Traffic & Transportation

A Quantitative Method for Assessing Freeway Driving Risk Based on Continuous Observation Data

Chonghao PANG[1], Peiqun LIN[2], Minping GONG[3], Qiang ZENG[4], Chuhao ZHOU[5]

Original Scientific Paper
Submitted: 23 Mar 2025
Accepted: 15 Oct 2025
Published: 29 June 2026

[1] 202110181554@mail.scut.edu.cn, School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[2] Corresponding author, pqlin@scut.edu.cn, School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[3] hn125985@outlook.com, Shenzhen Wanwuyun Technology Co., Ltd, Shenzhen, China
[4] zengqiang@scut.edu.cn, School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[5] zchuomi1017@scut.edu.cn, School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
Frequent freeway accidents cause significant casualties and economic losses, necessitating robust risk assessment methods. This study proposes a quantitative method for assessing freeway driving risk using continuous observational data from toll transactions. Based on toll data from the Yongguan Freeway in Guangdong Province, China (June–August 2022), 18 risk characteristic indicators for cargo vehicles and 13 for passenger vehicles were developed. Factor analysis reduced these indicators into five common factors, followed by K-means++ clustering to categorise vehicles into risk groups. The entropy weight method calculated risk scores, determining risk levels. The model identified 17.75% of cargo vehicles as high-risk and 14.03% as moderately high-risk, and 7.47% of passenger vehicles as high-risk and 1.08% as moderately high-risk. Validation using rescue events per 10,000 vehicles (RM) from a Guangdong Province accident database, due to limited crash data availability, confirmed consistency with model-assigned risk levels, supporting targeted safety interventions.

KEYWORDS
driving risk classification; driving risk features; vehicle user profile; freeway toll data; dimensionality reduction.

1. INTRODUCTION

Traffic crashes disrupt traffic operations, damage traffic flow and cause serious urban problems worldwide, resulting in significant economic losses to individuals, families and the entire country, accounting for 3% of the gross domestic product in most countries. Quantifying the risks associated with highway driving allows researchers to identify and analyse the various factors that contribute to crashes and incidents on the road. This includes factors such as road conditions, weather, driver behaviour, vehicle characteristics and traffic patterns. By quantifying these risks, researchers can gain a deeper understanding of their individual and collective impacts on highway safety. Quantitative risk assessment enables comparative analysis and benchmarking across different regions, road networks or time periods. By quantifying risks consistently and employing standardised methodologies, researchers can compare risk levels across various contexts. This comparative analysis facilitates the identification of best practices, the sharing of knowledge and experiences, and the development of comprehensive strategies that have proven effective in different settings.

According to statistics, road traffic crashes are the leading killer of children and young people aged 5 to 29 worldwide, with an estimated one person dying from traffic crashes every 24 seconds [1]. In the past five years, China's freeway mileage has increased by more than 4% annually, currently reaching 177,300 kilometres,

ranking first in the world, but the accompanying traffic crashes have also become more and more serious. In 2021, there were about 238,000 people injured and about 62,000 people killed in traffic crashes in China, resulting in a direct property loss of 1.45 billion yuan [2]. Since 1988, more than 5 million car crashes have occurred in the United States every year, of which about 30% result in death and injury [3], and in 2021, there were 38,824 car crashes on highways, accounting for 96.57% of the total number (40,205) of road traffic fatalities in the country [4]. In Canada, in 2021, the death rate per 100,000 population increased to 4.7 people, and the death rate per billion vehicle kilometres travelled increased to 4.8 [5]. Frequent highway traffic crashes threaten the safety of people's lives and property and limit the expected performance of highways.

Researchers from various countries have also proposed many methods to control traffic crashes, mainly using the characteristics of people, vehicles, roads and the environment to quantify driving risk. The aim of driving risk quantification is to use mathematical results to describe the crash risk of vehicles on the road. Driving risk quantification can establish risk assessment indicators based on traffic crashes, vehicle driving risk assessment models that consider various factors, and vehicle driving risk analysis that considers driver behaviour at macro, middle and micro levels.

After years of theoretical and practical developments, the research on highway driving risk assessment has yielded fruitful results. However, the existing studies still have certain limitations, which are mainly reflected in the following aspects:

1) The selection of assessment indicators has a certain degree of randomness. Highway driving risk assessment mainly uses data related to traffic density and speed, or historical crash data [6–8], to extract representative data, but there has been no large-scale quantitative analysis of the degree of influence of each indicator on crash risk assessment.
2) Risk assessment is often based on a single-vehicle perspective [9, 10]. Freeway vehicle risk assessment is primarily employed for collision warning. These studies are grounded in the analysis of actual driving vehicles and their corresponding road environments, evaluating the collision probability for vehicles of the same type when subsequently traversing the same or similar road segments. The focus is predominantly on a single vehicle type, with a notable absence of comprehensive risk assessment encompassing multiple vehicle types.

This study focuses on the risk assessment of different vehicle models on highways. Through factor analysis and clustering algorithms, it selects driving risk indicators that are well-aligned with vehicle models. Subsequently, the risk values of different highway vehicle models are quantitatively calculated, thereby obtaining a ranking of highway vehicle model risks. The main objective of this paper is to quantify the risks of highway vehicle models, with key innovations and improvements in the following aspects:

1) The study comprehensively selected 32 driving characteristic indicators for both passenger and commercial vehicles, and used factor analysis to identify the risk indicators that are closely related to highway driving risks. Through factor analysis, the selected indicators ensure universality and representativeness, which can serve as the data foundation for the highway risk assessment system.
2) By utilising clustering algorithms, the study constructed driver behaviour user profiles, uncovering the correlations between key indicators and vehicle types. This expands the vehicle risk assessment from a single-vehicle perspective to a broader vehicle type-level analysis, providing a robust theoretical foundation even in the face of data limitations.
3) The study proposes a quantitative method of freeway driving risk assessment based on continuous observation data. It calculates the probability of crash occurrence for different vehicle types on highways and compares the results with historical crash data. The ranking of vehicle types derived from the two approaches fully aligns with each other. The obtained results not only can serve as the basis for highway vehicle guidance, but also can provide the data foundation for other real-time vehicle risk assessment applications.

The remainder of this paper is organised as follows: Section 2 reviews the literature on vehicle model risk assessment and driver behaviour profiling. Section 3 details the construction process and methodological framework of the proposed highway vehicle risk assessment method. Section 4 presents a quantitative method for freeway driving risk assessment based on continuous observational data, supported by an example analysis, and evaluates the effectiveness of the proposed methods through comparison and discussion. Section 5 summarises the study and outlines directions for future research.

2. LITERATURE REVIEW

Traffic crash rate and crash frequency are commonly used quantitative indicators to evaluate traffic crash risk. Basu et al. [6] used a traffic crash rate indicator to evaluate highway safety under mixed traffic conditions. Anastasopoulos et al. [11] used the Tobit regression model to directly analyse the impact of various factors of time-continuous data on the crash rate on highway sections. Directly using the traffic crash rate and crash frequency can intuitively evaluate the design level of a road, but it cannot prevent or predict the real-time driving risk.

In naturalistic driving research, a suite of sensors collects real-time data on vehicle kinematics and environmental interactions, facilitating the quantitative assessment and prediction of vehicle trajectories to reduce highway collision risks. The vehicle driving risk assessment models consider various factors which generally include factors such as road conditions, vehicle speed, traffic flow and driver behaviour. At present, machine learning algorithms are usually used to integrate the weights of influencing factors and give a calculation method for predicting the driving risk of different vehicle types on specific road sections [9, 10]. Since the impact of various factors can be considered, the XGBoost algorithm has become an important method for building crash detection and prediction models [12, 13]. Liu et al. [1] proposed a novel collision risk assessment framework based on transfer learning, which is convenient for finding out the measurement criteria of data similarity and helps to adapt the model to other roads and periods. Mahmud et al. [14] proposed an evaluation model of risk factors for highway crashes under mixed traffic conditions, and Koçar et al. [15] based on a fuzzy inference system, proposed a traffic safety evaluation model including parameters such as vehicle speed, crash frequency, weather conditions, tyre tread depth and fatigue degree. The model design must balance data completeness with ensuring broad applicability across diverse scenarios.

Driving behaviour analysis assesses driver risk by evaluating key indicators such as speeding frequency, nighttime driving and fatigue-related behaviours. Naturalistic driving research enhances this process by collecting real-time data on driver actions and traffic conditions using sensors. These data enable detailed analysis of parameters, including acceleration, braking, steering, speed, lane-keeping and attention allocation, supporting the quantitative assessment of driving risks on freeways [16]. These data can be used to investigate driver responses to different traffic situations, reveal characteristics and variations in driving behaviour, and assess driver attention and emotional states. Naturalistic driving research helps evaluate the impact of driving behaviour on traffic safety. By analysing driving data collected on actual roads, researchers can identify high-risk driving behaviours and traffic conflicts, and study their relationship with crash risk. This contributes to understanding the causes of traffic crashes and the contribution of driving behaviour to crash risk, providing scientific evidence for traffic safety policies and interventions. [17] established a driver visual lane model to quantify the driver's visual perception, and formed a probabilistic neural network (PNN) to identify crash-prone points on double-lane mountain roads. It can be used to evaluate the candidate design of roads without historical traffic crash data. Wen et al. [18] found that the driver's familiarity with the road would significantly affect the crash rate. The less familiar a driver is with the road, the more they are influenced by external factors such as weather, road conditions, time of day, lighting and others. Zhao et al. [19] established a polynomial Logit model to explore the impact of various factors on traffic safety risk levels through driving behaviour data. Zhang et al. [20] evaluated the risk factors in the vehicle driving process by analysing the driver's driving style. The driver's driving behaviour is a direct factor leading to the crash rate. The driver's driving behaviour is not only affected by real-time road and environmental factors, but also by some factors other than driving behaviour, such as age, family, physical fitness, driving vehicle type, etc., which will have an impact.

