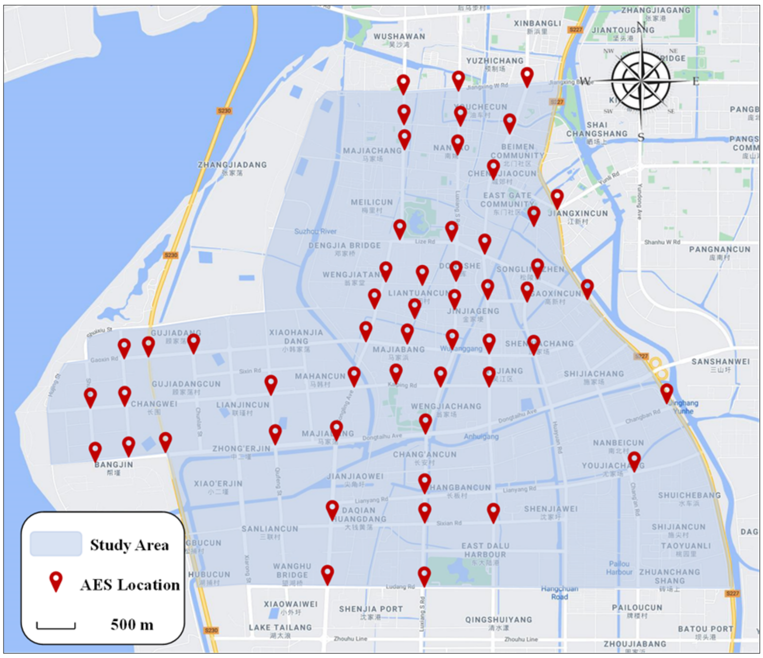


Figure 1 The locations of AES in Wujiang

Table 1 Violation type and number detected by AES

Violation Type	Violation Description	Violation Number
1	Traffic sign or marking violation	8890
2	Stopping beyond the stop line	5582
3	Red-light running	3947
4	Driving in bicycle lanes	728
5	Failure to yield to pedestrians	678
6	Unlawful cut-in	435
7	Wrong-way driving	408
8	Violation parking	321
9	Speeding	226
10	Seatbelt violation	112
11	Abuse of high beam lights	45
12	Failure to maintain lane	23

Table 2 Description of traffic violation influencing factors

Variables	Symbol	Violation	Non-violation

			N	%	*N*	%
			21525	9.9	194715	90.1
	Occurrence time					
Time factor	Nights	X_{10}	2794	1.3	87306	40.3
	Morning rush hour	X_{11}	3775	1.7	23255	10.7
	Midday	X_{12}	12679	5.9	59401	27.4
	Evening rush hour	X_{13}	2277	1.1	24753	11.4
	Occurrence day					
	Weekday	X_{20}	13934	6.4	131074	60.6
	Weekend	X_{21}	5580	2.6	45300	20.9
	Holiday	X_{22}	2011	0.9	18341	8.5
	Land use					
Space factor	Residential	X_{30}	10028	4.6	89048	41.1
	Commercial	X_{31}	2629	1.2	27927	12.9
	Administration	X_{32}	2323	1.1	17467	8.1
	Education	X_{33}	3661	1.7	32653	15.1
	Hospital	X_{34}	238	0.1	4582	2.1
	Recreational	X_{35}	2646	1.2	23038	10.6
	Road types					
	Ramps	X_{40}	588	0.3	6277	2.9
	Arterial road	X_{41}	3975	1.8	34121	15.8
	Collector road	X_{42}	16243	7.5	139861	64.6
	Local road	X_{43}	719	0.3	14456	6.7
	Traffic congestion					
Traffic factor	Free flow	X_{50}	21043	9.7	192985	89.2
	Slight congestion	X_{51}	122	0.1	825	0.4
	Severe congestion	X_{52}	360	0.2	905	0.4
			Minimum	Maximum	Average	Error
	Traffic flow(pcu/day)	X_6	325	54794	18863	10493
	10 min average wind (m/s)	X_7	0	5.9	1.918	0.933
Weather factor	Temperature (℃)	X_8	-1.8	31.9	14.858	6.862
	Rainfall (mm/h)	X_9	0	18.3	0.118	0.683

The AES can automatically detect traffic violations 24 hours a day, combing with the time, space and environmental data to constitute a continuous data environment. The previous study has proved that the optimal temporal search bandwidths of traffic violations were rounded to 30 minutes. Thus, the continuous data environment in this study can be divided into violation and non-violation within 30 minutes. Each dataset includes four major factor types: time factors, space factors, traffic factors, and weather factors. Together, these datasets include 24 independent factors such as violation occurrence time, violation occurrence day, and intersection surrounding land use, as shown in Table 2. Of the different types of factors, the time, traffic, and weather factors can be regarded as dynamic factors, while the space factor can be regarded as a static factor. For occurrence time, a day can be divided into four parts based on traffic flow, which includes morning rush hour (7:00-10:00), midday (10:00-17:00), evening rush hour (17:00-20:00), and night (20:00-7:00). Of the 85 collected occurrence days, 57 were weekdays, 20 were weekends, and 8 were holidays. Traffic violations occurring at a given AES might be correlated with the same land use and road types. Land use is based on the number and type of points of interest (POI) around AES. Each of the 53 intersections has four approaches, for which the land use is provided by Baidu Maps, a satellite mapping application. Land use for the total 212 approaches includes: 107 residential districts, 29 commercial districts, 18 Administration districts, 21 education districts, 20 hospital districts, and 17 recreational districts. Road type has always been the focus of AES safety monitoring, and, like land use, is assigned by intersection approach. The road types include: 11 ramps, 70 arterials, 112 collector roads, and 19 local roads. Traffic congestion data at intersections is also provided by Baidu Maps, and is divided into the free flow, slight congestion and severe congestion, with a congestion data update frequency of 30 minutes. The weather factor data is provided by air monitoring stations, which can offer 1-hour continuous data to capture the changes of weather conditions over a short time period. As shown in Table 2, the number of violation data (N=21525) is significantly less than non-violation data (*N*=194715). The imbalance ratio between non-violation data and violation data is 9.

4 Analysis of traffic violation influencing factors

4.1 Influencing factors analysis model

To analyze various factors affecting the probability of violations, a discrete choice model that could determine discrete outcomes was estimated. For traffic violation behavior under the continuous data environment, a binary model was needed since there were only two possible outcomes: occur and not occur. To accommodate the binary model, the response variable of occurrence was created, labeled 1 if a violation occured at the AES within 30 minutes, and zero otherwise. The utility function of whether or not a traffic violation occurred is shown:

$$y_i = \beta_i^{'} X_i + \varepsilon_i \quad i = 0,1 \tag{1}$$

where y_i is the response variable of violation occurrence; X_i is the vector of the explanatory variable; $\beta_i^{'}$ is the vector of coefficients for the explanatory variable; ε_i is the random component (error term) that explains the unobserved effect on the frequency of the violation observation. For the logistic model used in this study, the probability of a traffic violation occurring is shown:

$$P(y_i = 1 \mid X_i) = \frac{\exp(\beta_i^{'} X_i)}{1 + \exp \beta_i^{'} X_i}, \ (i = 1, \ldots, n) \tag{2}$$

4.2 Impact of influencing factors

The Python sklearn package can be used with the logistic model to estimate the model coefficients (Pedregosa et. al., 2011). Table 3 displays coefficient estimation results and adjusted odds ratios (95% CI) in stepwise logistic analysis. The adjusted odds ratio (OR) of significant factors and their 95% confidence intervals (CIs) were computed using a stepwise logistic model in which all factors were initially included and from which insignificant factors were subsequently removed by the stepwise procedure. Entry and removal probabilities for the stepwise procedure were both set at 0.05. As shown in Table 3, except for the administration district (X_{32}), recreational district (X_{35}), slight congestion condition (X_{51}), and rainfall (X_9), all factors were significantly correlated with traffic violations at intersections.

The time factors of occurrence time and day are represented by dummy variables. The coefficient of the occurrence time variable indicates that the probability of a violation during the daytime is significantly higher than at night. The OR value shows that the probability of a violation in the midday period (X_{12}) increases five times more than the night period (X_{10}), while increasing the probability in the morning (X_{11}) and evening (X_{13}) periods by 4.7 (95% CI=5.395-5.987) and 1.7 (95% CI=2.51-2.82), respectively. In addition, the coefficient on the occurrence day variable indicates that the probability of a violation increases slightly on weekends (X_{21}, OR=1.162, 95% CI=1.123-1.202) and holidays (X_{22}, OR=1.144, 95% CI=1.085-1.206) compared to weekdays (X_{20}). Among land-use factors, the probability of a violation is highest in the residential district (X_{30}). Collector roads (X_{42}, OR=1.216, 95% CI=1.109-1.334) exhibit a particularly higher risk of traffic violation than ramps (X_{40}). In contrast, the probability of a violation on a local road (X_{43}, OR=0.531, 95% CI=0.471-0.6) is the smallest. The P-value of the slight congestion conditions (X_{51}) is 0.1 >0.05, which means traffic slows down has no significant association with traffic violations. In contrast, severe congestion (X_{52}, OR=3.646, 95% CI=3.193-4.164) increases the probability of a violation. The OR of traffic flow (X_6) is 3.067 (95% CI=2.835-3.318) shows that with an increase in traffic flow, the probability of a violation also increases. Factors such as average wind (X_8) and temperature (X_9) also affect the probability of a violation to a certain extent. Rainfall has a P value of 0.2746 >0.05, showing no significant association with traffic violations.

Table 3 Coefficient estimation results and adjusted odds ratios (95% CI) in stepwise logistic analysis

Variables	Symbol	Coef.	Std.Err.	z	P>|z|	ORs	2.5%	97.5%
Intercept		-3.512	0.057	-61.307	0	0.03	0.027	0.033
Nights	X_{10}	(base variable)						
Morning rush hour	X_{11}	1.738	0.027	65.407	0	5.683	5.395	5.987
Midday	X_{12}	1.901	0.022	87.519	0	6.694	6.415	6.986
Evening rush hour	X_{13}	0.979	0.030	33.014	0	2.66	2.510	2.820
Weekday	X_{20}	(base variable)						
Weekend	X_{21}	0.150	0.017	8.651	0	1.162	1.123	1.202
Hoiliday	X_{22}	0.134	0.027	5.008	0	1.144	1.085	1.206
Residential	X_{30}	(base variable)						
Commercial	X_{31}	-0.198	0.025	-7.904	0	0.82	0.781	0.862
Administration	X_{32}	-0.035	0.027	-1.281	0.2001	-	-	-
Education	X_{33}	-0.103	0.024	-4.348	0	0.902	0.862	0.945
Hospital	X_{34}	-0.924	0.069	-13.412	0	0.397	0.347	0.454
Recreational	X_{35}	0.013	0.025	0.519	0.6039	-	-	-
Ramps	X_{40}	(base variable)						
Arterial road	X_{41}	0.154	0.050	3.067	0.0022	1.167	1.057	1.287
Collector road	X_{42}	0.196	0.047	4.166	0	1.216	1.109	1.334
Local road	X_{43}	-0.632	0.062	-10.245	0	0.531	0.471	0.600
Free flow	X_{50}	(base variable)						
Slight congestion	X_{51}	0.166	0.101	1.644	0.1001	-	-	-
Severe congestion	X_{52}	1.294	0.068	19.091	0	3.646	3.193	4.164
Traffic flow	X_6	1.121	0.040	27.893	0	3.067	2.835	3.318
Average wind	X_7	-0.181	0.046	-3.904	0.0001	0.834	0.762	0.914
Temperature	X_8	-0.890	0.040	-21.998	0	0.411	0.380	0.445
Rainfall	X_9	0.222	0.203	1.093	0.2746	-	-	-

Logistic regression model results:		No. Observations	216,240
Log-Likelihood:	-63,437	Wald chi2 (19)	77,534
Pseudo R-squared:	0.096	Prob > chi2	0.000

4.3 Discussion

Since the traffic violation dataset in this study is derived from AES, potential influencing factors, that the existing literature may not have considered, are made available. For example, some studies use police enforcement reports to analyze traffic violations. The sample of these data is incomplete, often lacking accurate time, space, and environmental data. In contrast, since a full dataset under the continuous data environment was included in this study, unobserved variability in the error component can be reduced, making the results more accurate and reliable.

Occurrence time is a significant factor in traffic violations. As the traffic flow increases during the morning and evening rush hours, the probability of traffic violations also increases. Morning rush hours exhibit a higher probability of violation than evening hours, possibly because drivers are anxious in the morning to arrive on time to work. There is a slight increase of violation probability on weekends and holidays. Since Wujiang is a tourist city, it will attract a lot of tourists on weekends and holidays. Tourists may be unfamiliar with the road environment and lead to some violations such as wrong-way driving and traffic marking violations.

As land use is a static factor, traffic demand for different land uses affects the probability of violations. For example, the residential district is widely distributed while the commercial, education, and hospital districts are concentrated in Wujiang. The more concentrated demand in these districts results in stricter traffic management, an important reason for these districts' lower probability of traffic violations than that of the residential district. Arterial and collector road types have higher rates of traffic violations, in contrast with local roads, which have the lowest probability. Although some studies have shown that speeding violations are more likely to occur on the ramps (Zhang et al., 2011, Cheng et al., 2019), it is worth noting that this study found fewer traffic violations on ramps likely because these roads have a limited number of at-grade intersections, and a greater number of physical separation barriers such as central reservations or green belts. Collector roads are more likely to occur violations. Usually, only one left-turn lane is set on collector roads. As one left-turn lane may be insufficient to satisfy the demand, the queue length becomes oversaturated and may prompt many drivers to violate the left-turn ban markings in other lanes.

Traffic congestion shows a particularly high rate of violations, as drivers are more likely to violate traffic signs and markings, and perform unlawful cut-in in this condition. Vehicles in slight congestion conditions may not significantly incur traffic violations because drivers are more cautious about keeping an interval between vehicles due to safety considerations. Higher wind speed may reduce smog and improve visibility of the color of signal lights at intersections, resulting in a lower violation rate. Lower temperature (the lowest temperature in February is around 0°C) may cause roads to freeze and increase braking distance, further increasing the probability of traffic violations at intersections, such as speeding, red-light running.

In summary, the results obtained in this study show that a traffic violation is often the result of a driver interacting with potential factors. In general, intersections are high-risk points for traffic violations and multiple types of violations. Although these results are entirely derived from continuously recorded traffic violation data by AES, they are consistent with the experience of traffic managers and police officers' on-site observations. Moreover, the analysis results of traffic violation influencing factors are useful for traffic managers and police officers developing control strategies for different types of risk at intersections.

5. Traffic violations prediction model

5.1 Random forest algorithm

Predicting whether or not a traffic violation occurs is a typical binary classification problem, and helps traffic management authorities to implement countermeasures to prevent traffic violations. The random forest (RF) algorithm is an effective way to improve the performance of a decision tree while retaining most of the latter's appealing properties (Liaw et al., 2002). In general, the RF classifier uses a set of classification and regression trees (CART) to make a prediction, trees that are created by drawing a subset of training samples through replacement (a bagging approach). About two-thirds of the samples (referred to as in-bag samples) are used to train the trees, and the remaining one-third of the samples (out-of-bag samples, OOB) is used in a type of cross-validation in parallel with the training step to estimate the RF algorithm performance. Each decision tree is independently produced without any pruning, and each node is split using a user-defined number of features, selected at random. According to grow up to a user-defined number of trees, the algorithm creates trees to get a high variance and low bias (Dogru et al., 2018). The final classification decision is taken by averaging (using the arithmetic mean) the class assignment probabilities calculated by all produced trees.