The method that quantifies vehicle risk types based on vehicle characteristics can provide data-driven insights for mitigating vehicle collisions. However, it is noteworthy that driving risk quantification methods typically rely on specific road conditions, assessing the overall environment after a vehicle enters the road to evaluate driving risks. This process demands high timeliness and often uses road segments as the evaluation unit, influenced by multiple factors. Therefore, this study proposes a method to extract vehicle characteristics from repeated driving behaviours over a period, utilising data such as travel time, payload, vehicle speed and origin-destination (OD) patterns.

The concept of vehicle feature extraction in this paper is derived from the user portrait. Cooper [21] first proposed the concept of user persona, which is a virtual representation of real users, facilitating the analysis and acquisition of user characteristics. For the specific connotation of user persona, scholars at home and abroad have different understandings, but its essence is consistent, that is, based on the available massive data, user-centred, according to their basic information, behaviour information and other dynamic attributes for classification, and then combined with certain data mining methods, such as logistic regression, decision tree, support vector machine, Bayesian network, k-means algorithm, etc., to extract user characteristics from them. After further abstraction and refinement, the user characteristics are labelled, thus obtaining a complete user characteristic portrait [22–24]. Clerck et al. [25] established a car owner portrait to evaluate the total social cost of vehicle purchase and use. Singh et al. [26] presented a systematic review of studies on driver behaviour profiling and found that speed, acceleration, braking, position, mileage and time changes were significantly correlated with collision risk. Maier et al. [27] analysed the relationship between personality traits and mobile phone use while driving. Due to the relatively late implementation of networked highway toll collection, vehicle profiling predominantly relies on data sources such as GPS trajectory data, on-board diagnostic equipment data, vehicle registration data, operational platform data and traffic violation records, with limited utilisation of networked highway toll collection data. From a research perspective, vehicle profiling studies primarily focus on analysing general driving behaviour characteristics, while investigations into vehicle driving risk characteristic profiling remain underexplored.

Given the limitations of existing studies on highway driving risk assessment, this study aims to systematically quantify the driving risks of different vehicle models using factor analysis and clustering algorithms. Based on the research background and objectives, the following hypotheses are proposed. First, the 32 driving characteristic indicators selected through factor analysis can significantly differentiate the highway driving risks of various vehicle models, exhibiting high universality and representativeness. Second, driver behaviour profiles constructed using clustering algorithms can reveal the associations between key risk indicators and vehicle models, thereby extending risk assessment from a single-vehicle perspective to a vehicle model-level analysis. Finally, the quantitative risk assessment method based on continuous observational data can accurately calculate the crash probabilities of different vehicle models on highways, with results demonstrating high consistency with historical crash data. These hypotheses will be empirically tested to provide theoretical and data support for highway vehicle risk assessment and real-time applications.

3. METHODOLOGY

3.1 Data source

This paper uses the toll transaction data of all vehicles that passed through the Changping-Guanzhang section of the Guangdong Yongguan freeway from 12 June to 20 August 2022 as the experimental data. Due to issues related to hardware, software or human factors, toll revenue data from highways often contain anomalies or missing values. To accurately demonstrate the effectiveness of the proposed model, this paper removes erroneous data, including records with incorrect licence plate numbers, entry gross weights (for cargo vehicles) less than 1,000 kg, travel distances less than 1 km or greater than 2,100 km, travel times less than 1 minute or greater than 1 day, and travel speeds less than 2 km/h or greater than 160 km/h. After data preprocessing, 55,524,158 records were retained, including toll records of 300,750 cargo vehicles and 1,530,294 passenger vehicles. The cleaned data are utilised for subsequent metric calculation and analysis. The relevant data have been anonymised.

3.2 Methodological framework

The objective of this paper is to recognise the driving risk level of vehicles based on the freeway network toll data. We construct a vehicle feature profile label system for driving risk and perform vehicle clustering based on this system. Then, we use the entropy weight method to construct a vehicle driving risk level recognition model based on the driving risk feature profile. The framework structure of this study is shown in *Figure 1,* and the specific process consists of the following steps:

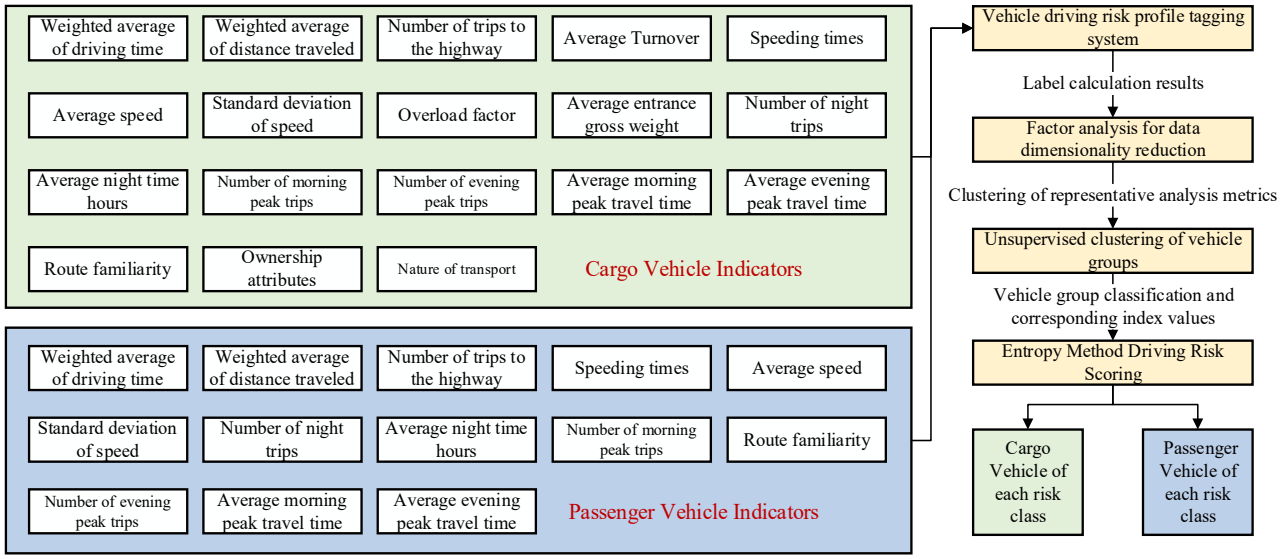

Figure 1 – Flow chart of freeway driving risk index calculation method based on continuous observation data

1) Factor analysis is conducted on the driving risk characteristic variables of cargo vehicles and passenger vehicles to derive factor scores for each vehicle. Factors with high scores are selected for their strong explanatory power, enabling dimensionality reduction of the overall risk characteristics for cargo vehicles and passenger vehicles. The detailed method is described in 3.3 and 3.4.
2) Using Python, the factor scores of cargo vehicles and passenger vehicles are used as the input data for clustering methods, respectively. This paper will use the K-means++ clustering method, confirm the best clustering number by the elbow method, and achieve the classification effect of vehicle group profiles for cargo vehicles and passenger vehicles. The detailed method is described in 3.5.
3) In order to improve the objectivity in defining user categories, the entropy weight method is used to score the risk of cargo vehicles and passenger vehicles. By calculating the entropy value and weight of the indicator data, the comprehensive score of each indicator is obtained. Combined with the relevant data indicator level of each category of vehicles, the differentiated driving risk feature performance of each category of vehicles is obtained, and finally, the risk level of each category of vehicles is determined. The detailed method is described in 3.6.
4) Using the crash database data of the same period to verify the model, comparing the predicted risk level of each category of vehicles in the model with the proportion of each category of vehicles in the crash database of the same period, to evaluate the accuracy and credibility of the model.

3.3 Construction of vehicle driving risk feature profile label system

The vehicle driving risk characteristic portrait label mainly includes the following aspects: driving intensity-related, speed-related, load-related (cargo vehicle specific), night driving-related, peak travel-related, spatial characteristics-related, vehicle attribute-related (cargo vehicle specific), etc. This paper constructs 18 vehicle driving risk characteristic labels for cargo vehicles and 13 vehicle driving risk characteristic labels for passenger cars.

1) Driving intensity-related

The single-road environment and less traffic interference on the freeway make the driver prone to fatigue, lack of concentration and reduced reaction ability. Fatigue driving as a result of long driving hours, long shifts and rigid schedules has been established as an important risk factor for road traffic crashes and fatalities [28, 29].

For the evaluation of driving intensity, this paper proposes a weighted average of driving duration $E_1(i)$, weighted average of driving distance $E_2(i)$, the number of trips on the freeway $E_3(i)$ and the average turnover $E_4(i)$ (for cargo vehicles) are used as four label indicators to measure the driving intensity. The driving duration weight coefficient TW_{ij} is determined by the single trip time t_{ij}, and the driving distance weight coefficient SW_{ij} is determined by the single trip distance s_{ij}.

The Regulations for the Implementation of the Road Traffic Safety Law of the People's Republic of China [30] stipulate that the driver should stop and rest after driving a motor vehicle for more than 4 hours

continuously, so the trips with driving time t_{ij} between 1 h and 4 h are assigned a weight coefficient TW_{ij} of 1, and those less than 1 h are assigned 0.5. According to the driving characteristic evaluation results of drivers after driving for a period T_c obtained from relevant studies, considering the situation of double drivers for medium and long-distance vehicles, the time in the evaluation results is multiplied by 2, corresponding to different weight levels, and finally, the weights are assigned according to the evaluation results scores. In addition, according to relevant data, the urban commuting radius generally does not exceed 50 km, the average transport distance of freeway freight transport in 2021 is 176 km, and referring to the above weight grading level difference, corresponding weights are assigned to the transport distance. The driving duration weight coefficient TW_{ij} and transport distance weight coefficient SW_{ij} are shown in *Table 1*.

The calculation method of driving duration weighted average $E_1(i)$ is shown in *Equation 1*, and the calculation method of driving distance weighted average $E_2(i)$ is shown in *Equation 2*.

$$E_1(i) = \frac{1}{N_i}\sum_{j=1}^{N_i}(t_{ij} \cdot TW_{ij}) \tag{1}$$

$$E_2(i) = \frac{1}{N_i}\sum_{j=1}^{N_i}(s_{ij} \cdot SW_{ij}) \tag{2}$$

The turnover of cargo vehicles is one of the important indicators of the development of the transportation industry. It can reflect the development level and industrial scale of the transportation industry. For individual cargo vehicles, the turnover can be used to describe the transportation intensity of cargo vehicles. The calculation equation of average turnover $E_4(i)$ is shown in *Equation 3*.

$$E_4(i) = \frac{1}{N_i}\sum_{j=1}^{N_i}(s_{ij} \cdot EW_{ij}) \tag{3}$$

where $E_4(i)$ is the average turnover of the $i-th$ vehicle, N_i is the number of trips of the $i-th$ vehicle, s_{ij} is the driving distance of the $j-th$ trip of the $i-th$ vehicle, and EW_ij is the total weight of the entrance of the $j-th$ trip of the $i-th$ vehicle.

2) Speed-related

Vehicle speed is one of the important criteria for measuring a driver's driving safety. The narrowing of driving sight and the lengthening of braking distance caused by excessive speed are the main causes of serious traffic crashes [31]. Based on the freeway toll data, this paper proposes three label indicators: vehicle speed means $E_5(i)$, vehicle speed standard deviation $E_6(i)$ and overspeed times $E_7(i)$. The vehicle speed mean indicates the speed of the vehicle, the standard deviation reflects the driver's stability in controlling the vehicle speed, and the overspeed times are used to evaluate the vehicle's overspeed behaviour. The calculation equations of vehicle speed mean and vehicle speed standard deviation are *Equations 4 and 5*.

$$E_5(i) = \frac{1}{N_i}\sum_{j=1}^{N_i} v_{ij} \tag{4}$$

$$E_6(i) = \sqrt{\frac{\sum_{j=1}^{N_i}(v_{ij}-\bar{v}_i)^2}{N_i}} \tag{5}$$

where $E_5(i)$ is the speed mean of the $i-th$ vehicle, N_i is the number of trips of the $i-th$ vehicle, v_{ij} is the driving speed of the $j-th$ trip of the i-th vehicle, and $E_6(i)$ is the speed standard deviation of the $i-th$ vehicle.

The traffic management department often uses the 85% percentile speed as the speed limit for some sections, that is, among all vehicles driving in this section, 85% of vehicles have a driving speed below this speed, and only 15% of vehicles have a driving speed higher than this value. Considering that the toll data obtained are the driving speed, this study uses the 85% percentile speed of all driving speeds in the sample as the overspeed threshold v_m, and identifies the trips with driving speed v_{ij} exceeding this threshold as overspeeding, and then counts the number of overspeeding times $E_7(i)$ for each vehicle.

3) Load-related (for cargo vehicles)

Overloading and over-limit transportation is an extremely dangerous behaviour, which often leads to traffic crashes. Studies have shown that for a cargo vehicle with a speed of 50 km/h when overloaded by 10%, the vehicle's emergency braking distance and risk factor will increase by 80%.

Based on the freeway toll data, this paper proposes two label indicators: overload coefficient $E_8(i)$ and average entrance total weight $E_9(i)$, which are used to evaluate the load behaviour of cargo vehicles. First, calculate the ratio k of the entrance total weight EW_{ij} and the approved total weight W_i'. If k is greater than 1, the trip Tr_{ij} is 1, otherwise it is 0. Second, according to the level of k, assign a weight TrW_{ij} to the trip Tr_{ij}. Finally, calculate the ratio of the weighted overload times of each cargo vehicle to the number of trips N_i on the freeway during the statistical time, and the result is the cargo vehicle overload coefficient $E_8(i)$, calculated using *Equations 6 and 7*.

$$k = \frac{EW_{ij}}{W_i'} \tag{6}$$

$$E_8(i) = \frac{1}{N_i}\sum_{j=1}^{N_i}(Tr_{ij} \cdot TrW_{ij}) \tag{7}$$

where the weight of overload times TrW_{ij} is based on the deduction standard stipulated in the Administrative Measures for Scoring Management of Road Traffic Safety Violations[32] according to the percentage of the load exceeding the maximum allowable total mass of driving cargo vehicles, as shown in *Table 1*.

Table 1 – Weight grading of driving duration, transport distance and overload times

$\mathbf{TW_{ij}}$				$\mathbf{SW_{ij}}$		$\mathbf{TrW_{ij}}$			
$2T_c$	rate	rank	weight	rank	weight	overload ratio	deduction of points	rank	weight
4 h	0.83	<1 h	0.5	<50 km	0.5	< 30%	1	<0.2	1
		[1 h,4 h)	1	[50 km,100 km)	1	30%~50%	3	[0.2,0.4)	1.5
5 h	0.73	[4 h,6 h)	1.5	[100 km,200 km)	1.5	>50%	6	[0.4,0.6)	2
7 h	0.53	[6 h,8 h)	2	[200 km,300 km)	2			≥0.6	2.5
9h	0.39	[8 h,10 h)	2.5	[300 km,400 km)	2.5				
		≥10 h	3	≥400 km	3				

Traffic crashes caused by heavy cargo vehicles are particularly serious on freeways. Large cargo vehicles have many blind spots and large inertia, and cause great harm and a high mortality rate when crashes occur. This paper uses the average total weight at the entrance $E_9(i)$ as a measure of the impact of cargo vehicle weight on pavement and driving safety. The calculation equation of the average total weight at the entrance $E_9(i)$ is given by *Equation 8*.

$$E_9(i) = \frac{1}{N_i}\sum_{j=1}^{N_i} EW_{ij} \tag{8}$$

4) Night driving-related

The Regulations on the Implementation of the Road Traffic Safety Law of the People's Republic of China [30]suggest that driving time at night should not exceed 23:00 as much as possible. Driving at night is more dangerous than during the day due to factors such as narrow driving vision, reduced vision, increased blind spots, drowsiness, difficulty concentrating, etc. This paper defines the time of vehicle driving at night as 0:00-5:00 in the morning. Based on freeway toll data, this paper proposes two label indicators, namely the number of night driving $E_{10}(i)$ and the average night driving duration $E_{11}(i)$, to evaluate the driver's night driving behaviour. Among them, the calculation method of the number of night driving $E_{10}(i)$ is to count the number of trips for which the night driving duration TN_{ij} of the $i-th$ vehicle is greater than 0. The night driving duration TN_{ij} is the time that the $i-th$ vehicle travels between 0:00 and 5:00 in the morning for the $j-th$ trip. The average night driving duration $E_{11}(i)$ is the average of the night driving duration of all trips. The calculation equation is *Equation 9*.

$$E_{11}(i) = \frac{1}{N_i}\sum_{j=1}^{N_i} TN_{ij} \tag{9}$$

where $E_{11}(i)$ represents the average night driving duration of vehicle i, N_i is the number of trips for vehicle i.

5) Peak travel-related

Traffic congestion is common during peak hours, when the distance between vehicles is shortened and crashes such as rear-end collisions are more likely to occur. In addition, traffic congestion can increase driving stress and anxiety, which can further reduce the driver's attention and reaction speed. This paper defines the morning peak hours as 7:00-9:00 a.m. and the evening peak hours as 17:00-19:00.

Based on freeway toll data, this paper proposes four indicators, namely the number of morning peak trips $E_{12}(i)$, the number of evening peak trips $E_{13}(i)$, the average morning peak travel time $E_{14}(i)$ and the average evening peak travel time $E_{15}(i)$, to describe the peak travel behaviour. Among them, the calculation method of the number of morning peak trips $E_{12}(i)$ is to count the number of trips for which the morning peak driving duration TM_{ij} of the $i-th$ vehicle is greater than 0. The morning peak driving duration TM_{ij} is the time that the $i-th$ vehicle travels between 7:00 and 9:00 a.m. for the j-th trip. The average morning peak travel time $E_{14}(i)$ is the average of the morning peak driving duration of all trips. The calculation equation is *Equation 10*.

$$E_{14}(i) = \frac{1}{N_i}\sum_{j=1}^{N_i} TM_{ij} \tag{10}$$

The calculation method of the number of evening peak trips $E_{13}(i)$ is to count the number of trips for which the evening peak driving duration TE_{ij} of the $i-th$ vehicle is greater than 0. The evening peak driving duration TE_{ij} is the time that the $i-th$ vehicle travels between 17:00 and 19:00 or the $j-th$ trip. The average evening peak travel time $E_{15}(i)$ is the average of the evening peak driving duration of all trips. The calculation equation is *Equation 11*.

$$E_{15}(i) = \frac{1}{N_i}\sum_{j=1}^{N_i} TE_{ij} \tag{11}$$

6) Spatial characteristics-related

The driver's visual perception is an important factor that affects road familiarity. When drivers drive on familiar roads, they may exhibit automated or procedural driving behaviour, which means that their brains automatically control the vehicle's movement without much cognition and thinking. This automated driving behaviour may lead to driver distraction and overconfidence, which can increase the risk of driving crashes. However, when drivers drive on unfamiliar roads, they may also feel uneasy and stressed, and make driving mistakes. Therefore, these familiarity factors should be included in the scope of traffic safety.