There are two random procedures in the RF algorithm: first, training sets are constructed by randomly using a bootstrap mechanism with replacement. Second, random features are selected with non-replacement from the total features when the nodes of the trees are split. Let the original dataset $T = \{(x_{i1}, x_{i2}, \ldots, x_{iM}, y_i)\}_{i=1}^{N}$, the vector $x_{i1}, x_{i2}, \ldots, x_{iM}$ denote the *M*-dimension attributes or features, $Y = \{y_i\}_i^N$ denotes classification labels, and a sample is deduced as label *c* by $y_i = c$. Thus, the pseudocode for the RF algorithm is showed below.

Algorithm 1: pseudocode for random forest algorithm

Input: training set, testing set, cluster number *c*, tree number *nTree*, deepness, hyper parameter κ, attribute select method, termination criteria

Procedure:

```
1.  for $i$ = 1 to nTree do
2.  Use the bootstrap method to produce training sets with size N for each tree
3.  Select κ attributes randomly building nodes and split the dataset by the best attribute
4.  Generate each tree recursively without pruning
5.  end for
6.  Calculate the probability of an unknown sample $x$ belonging to class $c$:
```

```
7.  $p(c \mid x) = (\frac{1}{nTree})\sum h_j(c \mid x)$
8.  Predict class through majority voting
9.  $c \leftarrow \arg\max p(c \mid x)$, and calculate $OOB\ error$
Output: random forest classification model and classification results
```

Similar to most classifiers, random forests can also suffer from the problem of learning from an extremely imbalanced training data set. The problem with the RF procedure (see Algorithm 1) in the presence of imbalanced data is two-fold. On one hand, successive partitioning of the dataspace results in fewer and fewer observations of minority class examples resulting in fewer leaves describing minority concepts and successively weaker confidence estimates. On the other hand, concepts that have dependencies on different feature space conjunctions can go unlearned by the sparseness introduced through partitioning. Thus, this study mainly uses the proximity weighted synthetic oversampling technique (ProWSyn) to reduce imbalanced data classification and thereby improve the RF algorithm performance.

5.2 ProWSyn algorithm

The imbalanced data problem significantly compromises the performance of most standard learning algorithms. The imbalance ratio (IR) is the widely accepted measure to determine imbalanced data. IR is the ratio between majority class and minority class:

$$IR = \frac{Majority\ class}{Minority\ class} \tag{4}$$

A dataset can be considered imbalanced if IR > 1.5. The synthetic minority oversampling technique (SMOTE) algorithm is a powerful solution to the problem that has shown success in various application domains. This technique mainly adds synthetic minority class samples to the original dataset to achieve a balanced dataset. Consequently, many extensions and alternative methods based on SMOTE have been proposed under different scenarios in recent years. For example, Yahaya (2019) established a framework for the combined use of variable selection and SMOTE data balance techniques to handle the dimensionality and imbalanced problems associated with the crash data. Cai (2020) used a deep convolutional generative adversarial network model to fully understand the traffic data leading to crashes. Islam (2021) used a variational autoencoder model to generate millions of crash samples from only a limited number of training data. Kovács (2019) used 104 imbalanced datasets to conduct a detailed comparison of 85 improved SMOTE methods, and thus gained insight into the performance of oversamples on various types of datasets and the performance of various operating principles. The results showed the proximity weighted synthesis

(ProWSyn) as top-performing compared in comparison with other methods.

The ProWSyn method, established by Barua et al. (2013), is an efficient oversampling approach even though the assumption made on the distribution of data is stronger than that of SMOTE. ProWSyn uses the distance information between the minority class samples and the majority samples in assigning weights to the minority class samples, and its effectiveness has been evaluated on several benchmark classification problems with high imbalance ratios. ProWSyn is therefore used in this paper to improve the imbalance ratio of the traffic violations dataset. The pseudocode for the ProWSyn algorithm is shown below.

Algorithm 2: pseudocode for ProWSyn

```
Input:
  $X_{origin}$ is the original imbalanced data with $N$ samples $x_i$, $i$=1,…,$N$, where $x_i$ is an instance in $m$ dmensional
  feature space.
  $S_{maj}$ and $S_{min}$ are majority and minority class sets, respectively.
  $N_{maj}$ and $N_{min}$ are the number of majority and minority class samples respectively.
Procedure:
  1. Calculate the number of synthetic samples that need to be generated for the minority class: $G=(N_{maj} - N_{min}) \times \beta$, where $\beta \in [0,1]$ is a parameter used to specify the desired balance level aftergeneration of
     the synthetic samples.
  2. Initialize, $P= S_{min}$.
  3. For i=1 to L-1
     a. From each majority sample y, find the nearest $K$ minority samples in $P$ according to Euclidean
        distance. Let, the set of these $K$ samples be $D_K(y)$.
     b. Form partition $P_i$ as the union of all $D_K(y)$s:
              $P_i = \bigcup_{y \in S_{maj}} D_K(y)$
     c. Set proximity level of each minority sample $x$ in partition $P_i$ to be $i$:
              $PL_x = i,\ \forall x \in P_i$
     d. Remove selected minority samples from $P$, $P$=$P$- $P_i$
     e. end for
  4. Form partition $P_L$ with the remaining unpartitioned samples in $P$.
  5. Set proximity level of each $x$ in $P_L$ to be $L$:
     $PL_x = L,\ \forall x \in P_L$
  6. For each $x$, calculate a weight $\omega_x$ from its proximity level $PL_x$ defined as:
     $\omega_x = \exp(-\theta \times (PL_x - 1))$
     where $\theta$ is a smoothing factor and $\omega_x \in [0,1]$.
  7. Normalize $\omega_x$ according to $\hat{\omega}_x = \omega_x / \sum_{z \in S_{\min}} \omega_z$
  8. Calculate the number of synthetic samples $g_x$ that need to be generated for $x$:
  $g_x = \hat{\omega}_x \times G$
  9. Find the clusters of minority set, $S_{min}$
  10. Initialize set, $S_{omin}$=$S_{min}$
```

```
11. For each $x$, generate $g_x$ synthetic minority class samples according to the following steps:
    Do the **Loop** from 1 to $g_x$
        a. Randomly select one minority sample y, from $x$'s cluster (as found in step 9).
        b. Generate a synthetic sample, s, according to
           $s = x + \alpha \times (y - x)$, where $\alpha$ is a random number in the range [0, 1].
        c. Add s to $S_{omin}$ : $S_{omin} = S_{omin} \cup \{s\}$
    end **Loop**
**Output:** Oversampled minority data set, $S_{omin}$
```

5.3 Performance measures

Accuracy, recall, specificity, precision, F1-score, geometric mean (G-mean), and area under curve (AUC) are commonly used performance measures, and are used to evaluate the effectiveness of the prediction model in this paper. The accuracy of a binary classifier is often described by a confusion matrix in which *TP*, *FN*, *FP* and *TN* are, respectively, true positive, false negative, false positive, and true negative. The accuracy, recall, specificity and precision measures are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The classifiers may have high overall accuracy, with 100% accuracy in the majority class while only 0-10% accuracy in the minority class, because the overall accuracy is biased towards the majority class. Hence, the accuracy measure is not a proper evaluation metric for the imbalanced class problem. Instead, the F1-score, G-mean, and AUC are used to evaluate imbalanced data. Usually, F1-score is a performance metric that links both precision and recall, and a larger F1-score indicates a better classifier. The F1-score measure is defined as follows:

$$F1 - score = \frac{2}{1 / Precision + 1 / Pecall} \tag{7}$$

The G-mean attempts to maximize the accuracy across the two classes with a good balance. Only when both sensitivity and specificity are high can the G-mean attain its maximum, which indicates a better classifier. The G-mean is defined as follows:

$$G - mean = \sqrt{Recall \bullet Specificity} \tag{8}$$

AUC is the area under the receiver operating characteristics (ROC) curve, and is commonly used to presents result for binary decision problems in machine learning. The ROC curves represent the trade-off between TP and FP rates (Fawcett, 2006), which shows how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. However, ROC curves can present an overly optimistic view of an algorithm's performance if there is a large imbalance in the class distribution data. Thus, Precision-Recall (PR) curves have been cited in this paper as a supplementary method to ROC curves to verify the performance of class distribution algorithms. One of the most important differences between ROC curves and PR curves is the visual representation of the curves. The PR curves can expose differences between algorithms that are not apparent in ROC curves. For example, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis. Precision by comparing false positives to true positives rather than true negatives captures the effect of a large number of negative examples on the algorithm's performance(Davis & Goadrich, 2006).