This paper proposes a road familiarity degree $E_{16}(i)$ indicator to measure road familiarity degree. $E_{16}(i)$ uses information entropy to calculate, which describes the uncertainty of an event. Entropy is a measure of how uncertain a random variable is, and it is the expectation of the information quantity produced by all possible events. First, calculate the frequency p_{ij} of OD pairs appearing, and then use *Equation 12* to calculate information entropy.

$$E_{16}(i) = -\sum_{j=1}^{NO_i} p_{ij}\, log_2 p_{ij} \tag{12}$$

where $E_{16}(i)$ represents the road familiarity degree of vehicle i, NO_i is the number of OD pairs for vehicle i, p_{ij} is the frequency of OD pair j for vehicle i.

7) Vehicle property-related (for cargo vehicles)

Private property refers to vehicle ownership belonging to an individual or enterprise, while public property refers to vehicle ownership belonging to a public institution or collective ownership. The private and public property of vehicles has complex effects on traffic safety. Private vehicles are used under the supervision of drivers and are easier to maintain and use, but drivers may ignore traffic rules and safety. Public vehicles may be used by multiple drivers, who may not care about vehicle maintenance and use, but these vehicles are subject to public institution supervision to ensure vehicle safety. In addition, cargo vehicles with a load capacity of more than 12 tons and cargo vehicles transporting dangerous goods are relatively large, slow-moving, have a long braking distance and require more road space, which can cause traffic congestion and inconvenience and safety hazards for other vehicles. To ensure traffic safety, these large cargo vehicles and dangerous goods transport vehicles need more strict management and supervision, and drivers and vehicles also need to strictly comply with traffic regulations and safety standards to avoid traffic crashes. Therefore, this paper constructs two indicator labels: owner property $E_{17}(i)$ and transport nature property $E_{18}(i)$. All symbols and their corresponding meanings are presented in *Table 2*.

Table 2 – Data fields corresponding to label indicators

Cargo vehicle	Passenger vehicle	Symbols
Weighted average of driving time	Weighted average of driving time	$E_1(i)$
Weighted average of driving distance	Weighted average of driving distance	$E_2(i)$
Number of trips to the freeway	Number of trips to the freeway	$E_3(i)$
Average turnover	/	$E_4(i)$
Average speed	Average speed	$E_5(i)$
The standard deviation of speed	The standard deviation of speed	$E_6(i)$
Number of speeding trips	Number of speeding trips	$E_7(i)$
Overload factor	/	$E_8(i)$
Average entrance gross weight	/	$E_9(i)$
Number of night trips	Number of night trips	$E_{10}(i)$
Average nighttime duration	Average nighttime duration	$E_{11}(i)$
Number of morning peak trips	Number of morning peak trips	$E_{12}(i)$
Number of evening peak trips	Number of evening peak trips	$E_{13}(i)$
Average morning peak travel time	Average morning peak travel time	$E_{14}(i)$
Average evening peak travel time	Average evening peak travel time	$E_{15}(i)$
Route familiarity	Route familiarity	$E_{16}(i)$
Ownership attributes	/	$E_{17}(i)$
Whether it is a cargo vehicle with a load of 12 tons or more	/	$E_{18}(i)$

3.4 Data dimensionality reduction

Factor analysis can reduce a large number of observed variables into a few factors, which can explain most of the variation in the original data, and help us discover the underlying structure and relationships behind the data. Let the original variables be $X_1, X_2, X_3 \ldots X_n$. Through factor analysis, the original variables can be explained from m aspects, and there are m common factors, namely $F_1, F_2, F_3 \ldots F_m$. The relationship between the original variable X and the common factor F can be expressed as *Equation 13.*

$$X = \Lambda F + \epsilon \tag{13}$$

where Λ is the factor loading matrix, F is the common factor vector, ϵ is the unique factor vector. F and ϵ are statistically independent.

The main idea of factor analysis is twofold: one is the selection of common factors, and the other is the interpretation of factors. The basic steps are as follows:

1) Data standardisation, to eliminate the differences in magnitude and dimension among variables.
2) Use the KMO test and Bartlett's test of sphericity to test the suitability of factor analysis. Factor analysis requires that the KMO value is generally greater than 0.5, and the significance level of Bartlett's test reaches 0.05.
3) Common factor selection, based on the eigenvalue size and cumulative variance contribution rate, to determine the common factors. Usually, factors with eigenvalues greater than 1 and a cumulative variance contribution rate greater than 80% are selected. Determine the final number of factors to retain, in order to explain most of the variance and covariance structure of the data.
4) Factor rotation and interpretation; rotate the factor loading matrix to make the relationship between each variable and each factor clearer and more intuitive while reducing the correlation between factors. Factor rotation can make factor interpretation avoid subjective factors and make interpretation more favourable. It mainly differentiates the squares of factor loadings in the factor loading matrix towards 0 and 1, making the loadings more distinct. In practical applications, there are mainly two ways of rotation: orthogonal and oblique.

5) Calculate the scores of each factor, which can be done by using a linear combination of the original indicators. Factor analysis is often used for dimensionality reduction, data cleaning, exploring latent structures and relationships, and extracting important features.

3.5 Cluster classification of vehicles using an unsupervised clustering method

This paper uses a partition-based clustering algorithm to classify the users. K-means clustering, the most classic one of the partition clustering algorithms, is widely used in current data clustering research. Compared with the K-means algorithm, the K-means++ algorithm has better clustering performance and scalability [33], and is more intelligent and reasonable in selecting initial centre points. Therefore, it is more commonly used in practical applications. The K-means++ clustering algorithm needs to determine the optimal number of clusters k, which ensures high intra-cluster similarity and low inter-cluster similarity. This paper will use the elbow method to ensure the selection of the optimal k value.

3.6 Driving risk scoring with the entropy method

The concept of "entropy" was originally a physical parameter in thermodynamics. In 1988, Shannon combined the concept of entropy with information theory and proposed the theory of the "entropy weight method". In this method, researchers can use entropy values to judge the degree of dispersion of a certain indicator variable, which is negatively correlated. Therefore, information entropy can be used to calculate the weight of each indicator variable and provide a scoring basis for multi-objective comprehensive evaluation.

The basic principle of the entropy weight method is to determine the target weight according to the variability of the indicators, which is more objective than expert scoring. It has been widely used in multi-objective system evaluation. Therefore, this paper will use the entropy weight method to calculate the weight W_j and score the cargo vehicle operation risk with *Equations 14–18*.

1) Normalise the vehicle risk indicators with different units using *Equation 14*.

$$a_{ij}' = 0.998 \cdot \frac{a_{ij} - min(a_j)}{max(a_j) - min(a_j)} + 0.002 \tag{14}$$

2) Calculate the information entropy of the risk indicators using *Equation 15*.

$$E_j = -ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \, ln \, p_{ij} \tag{15}$$

$$p_{ij} = \frac{a_{ij}'}{\sum_{j=1}^{n} a_{ij}'} \tag{16}$$

3) Calculate the weight of each indicator using *Equation 17*.

$$W_j = \frac{1 - E_j}{\sum_{j=1}^{m} 1 - E_j} \tag{17}$$

4) Calculate the vehicle driving risk score using *Equation 18.*

$$Z_i = \sum_{j=1}^{n} a_{ij}' \, W_j \tag{18}$$

where a_{ij} is the $j-th$ indicator of the $i-th$ vehicle type, a_{ij}' is the normalised a_{ij}, E_j is the information entropy of $j-th$ indicator, n is the total number of vehicle type, p_{ij} is the output probability, W_j is the weight of the $j-th$ indicator, Z_i is the score of the $i-th$ vehicle type.

To address the issue of inherent correlations among indicators, this study applied factor analysis to reduce the dimensionality of 18 cargo vehicle indicators and 13 passenger vehicle indicators into five mutually independent factors before employing the entropy weight method. Factor analysis, utilising principal component extraction and orthogonal rotation, consolidated the common variance of the original indicators into orthogonal factors, effectively mitigating multicollinearity. The entropy weight method then calculated weights based on the data distribution of these factors. Due to their orthogonality, no further consideration of inter-factor correlations was required, allowing the method to focus on the information contribution of each factor.

4. CASE STUDY

4.1 Vehicle travel risk profile label calculation

Based on the freeway toll data, 18 label indicators for cargo vehicles and 13 label indicators for passenger vehicles were extracted. The label indicator data are shown in *Table 3* and *Table 4*.

Table 3 – Cargo vehicle labelling indicator data sheet (excerpt)

Vehicle ID	$E_2(i)$	$E_3(i)$	$E_5(i)$	$E_7(i)$	$E_8(i)$	$E_9(i)$	$E_{16}(i)$	$E_{17}(i)$	$E_{18}(i)$
1	51390	192	57.04	13	0.904	5030.6	7.218	0	0
2	44673	294	51.59	2	1.337	31545.1	4.520	0	1
3	40001	278	70.09	62	0.050	14919.1	5.085	0	1
4	24444	211	76.54	48	0.498	16352.4	4.769	0	1
5	6742	144	66.24	34	0.000	8169.3	3.618	0	0
6	44512	194	66.22	36	0.951	4997.7	6.670	1	0
7	47292	67	64.43	6	0.396	4067.0	5.058	1	0
8	39214	89	65.34	18	0.556	4333.7	3.730	0	0
9	380520	90	63.12	8	1.389	6511.9	6.247	0	0
10	126087	58	66.27	4	0.121	3408.8	5.638	1	0
11	42848	117	62.03	11	0.286	2674.1	6.264	1	0
12	48599	97	63.24	5	0.124	10250.3	5.576	1	0

Table 4 – Passenger vehicle labelling indicator data sheet (excerpt)

Vehicle ID	$E_1(i)$	$E_3(i)$	$E_5(i)$	$E_7(i)$	$E_{11}(i)$	$E_{12}(i)$	$E_{13}(i)$	$E_{15}(i)$	$E_{16}(i)$
1	3957	9	81.86	2	0	0	0	715	3.17
2	13896	55	77.64	4	318	11	539	1102	5.50
3	724	97	70.63	2	74	11	100	199	5.35
4	2416	70	89.89	17	121	5	132	706	4.06
5	802	151	60.28	7	3	38	67	170	3.38
6	1503	129	75.70	16	60	5	70	262	4.98
7	856	244	85.56	51	22	50	256	198	4.32
8	1222	74	75.71	5	38	1	3	308	1.39
9	1031	48	83.53	15	26	2	11	226	3.69
10	3557	58	87.83	18	574	1	9	436	3.81
11	678	21	86.51	7	75	0	0	191	2.01
12	3915	143	88.25	59	3	5	109	119	5.87

4.2 Clustering analysis of vehicle driving risk profile

The data used in this paper are the label indicator data of 300,750 cargo vehicles and 1,530,294 passenger vehicles in Section 4.1. In order to improve the clustering accuracy and avoid the impact of different orders of magnitude of each indicator on the clustering results, the Z-score algorithm was used to standardise the indicators first, and then the KMO test and Bartlett's test of sphericity were performed on the label indicators of cargo vehicles and passenger vehicles. Factor analysis requires that the KMO value is generally greater than 0.5, and the significance level of Bartlett's test reaches 0.05. The KMO values of cargo vehicles and passenger vehicles are 0.844 and 0.809, respectively, and the significance of Bartlett's spherical test is 0, which meets the requirements.

By analysing the eigenvalues of the correlation coefficient of explanatory variables, common factors were determined based on the eigenvalue size and cumulative variance contribution rate. As shown in *Figure 2* and *Table 5*, the eigenvalues of the first five factors for cargo vehicles are all greater than 1, with a cumulative variance contribution of 79.8%, approaching 80%. For passenger vehicles, the eigenvalues of the first five factors are significantly greater than 0.9, slightly below 1, yet they retain substantial explanatory power in practical analysis. Their cumulative variance contribution reaches 80.1%, exceeding 80%. Consequently, the first five factors were chosen as the results of dimensionality reduction in this study to reflect the meaning represented by the overall variable indicators of vehicle driving risk characteristics.

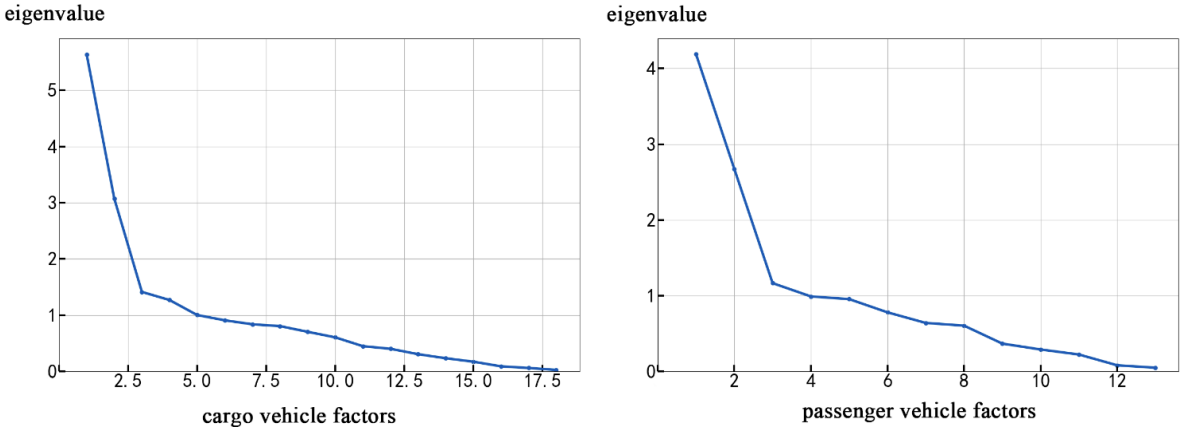

Figure 2 – Relationship between eigenvalues and changes in the number of factors

Table 5 – Cargo vehicle and passenger vehicle labelling indicator data sheet

Cargo vehicle				**Passenger vehicle**			
factors	Eigenvalue	Variance contribution rate	Cumulative variance contribution rate	factors	Eigenvalue	Variance contribution rate	Cumulative variance contribution rate
1	5.639	0.363	0.363	1	4.187	0.347	0.322
2	3.074	0.198	0.561	2	2.672	0.222	0.544
3	1.414	0.091	0.652	3	1.165	0.097	0.640
4	1.269	0.082	0.734	4	0.990	0.082	0.722
5	1.003	0.065	0.798	5	0.954	0.079	0.801
6	0.909	0.058	0.857	6	0.780	0.065	0.866
7	0.838	0.054	0.911	7	0.640	0.053	0.919
8	0.271	0.017	0.928	8	0.230	0.019	0.963
9	0.201	0.013	0.941	9	0.193	0.016	0.979
10	0.196	0.013	0.954	10	0.108	0.009	0.988
11	0.186	0.012	0.966	11	0.101	0.008	0.996
12	0.155	0.010	0.976	12	0.028	0.002	0.998
13	0.132	0.009	0.984	13	0.019	0.002	1.000
14	0.128	0.008	0.992	/	/	/	/
15	0.061	0.004	0.996	/	/	/	/
16	0.028	0.002	0.998	/	/	/	/
17	0.021	0.001	0.999	/	/	/	/
18	0.010	0.001	1.000	/	/	/	/

Due to the fact that there is no factor loading greater than 0.7 for some factors on some indicators, the interpretability of the actual meaning of indicators is not enough, so factor rotation is needed to enhance the explanatory ability of the first five factors for each indicator. According to the factor loading matrix of cargo vehicles' 18 indicators and passenger vehicles' 13 indicators on their respective first five factors, factor rotation was performed to make it easier to explain their positions, so as to better understand the structure and meaning of the data. In this paper, the variance orthogonal rotation method was used to rotate the factor loading matrix. The results are shown in *Figure 3.*

Based on the performance of cargo vehicles' five factors $(F1, F2, F3, F4, F5)$ and passenger vehicles' five factors $(Q1, Q2, Q3, Q4, Q5)$ in terms of representativeness and explanatory ability after factor rotation, it is proved that cargo vehicles' five common factors and passenger vehicles' five common factors are objectively feasible.

Figure 3 – Load matrix after rotation of cargo vehicle and passenger vehicle factors: a) cargo vehicle; b) passenger vehicle

The scores of cargo vehicles on cargo vehicles' common factors $(F1, F2, F3, F4, F5)$ and passenger vehicles on passenger vehicles' common factors $(Q1, Q2, Q3, Q4, Q5)$ were calculated respectively as inputs for K-means++ clustering algorithm. The factor score situation is shown in *Table 6.*

Table 6 – Public factor score table for vehicles (excerpt)

Cargo vehicle						**Passenger vehicle**					
Vehicle ID	F1	F2	F3	F4	F5	Vehicle ID	Q1	Q2	Q3	Q4	Q5
1	-0.308	1.695	-0.257	1.174	-0.736	1	-0.466	-0.482	0.288	0.179	0.001
2	-0.373	0.361	-0.676	2.443	0.932	2	1.291	0.317	-0.314	0.735	0.076
3	-0.088	0.604	0.793	1.100	0.586	3	1.740	-0.298	-0.802	0.105	0.218
4	-0.368	1.649	-1.275	0.685	-0.700	4	-0.740	-0.052	0.870	-0.565	0.440
5	0.062	4.392	-1.521	-0.266	0.816	5	1.552	-0.464	1.155	0.452	0.712
6	-0.382	0.074	-0.469	1.124	1.178	6	1.756	-0.424	-0.365	0.169	-0.427
7	-0.438	0.257	-0.560	-0.073	-0.428	7	4.624	0.153	-1.741	-0.049	-2.075
8	-0.585	-0.172	-0.860	0.339	-0.926	8	0.469	-0.529	-1.183	0.173	0.268
9	-0.621	2.591	2.269	1.196	-2.293	9	0.087	-0.524	-0.987	0.182	0.473
10	-0.603	-1.212	2.141	-0.044	0.467	10	-0.238	-0.255	0.202	-0.321	-0.005

Based on the elbow method, the optimal clustering effect is achieved when the number of clusters k for trucks and passenger cars is determined to be 5 each. Set $k = 5$ for the K-means++ clustering algorithm, and then combine it with the T-SNE dimensionality reduction algorithm for visualization. T-SNE algorithm first calculates the similarity between high-dimensional data points, usually using a Gaussian distribution to measure similarity between data points. Then it tries to find a set of new coordinates in low-dimensional space such that points that are similar in high-dimensional space are also similar in low-dimensional space, while dissimilar points are as far away as possible in low-dimensional space. The random sampling part of the samples' clustering effect is shown in *Figure 4*. The boundary between the cargo vehicles' five clusters and the passenger vehicles' five clusters obtained by the K-means++ clustering algorithm is obvious, and the clustering effect is good.