6. Results analysis

6.1 Model performance

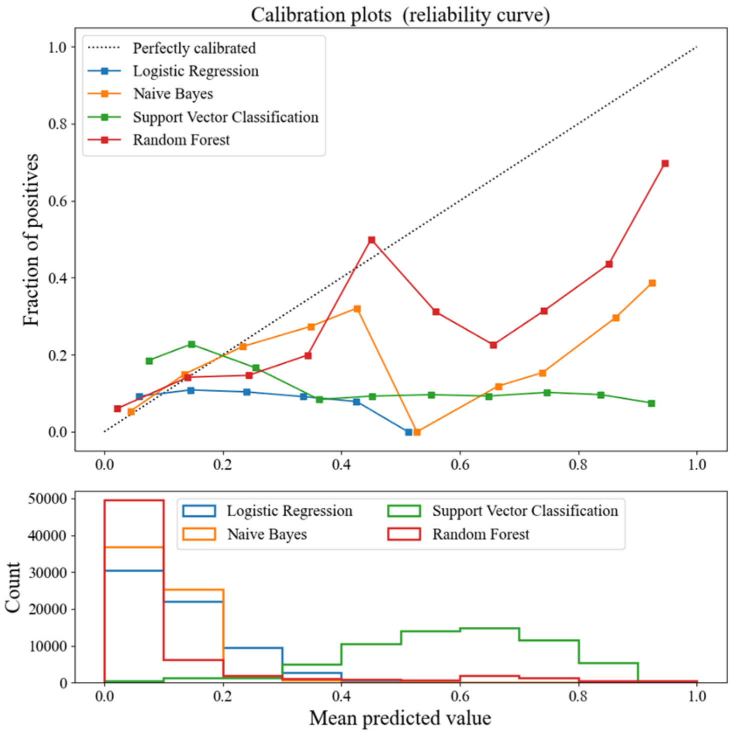
The traffic violation data is as described in Section 3, with each of the four datasets containing 9 feature variables. The full dataset under the continuous data environment contains 216,120 violation occurrences. As mentioned in Section 4.1, label 1 means a traffic violation occurred at the AES within 30 minutes, and label 0 means there was no violation. Accordingly, the classes labeled 1 and 0 are 21,445 and 194,675, respectively. To simulate the actual situation appropriately and preserve the degree of imbalance of the original data, the training and testing sets were divided using stratified random sampling at a ratio of 3:1. The results of each experiment were averaged over 5 times with 10-fold cross-validation to eliminate random effects.

A well-calibrated classifier is essential for the accuracy of prediction models. In this study, four widely used classifier models, Logistic Regression (LR), Gaussian NaiveBayes (GNB), Support Vector Machine (SVM), and Random Forest, were examined using data and ProWSyn oversampling data, respectively. In Figure 2, the reliability curves of each classifier are presented in the top graphs, and the histograms of the mean predicted value are below. As shown in Figure 2a, the original data performance of the four uncalibrated models, especially the logistic regression and support vector machine cannot represent the true probabilities well. Although the LR model is simple and easy to get a good probability interpretation, it is difficult to handle the large feature space and imbalanced data. Similarly, the prediction performance of

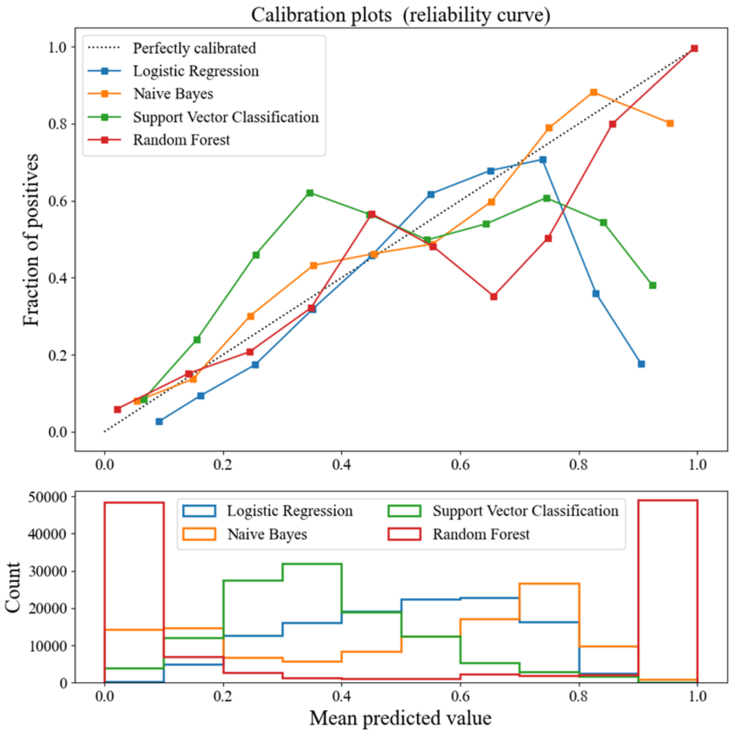
SVM model is not well when training large-scale data. When imbalance ratio from 9 to 1, as shown in Figure 2b, models' reliability curves show that the predicted probabilities are optimized closer to the perfectly calibrated line. It is worth noting that the RF algorithm shows the best performance, with its histogram peaks at 0 and 1 probabilities, while other probabilities are very rare, likely because the base-level trees trained with random forests have relatively high variance due to feature subsetting. As a result, the reliability curve shows a characteristic sigmoid shape, confirming that the RF returns probabilities typically closer to 0 or 1. Thus, the RF algorithm can be considered as an optimal traffic violation prediction model, and reducing the imbalance ratio can effectively improve its performance.

As mentioned in Sections 5.1 and 5.2, the RF algorithm was proposed for predicting traffic violations, and the logistic regression model was used as a comparison. Each of the two models used the 0.7 proportion of the original data, undersampling and oversampling data as the train data, and 0.3 proportion of the original data as the test data. Figure 3 shows the ROC curves of logistic regression and random forest. The blue line indicates the ROC curve for the original data, and the brown line and the green line indicate the ROC curve for the undersampling data (decreasing the majority class) and oversampling data (increasing the minority class), respectively. It can be observed that all ROC curves are higher than 0.5, indicating a better prediction performance of the RF algorithm. For example, the AUC of RF using original data is 0.763, higher than that of logistic regression (AUC=0.516), and the AUC of RF using oversampling data (AUC=0.913) is much higher than logistic regression (AUC=0.66). Both oversampling data and undersampling data show similar ROC curves in logistic regression, but the prediction performance of the RF algorithm using oversampling data is significantly higher than the undersampling and original data. Although Figure 3 illustrates that both logistic regression and the RF algorithm can predict the probability of traffic violations, the PR curves in Figure 4a show that the logistic regression presents an extremely low precision performance in the class distribution. Regardless of whether change the imbalance ratio or not, the average precision (AP) is less than 0.15. In contrast, the PR curves in Figure 4b show that the area under the green line is significantly higher than the brown and blue lines. The AP of RF using oversampling data (AP=0.635) is about twice as high as the original data (AP=0.315). Figure 4b also shows that increasing the minority class is more helpful to improve the prediction performance of the RF algorithm than decreasing the majority class.