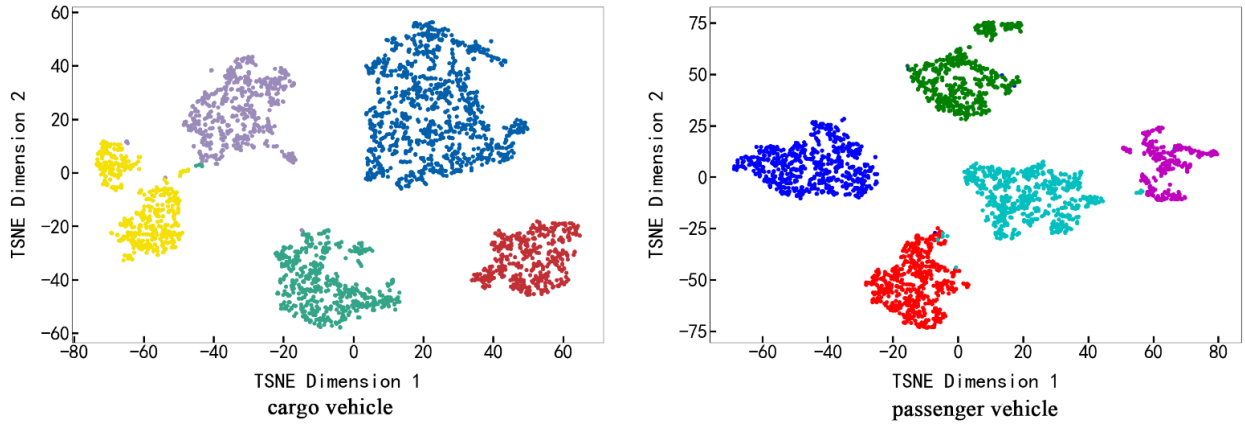

Figure 4 – The result of the T-SNE data dimensionality reduction algorithm on public factor clustering

According to clustering results, an average level of label indicators for each type of vehicle was counted to analyse the driving risk characteristics of each type of vehicle. Vehicle group segmentation results are shown in *Table 7* and *Table 8.* The number that ranks first for each metric across the five vehicle risk categories is highlighted in green.

Table 7 – Cargo vehicle clustering results

Indicators	Clustering results				
	Category 1	Category 2	Category 3	Category 4	Category 5
Number	42195	107154	53528	53379	44494
Weighted average of driving time (s)	72069.44	6305.62	4393.9	4019.5	12833.71
Weighted average of driving distance (km)	1104.05	117.43	99.05	75.4	204.23
Number of trips to the freeway	18.31	30.28	52.46	145.25	62.16
Average turnover	10896.67	823.57	382.86	488.43	3788.54
Average speed (km/h)	58.8	61.75	66.65	64.04	61.37
The standard deviation of speed	14.93	14.07	16.86	17.92	16.02
Number of speeding trips	1.32	3.43	10.56	24.69	6.62
Overload coefficient	1.14	0.4	0.8	0.58	1.31
Average entrance gross weight (kg)	22528.89	8130.76	5078.4	6879.24	30118.99
Number of night trips	7.72	1.93	3.88	9.44	11.98
Average nighttime duration (s)	6115.96	540.59	378.76	375.56	1610.85
Number of morning peak trips	6.6	3.44	5.52	17.68	9.86
Number of evening peak trips	6.55	5.11	8.89	25.18	9.2

Indicators	Clustering results				
	Category 1	**Category 2**	**Category 3**	**Category 4**	**Category 5**
Average morning peak travel time (s)	2198.51	298.14	247.29	271.98	577.24
Average evening peak travel time (s)	2147.23	471.43	415.86	390.19	530.62
Route familiarity	3.19	3.42	3.65	5.26	4.1
Percentage of private properties (%)	11.4	36.43	38.82	77.98	32.07
Percentage of cargo vehicles with a load capacity of 12 tons and above (%)	62.89	21.14	27.03	32.44	56.03
Vehicle share (%)	14.03	35.63	17.8	17.75	14.79

Based on relevant indicator data, the main driving risk characteristics of different vehicle categories for cargo vehicles were analysed. The category ranking based on the number of first-place rankings in 18 vehicle risk indicators is: Category 4 (7) = Category 1 (7) > Category 5 (3) > Category 3 (1) > Category 2 (0). This study analyses the vehicle composition of the five categories as follows:

— **Category 1**: Primarily consists of long-haul heavy-duty cargo vehicles, commonly utilised for cross-regional, long-distance, high-load transportation.
— **Category 2**: Primarily comprises medium-haul medium-duty cargo vehicles, typically employed for short-distance logistics of large cargo within urban areas, with moderate load capacity, short trip durations, infrequent freeway usage and no focus on transport efficiency.
— **Category 3**: Primarily includes short-haul light-duty cargo vehicles with limited load capacity and short transport distances, operating at high speeds. Other indicators are generally at low to medium levels, and these cargo vehicles are mainly used for transporting general cargo.
— **Category 4**: Primarily consists of short-haul medium-duty cargo vehicles, predominantly green-channel vehicles and small-scale urban delivery logistics vehicles, requiring high transport efficiency. As a result, they frequently access freeways even during peak hours.
— **Category 5**: Primarily comprises medium- to short-haul heavy-duty cargo vehicles, used for transporting extremely heavy cargo, mainly construction and industrial materials, within provincial boundaries.

Table 8 – Passenger vehicle clustering results

Indicators	Clustering results				
	Category 1	**Category 2**	**Category 3**	**Category 4**	**Category 5**
Number	675209	56569	114294	667624	16598
Weighted average of driving time(s)	3646.76	32622.8	1821.21	4896.92	83790.77
Weighted average of driving distance(km)	104.73	1023.6	60.43	191.94	1922.87
Number of trips to the freeway	22.23	3.24	112.71	15.07	2.31
Average speed (km/h)	69.98	81.82	77.96	85.45	77.97
The standard deviation of speed	21.49	8.66	21.07	14.56	7.11
Number of speeding trips	2.07	0.34	18.27	3.34	0.38
Number of night trips	1.13	0.26	5.23	1.01	1.19
Average nighttime duration (s)	180.98	530.58	77.63	238.57	7582.24
Number of morning peak trips	1.83	0.21	17.07	1.6	1.15
Number of evening peak trips	4.19	1.9	20.73	2.7	0.54
Average morning peak travel time (s)	156.35	192.6	205.97	276.19	3829.76
Average evening peak travel time (s)	438.77	3679.34	318.54	443.6	1576.84
Route familiarity	3.12	1.21	4.61	2.55	0.69
Vehicle share (%)	44.12	3.7	7.47	43.63	1.08

Based on relevant indicator data, the main driving risk characteristics of different vehicle categories for passenger vehicles were analysed. The category ranking based on the number of first-place rankings in 13 vehicle risk indicators is: Category 3 (6) > Category 5 (4) > Category 1 (1) = Category 4 (1) = Category 2 (1). This study analyses the vehicle composition of the five categories as follows:

— **Category 1**: Primarily consists of commuter-type small passenger vehicles, characterised by significant speed variability, unremarkable average speed and driving duration, high route familiarity and travel during peak hours.
— **Category 2 and Category 5**: Exhibit high similarity, primarily comprising long-haul passenger vehicles. The key difference is that Category 2 passenger vehicles have significantly lower driving intensity compared to Category 5. Category 2 passenger vehicles frequently travel on highways in the evening, while Category 5 passenger vehicles typically operate across days, with travel times including morning and evening peak hours. Consequently, drivers of Category 5 passenger vehicles are more prone to fatigue.
— **Category 3**: Primarily consists of short-haul commercial passenger vehicles with extremely high route familiarity. To ensure the timeliness of operational journeys, these passenger vehicles travel during morning and evening peak hours, frequently use highways and exhibit frequent speeding behaviour.
— **Category 4**: Shares high similarity with Category 1 but is characterised by longer driving times and higher speeds, with slightly lower route familiarity compared to Category 1. It primarily comprises tidal commuter-type small passenger vehicles.

4.3 Risk level identification for the vehicle

The data used are the common factor score data of the above text. The entropy weight method was used to calculate the indicator weights of cargo vehicles' five common factors and passenger vehicles' five common factors, respectively. The results are shown in *Table 9*. The weight results are consistent with the actual situation of freeway vehicles. Then, *Equations 14–18* were used to calculate the freight risk score of each driver. The results are shown in *Figure 5*.