The prediction results of the model are shown in Table 4. The accuracy of logistic regression using original data is 0.9. Because the IR of the original data is 9 (i.e., the original traffic violation data only accounted for 9.9% of the full data sample), logistic regression's prediction accuracy can achieve 90% even though the method's results are wrong. Despite the values of recall of logistic regression using undersampling data is high, the low precision and specificity indicate the number of negative examples in this dataset greatly exceeds the number of positives examples. Thus, the values of F1-score and G-mean also indicate that the logistic regression cannot solve the problem of imbalanced data classification. For comparison, the RF algorithm has a better classifier's prediction performance for imbalanced data not only in AUC and AP but also in F1-score and G-mean.

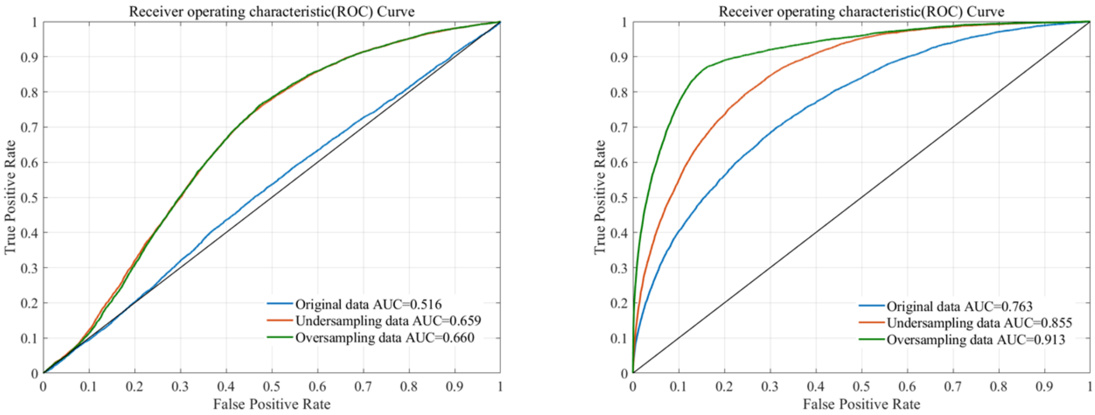


(a) Reliability Curves with original data



(b) Reliability Curves with oversampling data

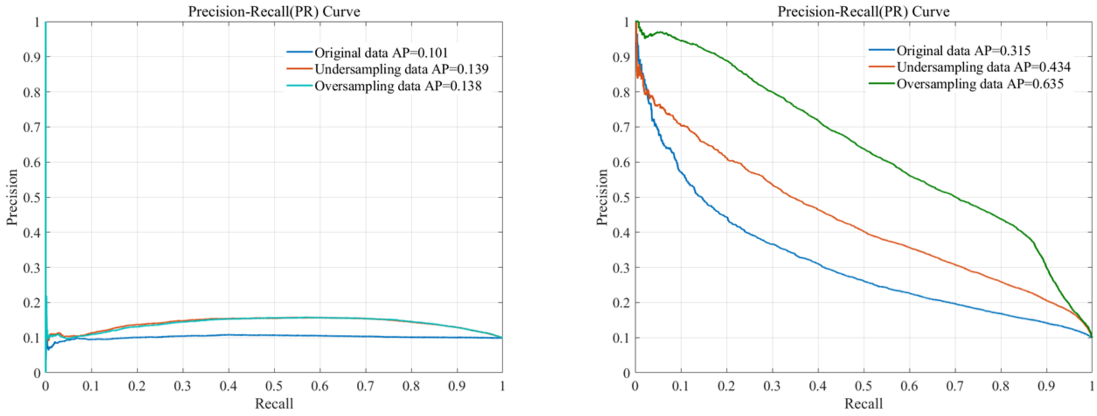
Figure 2 Reliability curves with calibration method



(a) Logistic regression

(b) Random forest

Figure 3 ROC curves of logistic regression and random forest



(a) Logistic regression (b) Random forest

Figure 4 Precision-Recall curves of logistic regression and random forest

Table 4 Prediction results with logistic regression and random forest

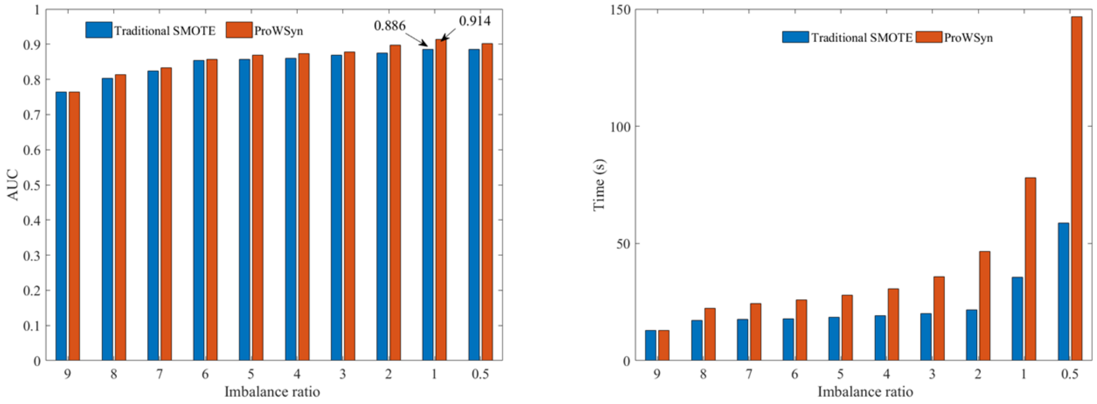
Performance	Logistic regression			Random forest		
measures	Original	Underampling	Oversampling	Original	Underampling	Oversampling
AUC	0.5165	0.6601	0.6590	0.7635	0.8559	**0.9135**
AP	0.1015	0.1397	0.1384	0.3149	0.4340	**0.6351**
Accuracy	0.9004	0.5907	0.5932	0.9008	0.6958	**0.9203**
Precision	0.1002	0.1541	0.1547	0.5039	0.2282	**0.6157**
Recall	0.0008	**0.6945**	0.6925	0.1392	0.8648	0.5280
Specificity	0.0602	0.1541	0.1547	0.5039	0.2282	**0.6157**
F1-score	0.0015	0.2523	0.2529	0.2182	0.3611	**0.5685**
G-mean	0.0069	0.3272	0.3273	0.2649	0.4443	**0.5702**