Table 9 – Indicator weights

Cargo vehicle			Passenger vehicle		
Common factor	weight	Main explanatory label indicators	Common factor	weight	Main explanatory label indicators
F1	0.270	Travel time, travel distance, turnover, night and peak driving time	Q1	0.303	Number of trips to the freeway, peak time travelling
F2	0.320	Speeding times, number of trips to the freeway	Q2	0.133	Travelling time, travelling distance
F3	0.209	Overload factor, average entrance gross weight, 12 tons and above type	Q3	0.180	Average speed
F4	0.166	Average speed, route familiarity	Q4	0.150	Peak driving time
F5	0.135	Ownership attributes	Q5	0.233	Speeding times

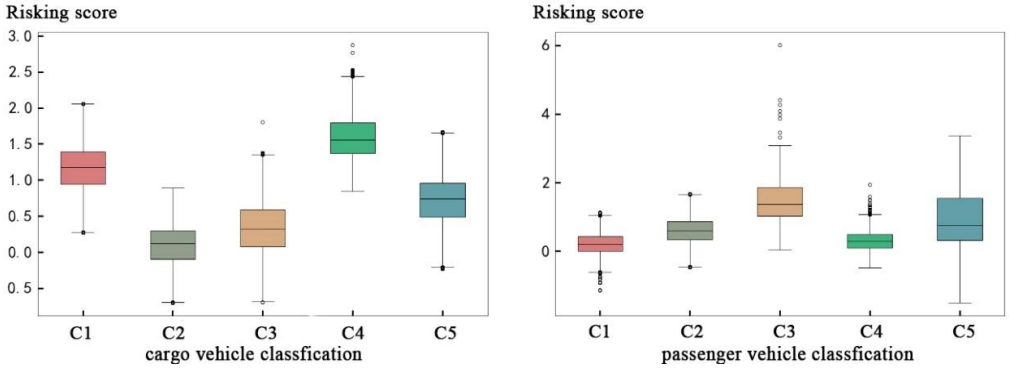

Figure 5 – Box plot of driving risk score

We can tell the risk level ranking of different types of cargo vehicles is **C4>C1>C5>C3>C2**. The risk level ranking of different types of passenger vehicles is **C3>C5>C2>C4>C1**.

5. RESULTS

5.1 Method evaluation

This paper uses actual feedback data for comparison to evaluate the effect of the proposed model. In the same period, the Guangdong Province crash database was queried to obtain the proportion of cargo vehicles involved in crashes among five risk levels and the proportion of passenger vehicles involved in crashes among five risk levels as vehicle risk coefficients. Using the rescue rate per 10,000 vehicles as a comparative metric, the calculation equation is *Equation 19*.

$$R_m = \frac{D}{M} \times 10^4 \tag{19}$$

where R_m represents the number of rescue events per 10,000 vehicles in a year, D denotes the number of rescue events for the sample vehicles in a year, and M is the total number of sample vehicles.

The calculation results are presented in *Figure 6*. We can observe that the risk level ranking of different types of cargo vehicles is **C4>C1>C5>C3>C2**. The risk level ranking of different types of passenger vehicles is **C3>C5>C2>C4>C1**, which aligns with the analytical results of the model.

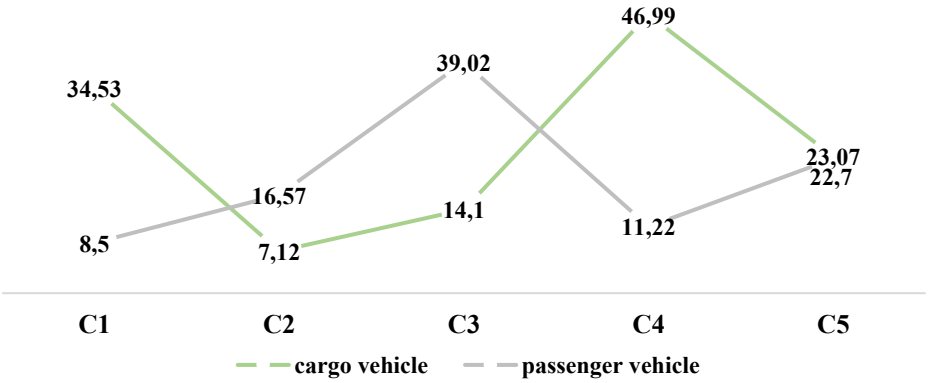

Figure 6 – The result of relative risk coefficients for five categories of cargo vehicles and passenger vehicles

To quantify the association between risk scores and observed risk, we calculated the relative risk coefficient (RRC) for each group relative to the low-risk baseline. RRC measures the fold increase in risk and is computed with *Equation 20*.

$$RRC = \frac{R_m^C}{R_m^1} \tag{20}$$

where R_m^1 is the R_m for the low-risk group, R_m^C represents the R_m across all vehicle risk categories. The vehicles in each risk category were ranked in descending order of RRC, resulting in *Table 10*.

Table 10 – Vehicle risk categories for passenger and cargo vehicles ranked by descending RRC and cumulative share

Cargo vehicle					Passenger vehicle				
Category	R_m	RRC	Vehicle share	Cumulative share	Category	R_m	RRC	Vehicle share	Cumulative share
C4	46.99	6.60	17.75%	17.75%	C3	39.02	4.59	7.47%	7.47%
C1	34.53	4.85	14.03%	31.78%	C5	22.7	2.67	1.08%	8.55%
C5	23.07	3.24	14.79%	46.57%	C2	16.57	1.95	3.70%	12.25%
C3	14.1	1.98	17.80%	64.37%	C4	11.22	1.32	43.63%	55.88%
C2	7.12	1.00	35.63%	100.00%	C1	8.5	1.00	44.12%	100.00%

As shown in *Table 10*, for cargo vehicles, categories C4, C1 and C5 have an R_m > 20, with these high-risk cargo vehicles accounting for a cumulative proportion of 46.57%, nearly half of the total cargo vehicles. Their relative risk coefficient (RRC) is at least 3.24 times higher than that of low-risk cargo vehicles. For passenger vehicles, categories C3 and C5 have an R_m > 20, with these high-risk passenger vehicles comprising only 8.55% of the cumulative proportion, less than 10% of the total passenger vehicles. However, their RRC is at least 2.67 times higher than that of low-risk passenger vehicles. These high-risk vehicles warrant particular attention.

5.2 High-risk class vehicles

Based on the above clustering results and driving risk score situation, as well as relevant indicator data, the driving risk characteristics of different vehicle categories were analysed, the vehicle driving risk level was determined, and high-driving risk vehicles were extracted. Based on the clustering results of the five vehicle category distribution, this study classifies vehicle risk into five levels from high to low: high-risk, moderately high-risk, moderate-risk, moderately low-risk and low-risk. The following lists the vehicle categories with R_m exceeding 20.

1) High-risk cargo vehicles: cargo vehicle Category 4

This type of cargo vehicle has the highest driving risk score, accounting for 17.75% of all cargo vehicles. This type of cargo vehicle primarily consists of short-haul medium-duty cargo vehicles, with relatively small amounts of goods transported, complex and diversified transportation routes, and more individual owners' operations, focusing on expanding the coverage of transportation objects and increasing the number of transportation times to improve efficiency and income. In order to increase the number of transportation times per unit time, speeding behaviour is common. The relevant labels for this type of vehicle are short-distance transportation, frequent trips on the freeway, high vehicle private attributes, frequent night driving and peak driving times, obvious speeding behaviour and unstable transportation routes.

2) Moderately high-risk cargo vehicles: cargo vehicle Category 1

This type of cargo vehicle ranks second in driving risk score among cargo vehicles, accounting for 14.03% of all cargo vehicles. This type of cargo vehicle primarily consists of long-haul heavy-duty cargo vehicles, with most vehicles belonging to large logistics companies or transporters. This type of cargo vehicle tends to transport as much and as far as possible in a single trip, thereby reducing the number of trips on the freeway and having relatively stable transportation employers. Vehicles will try to maintain full load, and these vehicles will be subject to public supervision; speeding behaviour is not obvious. The relevant labels for this type of vehicle are high single-trip transportation intensity, obvious public attributes, long night driving time and peak driving time, high proportion of cargo vehicles with loads over 12 tons, stable transportation route and infrequent trips on the freeway.

3) Moderate-risk cargo vehicles: cargo vehicle Category 5

This type of cargo vehicle ranks third in driving risk score among cargo vehicles, accounting for 14.79% of all cargo vehicles. This type of cargo vehicle primarily comprises medium- to short-haul heavy-duty cargo vehicles, with large amounts of goods transported and obvious overload behaviour. The relevant labels for this type of vehicle are relatively high single-trip transportation intensity, serious overload, and a high proportion of cargo vehicles with loads over 12 tons.

4) High-risk passenger vehicles: passenger vehicle Category 3

This type of passenger vehicle has the highest driving risk score, accounting for 7.47% of all passenger vehicles. This type of passenger vehicle primarily consists of short-haul commercial passenger vehicles with extremely high route familiarity. It has a high number of trips on the freeway, a high number of peak trips, obvious speeding behaviour and obvious commuting characteristics. The vehicle travels frequently during peak hours and during traffic congestion peak hours. Speeding can break through congestion to some extent, avoid falling into the congested traffic flow, and thus shorten travel time, resulting in a widespread speeding phenomenon. The relevant labels for this type of vehicle are obvious commuting characteristics, frequent trips on the freeway, frequent peak driving times and obvious speeding behaviour.

5) Moderately high-risk passenger vehicles: passenger vehicle Category 5

This type of passenger vehicle has the second-highest driving risk score among passenger vehicles, accounting for 1.08% of all passenger vehicles. This type of passenger vehicle primarily consists of long-haul passenger vehicles. It has a low number of trips on the freeway, a high single-trip driving intensity, a high

fatigue driving risk, long night driving time and peak driving time, and generally travels across provinces. The relevant labels for this type of vehicle are high driving intensity, high fatigue driving risk and frequent peak driving times.