6.2 Impact of imbalance ratio

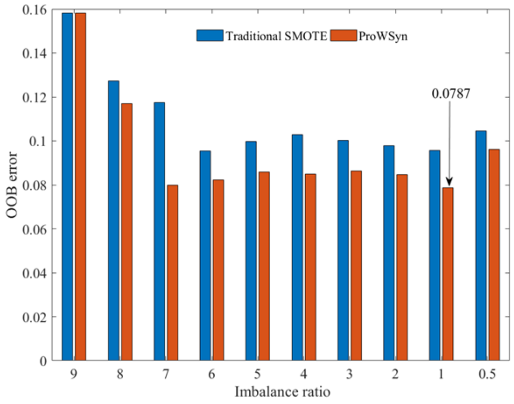
To verify the impact of the imbalance ratio (IR) on model prediction performance, the minority class of the training subset is oversampled using traditional SMOTE and ProWSyn. The model performance results are shown in Table 5 for the oversampling from 9 to 1 and 0.5. As can be seen in the table, the optimal IR=1 for RF, and the prediction performance of the ProWSyn method (AUC=0.914) is the best than traditional SMOTE (AUC=0.886), in Figure 5a. The model results show that the prediction performance will improve when the distribution of violation and non-violation is more balanced. When IR is high, the values of F1-score (0.2182), and G-mean (0.2649) are generally low because of the minority class of traffic violations. Compare with traditional SMOTE, the ProWSyn method effectively increases the proportion of violations in the full dataset and helps predict results biased towards the violation class. At the same time, the ProWSyn method also improves the prediction performance of the RF algorithm. However, the ProWSyn method uses the distance information between the minority class samples and the majority samples in assigning weights to the minority class samples. The training time of ProWSyn is higher than traditional SMOTE, especially when IR<1, in Figure 5b. Thus, the computational efficiency of the model will be further considered. It is worth noting that the performance of the RF algorithm is not linearly related to IR. For example, when IR changes from 9 to 8, the AUC of the model increases by 6.5% (0.764 to 0.814); when IR changes from 2 to 1, the AUC only increases by 1.8% (0.897 to 0.914). When IR approximately 1, the AUC of the RF algorithm gradually converges and reached a limiting value.

Table 5 Prediction results with different imbalance ratios

	Traditional SMOTE					ProWSyn				
IR	AUC	F1-score	G-mean	OOB error	Time(s)	AUC	F1-score	G-mean	OOB error	Time(s)
9	0.7635	0.2182	0.2649	0.1581	13.0165	0.7635	0.2182	0.2649	0.1581	13.0165
8	0.8024	0.3135	0.3648	0.1274	17.2003	0.8135	0.3973	0.4260	0.1171	22.4237
7	0.8244	0.3295	0.3807	0.1175	17.6711	0.8328	0.4154	0.4409	0.0800	24.3647
6	0.8539	0.3209	0.3721	0.0954	17.8745	0.8570	0.4235	0.4472	0.0823	25.9142
5	0.8574	0.3280	0.3788	0.0998	18.5368	0.8696	0.4323	0.4545	0.0858	27.8603
4	0.8595	0.3279	0.3800	0.1028	19.1823	0.8737	0.4472	0.4663	0.0848	30.7194
3	0.8683	0.3467	0.3975	0.1002	20.0771	0.8779	0.4584	0.4744	0.0864	35.8337
2	0.8746	0.3782	0.4260	0.0978	21.7353	0.8973	0.4834	0.4945	0.0847	46.5519
1	0.8858	0.4084	0.4521	0.0957	35.5334	**0.9135**	**0.5685**	**0.5702**	**0.0787**	77.9217
0.5	0.8855	0.4336	0.4710	0.1046	58.6838	0.9020	0.5259	0.5326	0.0961	146.692



(a) AUC with different IRs (b) Training times with different IRs



(c) OOB error with different IRs

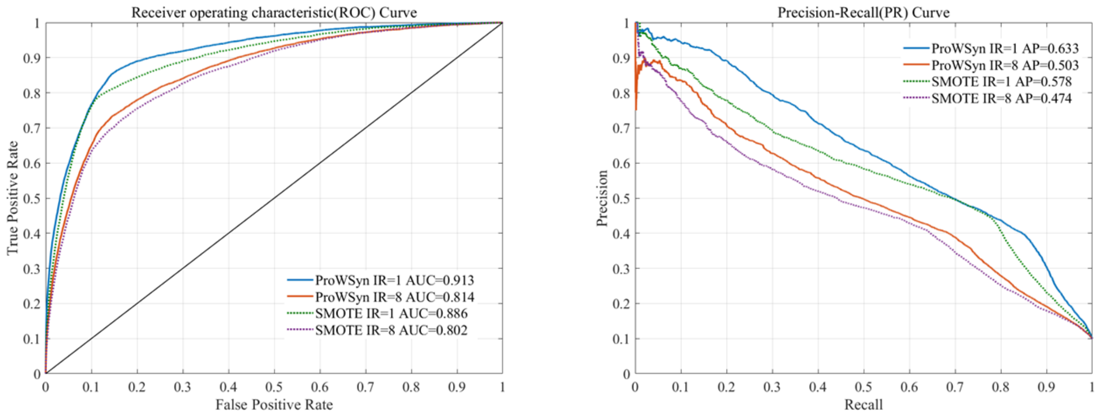
Figure 5 Comparison of the ProWSyn method and traditional SMOTE method

Figure 6 shows the ROC curves and PR curves of the RF algorithm using oversampling data by the traditional SMOTE and ProWSyn method. The goal in the area under the ROC curve is to be in the upper-left-hand corner. Both traditional SMOTE and ProWSyn methods of ROC curves in Figure 6a appear to be fairly close to optimal, but the PR curves in Figure 6b show that there is still room for improvement. At the same time, the PR curves show that ProWSyn can statistically outperform traditional SMOTE and effectively solve the problem of imbalanced data classification in traffic violation prediction.

In addition, RF usually performs a type of cross-validation in parallel with the training step by using the out-of-box (OOB) samples. OOB estimation is an unbiased estimate of the RF algorithm and can be used to measure the classifier's generalization ability. A smaller OOB error indicates a better classification performance. OOB error is defined as follows:

$$OOB\ error = \sum_{i}^{nTree} OOB\ error_i \,/\, nTree \qquad (9)$$

where *nTree* is the number of trees. Figure 7 shows the relationship between OOB error and the number of trees under four imbalance ratios. As can be seen, the OOB error rate decreases dramatically as more trees are added to the forest, and a limiting value of the OOB error rate is eventually reached. Figure 7 also shows that the RF algorithm converges as more than 60 trees are added. In addition, when IR>1, the OOB error (0.096) with IR=0.5 is significantly higher than the OOB error (0.079) with IR=1, in Figure 5c.



(a) ROC curves

(b) PR curves

Figure 6 ROC and PR curves of RF using oversampling data

In summary, IR=1 is the best imbalance ratio for predicting traffic violations, and ProWSyn can statistically outperform traditional SMOTE. The RF algorithm can effectively predict traffic violations, and the AUC of the model exceeds 0.91.

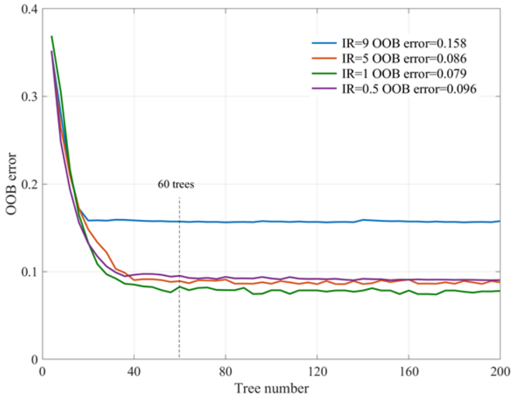


Figure 7 OOB errors with different imbalance ratios

7. Conclusions

The purpose of this study has been to analyze and predict traffic violation behavior at intersections. No doubt studying the influencing factors of violations can not only improve traffic safety but also may reduce some traffic crashes at the source. However, the traffic violation data of previous studies were primarily based on traffic crash reports or field observations. These data had shortcomings such as small sample size, low accuracy, low timeliness, and strong subjectivity (Stylianou et al., 2019, Mannering et al., 2020). Additionally, these data are not comprehensive and might have sampling errors that cause researchers to miss many important variables, leading to estimate errors in analyzing the factors involved. Therefore, traffic violation data in this study were completely derived from automated enforcement system (AES) raw data.

These violation data were used in this study for two purposes: (i) analyzing the significant factors associated with traffic violations in China, and (ii) predicting the probability of traffic violations at intersections.