6. CONCLUSION

This study developed a driving risk level identification model for cargo vehicles and passenger vehicles on highways, leveraging toll data mining and feature extraction from freeway entrances and exits in the Changping-Guanzhang section of the Yongguan Freeway in Guangdong Province (12 June to 20 August 2022). By applying factor analysis and clustering algorithms, the model quantified driving risk characteristics and categorised vehicles into risk levels. The results identified 17.75% of cargo vehicles as "high-risk" and 14.03% as "moderately high-risk," alongside 7.47% of passenger vehicles as "high-risk" and 1.08% as "moderately high-risk." Validation against crash data from the same period in Guangdong Province confirmed the model's effectiveness in identifying vehicles with elevated driving risk characteristics, providing a robust data foundation for highway safety management.

Based on the preliminary calculation of 18 cargo vehicle characteristic indicators and 13 passenger vehicle characteristic indicators, factor analysis was conducted to obtain representative influencing factors after dimensionality reduction. Clustering algorithms successfully constructed driver behaviour profiles, revealing significant correlations between key risk indicators and vehicle types, thus extending risk assessment from a single-vehicle perspective to a vehicle model-level analysis. The proposed quantitative method, based on continuous observational data, accurately calculated crash probabilities for different vehicle models, with risk rankings aligning closely with historical crash data, validating its predictive accuracy and practical utility.

The model relies on toll data from a specific highway section, which may limit its generalisability to other road types or regions. The 31 driving characteristic indicators, while comprehensive, may not fully capture emerging factors associated with new energy vehicles and autonomous driving technologies, such as battery performance or automation levels. The model's reliance on historical crash data for validation assumes data completeness and accuracy, which may be affected by underreporting or inconsistencies. The focus on highway environments may not account for the unique traffic dynamics of urban or rural roads, necessitating further adaptation of the indicator system.

Future research should expand the model to include diverse road types and regions to enhance its generalisability. Incorporating indicators specific to new energy vehicles and autonomous vehicles, such as energy consumption patterns or automation response times, could improve the model's relevance. Refining the indicator system for non-highway environments and integrating real-time data sources, such as in-vehicle sensors, could enhance the model's timeliness and applicability. The results enable traffic management authorities to implement targeted interventions, including high-risk vehicle identification, focused inspections for overloading and tailored safety guidance, thereby enhancing overall highway safety management.

ACKNOWLEDGEMENTS

This work was supported by the General Programme of the National Natural Science Foundation of China (No. 52072130) and the Natural Science Foundation of China (No. 72471091).

REFERENCES

[1] Liu Q, et al. Transfer learning-based highway crash risk evaluation considering manifold characteristics of traffic flow. *Accident Analysis & Prevention*. 2022;168:106598–106598. DOI: 10.1016/J.AAP.2022.106598.

[2] China NBOS 2021. Traffic Accidents 2021. https://data.stats.gov.cn/easyquery.htm?cn=C01&zb=A0S0D04&sj=2021 [Accessed 20th Mar. 2023].

[3] US Department Of Transportation NHTS. Traffic Safety Facts - 2015 Data - NHTSA 2017. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/812409_tsf2015dataspeeding.pdf [Accessed 20th Mar. 2023].

[4] Bureau Of Transportation Statistics USDOT 2021. Transportation Fatalities by Mode n.d.. https://www.bts.gov/content/transportation-fatalities-mode [Accessed 20th Mar. 2023].

[5] Database TCSN 2021. Canadian Motor Vehicle Traffic Collision Statistics: 2021 n.d.. https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2021 [Accessed 20th Mar. 2023].

[6] Basu S, Saha P. Evaluation of risk factors for road accidents under mixed traffic: Case study on Indian highways. *IATSS Research*. 2022;46:559–73. DOI: 10.1016/j.iatssr.2022.09.004.

[7] Zhang R, et al. Driver's journey from historical traffic violations to future accidents: A China case based on multilayer complex network approach 2024. DOI: 10.1016/j.aap.2024.107901.

[8] Zhang R, et al. Innovative prediction and causal analysis of accident vehicle towing probability using advanced gradient boosting techniques on extensive road traffic scene data 2025. DOI: 10.1016/j.aap.2024.107909.

[9] Zeng Q, et al. Incorporating real-time weather conditions into analyzing clearance time of freeway accidents: A grouped random parameters hazard-based duration model with time-varying covariates. *Analytic Methods in Accident Research*. 2023;38:100267. DOI: 10.1016/j.amar.2023.100267.

[10] AlKheder S, AlRukaibi F, Aiash A. Risk analysis of traffic accidents' severities: An application of three data mining models. *ISA Transactions*. 2020;106:213–20. DOI: 10.1016/j.isatra.2020.06.018.

[11] Anastasopoulos PC, et al. A study of factors affecting highway accident rates using the random-parameters tobit model. *Accident Analysis & Prevention*. 2012;45:628–33. DOI: 10.1016/J.AAP.2011.09.015.

[12] Qi H, et al. BGCP-based traffic data imputation and accident detection applications for the national trunk highway. *Accident Analysis & Prevention*. 2023;186:107051–107051. DOI: 10.1016/J.AAP.2023.107051.

[13] Parsa AB, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*. 2020;136:105405–105405. DOI: 10.1016/J.AAP.2019.105405.

[14] Mahmud SMS, et al. Micro-level safety risk assessment model for a two-lane heterogeneous traffic environment in a developing country: A comparative crash probability modeling approach. *Journal of Safety Research*. 2019;69:125–34. DOI: 10.1016/j.jsr.2019.03.008.

[15] Koçar O, Dizdar E. A risk assessment model for traffic crashes problem using fuzzy logic: A case study of Zonguldak, Turkey. *Transportation Letters*. 2022;14:492–502. DOI: 10.1080/19427867.2021.1896062.

[16] Raskar C, Nema S. Metaheuristic enabled modified hidden Markov model for traffic flow prediction. *Computer Networks*. 2022;206:108780. DOI: 10.1016/j.comnet.2022.108780.

[17] Yu B, et al. Quantifying drivers' visual perception to analyze accident-prone locations on two-lane mountain highways. *Accident Analysis & Prevention*. 2018;119:122–30. DOI: 10.1016/J.AAP.2018.07.014.

[18] Wen H, Xue G. Injury severity analysis of familiar drivers and unfamiliar drivers in single-vehicle crashes on the mountainous highways. *Accident Analysis & Prevention*. 2020;144:105667. DOI: 10.1016/j.aap.2020.105667.

[19] Zhao X, et al. A multinomial logit model: Safety risk analysis of interchange area based on aggregate driving behavior data. *Journal of Safety Research*. 2022;80:27–38. DOI: 10.1016/j.jsr.2021.11.002.

[20] Zhang Y, et al. A proactive crash risk prediction framework for lane-changing behavior incorporating individual driving styles. *Accident Analysis & Prevention*. 2023;188:107072. DOI: 10.1016/j.aap.2023.107072.

[21] Cooper A. The inmates are running the asylum. In: Arend U, Eberleh E, Pitschke K, editors. Software-Ergonomie '99: Design von Informationswelten, Wiesbaden: Vieweg+Teubner Verlag; 1999, p. 17–17. Wiesbaden: Vieweg+Teubner Verlag Publishing; 1999.

[22] Massanari A. Designing for imaginary friends: Information architecture, personas and the politics of user-centered design. *New Media & Society - NEW MEDIA SOC*. 2010;12:401–16. DOI: 10.1177/1461444809346722.

[23] Miaskiewicz T, Kozar K. Personas and user-centered design: How can personas benefit product design processes? *Design Studies*. 2011;32:417–30. DOI: 10.1016/j.destud.2011.03.003.

[24] Holden R, et al. Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *International Journal of Medical Informatics*. 2017;108. DOI: 10.1016/j.ijmedinf.2017.10.006.

[25] De Clerck Q, et al. Total cost for society: A persona-based analysis of electric and conventional vehicles. *Transportation Research Part D: Transport and Environment*. 2018;64:90–110. DOI: 10.1016/j.trd.2018.02.017.

[26] Singh H, Kathuria A. Profiling drivers to assess safe and eco-driving behavior – A systematic review of naturalistic driving studies. *Accident Analysis & Prevention*. 2021;161:106349. DOI: 10.1016/j.aap.2021.106349.

[27] Maier C, et al. Smartphone use while driving: A fuzzy-set qualitative comparative analysis of personality profiles influencing frequent high-risk smartphone use while driving in Germany. *International Journal of Information Management*. 2020;55:102207. DOI: 10.1016/j.ijinfomgt.2020.102207.

[28] Zhou F, et al. Driver fatigue transition prediction in highly automated driving using physiological features. *Expert Systems with Applications*. 2020;147:113204. DOI: 10.1016/j.eswa.2020.113204.

[29] Useche SA, et al. Multidimensional prediction of work traffic crashes among Spanish professional drivers in cargo and passenger transportation. *International Journal of Occupational Safety and Ergonomics*. 2022;28:20–7. DOI: 10.1080/10803548.2020.1732102.

[30] China TSC 2019. The Regulations for the Implementation of the Road Traffic Safety Law of the People's Republic of China (in Chinese) n.d. The Central People's Government of the People's Republic of China [Accessed 20th Mar. 2023].

[31] Kong X, et al. Understanding speeding behavior from naturalistic driving data: Applying classification based association rule mining. *Accident Analysis & Prevention*. 2020;144:105620. DOI: 10.1016/j.aap.2020.105620.

[32] China TSC 2022. Administrative Measures for Scoring Management of Road Traffic Safety Violations (in Chinese) 2022. https://www.gov.cn/gongbao/content/2022/content_5679697.htm [Accessed 20th Mar. 2023].

[33] Arthur D, Vassilvitskii S. K-Means++: The Advantages of Careful Seeding. Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms, vol. 8, 2007, p. 1027–35. DOI: 10.1145/1283383.1283494.