According to the logistic regression, this study had established that traffic violations were related to 20 factors. The midday period, weekend days, residential districts, collector roads, congested conditions, high traffic flow, lower wind speed, and low temperature would increase the probability of traffic violations. These results could help traffic managers and police officers to implement countermeasures to mitigate specific traffic hazards. For example, a set of measures such as drivers' cautious behavior, reasonable road facilities and strict traffic enforcement had successfully reduced traffic violations, and the impact of traffic crashes.

The AES data were further used to predict the probability of violations at intersections. The random forest was found to be the best traffic violation prediction model compared with Logistic Regression, Gaussian naïve Bayes, Support Vector Machine. The model results showed that the RF algorithm outperforms logistic regression in terms of many performance metrics such as AUC, AP, precision, recall, specificity, F1-score, and G-mean. To improve the imbalance ratio of non-violation data and violation data, the ProWSyn, an effective oversampling method, was used to increases the proportion of traffic violations in the full dataset and help predict results biased towards the violation class. Hence, the combined algorithm of RF and ProWSyn could not only improved the prediction of traffic violations, but also assisted the active traffic management system to reduce some traffic crashes.

Future research should be conducted to address the following two concerns. AES can detect many types of traffic violations, such as violating traffic signs and markings, red-light running, speeding, on which influencing factors may have different weights. Thus, it is necessary to quantitatively analyze the sensitivity of different types of traffic violations at the same intersection. Additionally, some deep learning methods such as ConvLSTM, ConvGRU, and ST-LSTM can be applied to investigate hidden variables that can improve the overall predictive ability of the model.

Acknowledgments

The research is supported by grants from National Key R&D Program of China (2018YFB1601600).

Reference

1. Adanu, E. K., Lidbe, A., Tedla, E., & Jones, S. (2021). Factors associated with driver injury severity of lane changing crashes involving younger and older drivers. Accident Analysis & Prevention, 149, 105867. http://doi.org/https://doi.org/10.1016/j.aap.2020.105867

2. Ahmed, M. M., & Abdel-Aty, M. (2015). Evaluation and spatial analysis of automated red-light running enforcement cameras. Transportation Research Part C: Emerging Technologies, 50, 130-140. http://doi.org/https://doi.org/10.1016/j.trc.2014.07.012
3. Amiruzzaman, M. (2019). Prediction of Traffic-Violation Using Data Mining Techniques. In Proceedings of the Future Technologies Conference (pp. 283-297). Springer, Cham. https://doi.org/10.1007/978-3-030-02686-8_23
4. Arhin, S. A., & Gatiba, A. (2020). Predicting crash injury severity at unsignalized intersections using support vector machines and naïve Bayes classifiers. Transportation Safety and Environment, 2(2), 120-132. http://doi.org/10.1093/tse/tdaa012
5. Ayuso, M., Sánchez, R., & Santolino, M. (2020). Does longevity impact the severity of traffic crashes? A comparative study of young-older and old-older drivers. JOURNAL OF SAFETY RESEARCH, 73, 37-46. http://doi.org/https://doi.org/10.1016/j.jsr.2020.02.002
6. Balasubramanian, V., & Sivasankaran, S. K. (2021). Analysis of factors associated with exceeding lawful speed traffic violations in Indian metropolitan city. Journal of Transportation Safety & Security, 13(2), 206-222. http://doi.org/10.1080/19439962.2019.1626962
7. Baratian-Ghorghi, F., Zhou, H., & Wasilefsky, I. (2016). Effect of Red-Light Cameras on Capacity of Signalized Intersections. JOURNAL OF TRANSPORTATION ENGINEERING, 142(1), 4015035. http://doi.org/10.1061/(ASCE)TE.1943-5436.0000804
8. Barua, S., Islam, M. M., & Murase, K. (2013, 2013-01-01). ProWSyn: Proximity Weighted Synthetic Oversampling Technique for Imbalanced Data Set Learning. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 317-328). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_27
9. Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. Transportation Research Part C: Emerging Technologies, 117, 102697. http://doi.org/https://doi.org/10.1016/j.trc.2020.102697
10. Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., & Ratanavaraha, V. (2020). Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. Accident Analysis & Prevention, 141, 105537. http://doi.org/https://doi.org/10.1016/j.aap.2020.105537
11. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH, 16, 321-357. http://doi.org/10.1613/jair.953
12. Cheng, Z., Lu, J., Zu, Z., Li, Y., & Lambert, A. (2019). Speeding Violation Type Prediction Based on Decision Tree Method: A Case Study in Wujiang, China. JOURNAL OF ADVANCED TRANSPORTATION, 2019, 8650845. http://doi.org/10.1155/2019/8650845
13. Davis, J., & Goadrich, M. (2006, 2006-01-01). The relationship between Precision-Recall and ROC curves. Paper presented at the Proceedings of the 23rd International Conference on Machine Learning - ICML '06.
14. Dogru, N., & Subasi, A. (2018). Traffic accident detection using random forest classifier. In 2018 15th learning and technology conference (L&T) (pp. 40-45). IEEE. DOI: 10.1109/LT.2018.8368509
15. Eccles, K. A. (2012). Automated enforcement for speeding and red light running (Vol. 729). Transportation Research Board. https://doi.org/10.17226/22716
16. Elamrani Abou Elassad, Z., Mousannif, H., & Al Moatassime, H. (2020). A proactive decision support system for predicting traffic crash events: A critical analysis of imbalanced class distribution. KNOWLEDGE-BASED SYSTEMS, 205, 106314. http://doi.org/https://doi.org/10.1016/j.knosys.2020.106314
17. Erke, A. (2009). Red light for red-light cameras?: A meta-analysis of the effects of red-light cameras on crashes. Accident Analysis & Prevention, 41(5), 897-905. http://doi.org/https://doi.org/10.1016/j.aap.2008.08.011
18. Fang, A., Qiu, C., Zhao, L., & Jin, Y. (2018). Driver Risk Assessment Using Traffic Violation and Accident Data by

Machine Learning Approaches. Paper presented at the 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE).

19. Fawcett, T. (2006). An introduction to ROC analysis. PATTERN RECOGNITION LETTERS, 27(8), 861-874. http://doi.org/10.1016/j.patrec.2005.10.010

20. Fei, G., Li, X., Sun, Q., Qian, Y., Stallones, L., Xiang, H., & Zhang, X. (2020). Effectiveness of implementing the criminal administrative punishment law of drunk driving in China: An interrupted time series analysis, 2004-2017. Accident Analysis & Prevention, 144, 105670. http://doi.org/https://doi.org/10.1016/j.aap.2020.105670

21. FHWA, N. (2007). Red light camera systems operational guidelines.

22. Goldenbeld, C., Daniels, S., & Schermers, G. (2019). Red light cameras revisited. Recent evidence on red light camera safety effects. Accident Analysis & Prevention, 128, 139-147. http://doi.org/https://doi.org/10.1016/j.aap.2019.04.007

23. Islam, Z., Abdel-Aty, M., Cai, Q., & Yuan, J. (2021). Crash data augmentation using variational autoencoder. Accident Analysis & Prevention, 151, 105950. http://doi.org/https://doi.org/10.1016/j.aap.2020.105950

24. Jahangiri, A., Rakha, H., & Dingus, T. A. (2016). Red-light running violation prediction using observational and simulator data. Accident Analysis & Prevention, 96, 316-328. http://doi.org/https://doi.org/10.1016/j.aap.2016.06.009

25. Jiang, X., Abdel-Aty, M., Hu, J., & Lee, J. (2016). Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. NEUROCOMPUTING, 181, 53-63. http://doi.org/https://doi.org/10.1016/j.neucom.2015.08.097

26. Jovanović, D., Lipovac, K., Stanojević, P., & Stanojević, D. (2011). The effects of personality traits on driving-related anger and aggressive behaviour in traffic among Serbian drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 14(1), 43-53. http://doi.org/https://doi.org/10.1016/j.trf.2010.09.005

27. Kerimov, M., Safiullin, R., Marusin, A., & Marusin, A. (2017). Evaluation of Functional Efficiency of Automated Traffic Enforcement Systems. Transportation Research Procedia, 20, 288-294. http://doi.org/https://doi.org/10.1016/j.trpro.2017.01.025

28. Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. APPLIED SOFT COMPUTING, 83, 105662. http://doi.org/https://doi.org/10.1016/j.asoc.2019.105662

29. Lee, D., Derrible, S., & Pereira, F. C. (2018). Comparison of Four Types of Artificial Neural Network and a Multinomial Logit Model for Travel Mode Choice Modeling. TRANSPORTATION RESEARCH RECORD, 2672(49), 101-112. http://doi.org/10.1177/0361198118796971

30. Li, Y., Abdel-Aty, M., Yuan, J., Cheng, Z., & Lu, J. (2020). Analyzing traffic violation behavior at urban intersections: A spatio-temporal kernel density estimation approach using automated enforcement system data. Accident Analysis & Prevention, 141, 105509. http://doi.org/https://doi.org/10.1016/j.aap.2020.105509

31. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

32. Lipovac, K., Đerić, M., Tešić, M., Andrić, Z., & Marić, B. (2017). Mobile phone use while driving-literary review. Transportation Research Part F: Traffic Psychology and Behaviour, 47, 132-142. http://doi.org/https://doi.org/10.1016/j.trf.2017.04.015

33. Long, K., Liu, Y., & Han, L. D. (2013). Impact of countdown timer on driving maneuvers after the yellow onset at signalized intersections: An empirical study in Changsha, China. SAFETY SCIENCE, 54, 8-16. http://doi.org/https://doi.org/10.1016/j.ssci.2012.10.007

34. Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC BIOINFORMATICS, 18(1), 169. http://doi.org/10.1186/s12859-017-1578-z

35. Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research, 25, 100113.

http://doi.org/https://doi.org/10.1016/j.amar.2020.100113

36. Mårdh, S. (2016). Identifying factors for traffic safety support in older drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 38, 118-126. http://doi.org/https://doi.org/10.1016/j.trf.2016.01.010

37. Mostyn Sullivan, B., George, A. M., & Brown, P. M. (2021). Impulsivity facets and mobile phone use while driving: Indirect effects via mobile phone involvement. Accident Analysis & Prevention, 150, 105907. http://doi.org/https://doi.org/10.1016/j.aap.2020.105907

38. Oppenheim, I., Oron-Gilad, T., Parmet, Y., & Shinar, D. (2016). Can traffic violations be traced to gender-role, sensation seeking, demographics and driving exposure? Transportation Research Part F: Traffic Psychology and Behaviour, 43, 387-395. http://doi.org/https://doi.org/10.1016/j.trf.2016.06.027

39. Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., & Summala, H. (2006). Cross-cultural differences in driving skills: A comparison of six countries. Accident Analysis & Prevention, 38(5), 1011-1018. http://doi.org/https://doi.org/10.1016/j.aap.2006.04.006

40. Park, S. H., Park, S. H., Kwon, O. H., & Sung, Y. (2019). Continuous risk profile and clustering-based method for investigating the effect of the automated enforcement system on urban traffic collisions. The Journal of Supercomputing, 75(8), 4350-4371. http://doi.org/10.1007/s11227-019-02752-6

41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

42. Porter, B. E., Johnson, K. L., & Bland, J. F. (2013). Turning off the cameras: Red light running characteristics and rates after photo enforcement legislation expired. Accident Analysis & Prevention, 50, 1104-1111. http://doi.org/https://doi.org/10.1016/j.aap.2012.08.017

43. Reason, J., Parker, D., & Lawton, R. (1998). Organizational controls and safety: The varieties of rule-related behaviour. JOURNAL OF OCCUPATIONAL AND ORGANIZATIONAL PSYCHOLOGY, 71(4), 289-304. http://doi.org/https://doi.org/10.1111/j.2044-8325.1998.tb00678.x

44. Regan, M. A., Hallett, C., & Gordon, C. P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. Accident Analysis & Prevention, 43(5), 1771-1781. http://doi.org/https://doi.org/10.1016/j.aap.2011.04.008

45. Retting, R. A., Ferguson, S. A., & Hakkert, A. S. (2003). Effects of Red Light Cameras on Violations and Crashes: A Review of the International Literature. Traffic Injury Prevention, 4(1), 17-23. http://doi.org/10.1080/15389580309858

46. 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. Accident Analysis & Prevention, 127, 134-149. http://doi.org/https://doi.org/10.1016/j.aap.2019.02.008

47. Smith, D. M., McFadden, J., & Passetti, K. A. (2000). Automated Enforcement of Red Light Running Technology and Programs: A Review. TRANSPORTATION RESEARCH RECORD, 1734(1), 29-37. http://doi.org/10.3141/1734-05

48. Stringer, R. J. (2018). Exploring traffic safety culture and drunk driving: An examination of the community and DUI related fatal crashes in the U.S. (1993–2015). Transportation Research Part F: Traffic Psychology and Behaviour, 56, 371-380. http://doi.org/https://doi.org/10.1016/j.trf.2018.05.014

49. Stylianou, K., Dimitriou, L., & Abdel-Aty, M. (2019). Chapter 12 - Big Data and Road Safety: A Comprehensive Review. In C. Antoniou, L. Dimitriou, & F. Pereira (Eds.), Mobility Patterns, Big Data and Transport Analytics (297-343). Elsevier. http://doi.org/https://doi.org/10.1016/B978-0-12-812970-8.00012-9

50. Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. Accident Analysis & Prevention, 72, 244-256. http://doi.org/https://doi.org/10.1016/j.aap.2014.06.017

51. Xing, F., Huang, H., Zhan, Z., Zhai, X., Ou, C., Sze, N. N., & Hon, K. K. (2019). Hourly associations between weather factors and traffic crashes: Non-linear and lag effects. Analytic Methods in Accident Research, 24, 100109. http://doi.org/https://doi.org/10.1016/j.amar.2019.100109

52. Yahaya, M., Jiang, X., Fu, C., Bashir, K., &amp; Fan, W. (2019). Enhancing Crash Injury Severity Prediction on Imbalanced Crash Data by Sampling Technique with Variable Selection. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). https://doi.org/10.1109/itsc.2019.8917223

53. Ye, F., & Lord, D. (2011). Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit, and Mixed Logit. TRANSPORTATION RESEARCH RECORD, 2241(1), 51-58. http://doi.org/10.3141/2241-06

54. Yuan, J., Abdel-Aty, M., Gong, Y., & Cai, Q. (2019). Real-Time Crash Risk Prediction using Long Short-Term Memory Recurrent Neural Network. TRANSPORTATION RESEARCH RECORD, 2673(4), 314-326. http://doi.org/10.1177/0361198119840611

55. Zhang, G., Tan, Y., & Jou, R. (2016). Factors influencing traffic signal violations by car drivers, cyclists, and pedestrians: A case study from Guangdong, China. Transportation Research Part F: Traffic Psychology and Behaviour, 42, 205-216. http://doi.org/https://doi.org/10.1016/j.trf.2016.08.001

56. Zhang, G., Yau, K. K. W., & Chen, G. (2013). Risk factors associated with traffic violations and accident severity in China. Accident Analysis & Prevention, 59, 18-25. http://doi.org/https://doi.org/10.1016/j.aap.2013.05.004

57. Zhang, Q., Ge, Y., Qu, W., Zhang, K., & Sun, X. (2018). The traffic climate in China: The mediating effect of traffic safety climate between personality and dangerous driving behavior. Accident Analysis & Prevention, 113, 213-223. http://doi.org/https://doi.org/10.1016/j.aap.2018.01.031