Xiaoxia XIONG, Ph.D.[1]
E-mail: x_xiong623@163.com
Yu HE, M.S. candidate[1]
E-mail: 2212104009@stmail.ujs.edu.cn
Xiang GAO, Ph.D. candidate[1]
E-mail: motgaoxiang@foxmail.com
Yeling ZHAO, M.S.[2]
(Corresponding author)
E-mail: zhaoyeling2006@ujs.edu.cn
[1] School of Automotive and Traffic Engineering
  Jiangsu University
  301 Xuefu Road, Zhenjiang 212013, P.R. China
[2] School of Finance and Economics
  Jiangsu University
  301 Xuefu Road, Zhenjiang 212013, P.R. China

A MULTI-LEVEL RISK FRAMEWORK FOR DRIVING SAFETY ASSESSMENT BASED ON VEHICLE TRAJECTORY

ABSTRACT

Few existing research studies have explored the relationship of road section level, local area level and vehicle level risks within the highway traffic safety system, which can be important to the formation of an effective risk event prediction. This paper proposes a framework of multi-level risks described by a set of carefully selected or designed indicators. The interrelationship among these latent multi-level risks and their observable indicators are explored based on vehicle trajectory data using the structural equation model (SEM). The results show that there exists significant positive correlation between the latent risk constructs that each have adequate convergent validity, and it is difficult to completely separate the local traffic level risk from both the road section level risk and vehicle level risk. The local and road level indicators are also found to be of more importance when risk prediction time gets earlier based on feature importance scoring of the LightGBM. The proposed conceptual multi-level indicator based latent risk framework generally fits with the observed results and emphasises the importance of including multi-level indicators for risk event prediction in the future.

KEYWORDS

traffic safety; multi-level risk; safety indictor; SEM; vehicle trajectory.

1. INTRODUCTION

Highway traffic safety has been a critical issue affecting public health globally for a long time [1], which is a hot and challenging topic for a variety of researchers and technical staff. Many previous studies on traffic safety have relied on historical traffic accident data, which has the advantage of intuitiveness, but is usually limited in sample size (as a traffic accident is a small probability event) and might ignore some near-crash events (or serious conflicts) that can evolve into crashes due to the high uncertainty during vehicle driving [2]. Thanks to the development of the information acquisition technology, more driving data with traffic conflicts of different severity have become available, such as naturalistic driving data [3] and vehicle trajectory data [4]. Naturalistic driving data are usually collected by sensor-equipped vehicles driven by recruited participants, and can provide a moving status of the subject vehicle and its surrounding vehicles, as well as driver behavioural data of the subject vehicle [5]. Vehicle trajectory data are generally obtained by video camera placed on the nearby building or more recently by drones, which provides a wider range trajectory data that covers all vehicles in the road section under surveillance [4]. Such enriched data sources have facilitated more diversified techniques in analysing driving safety problem.

Traffic conflict technique (TCT) is one of the most popular techniques to diagnose a driving safety problem at the vehicle level [6]. TCT generally utilises surrogate safety measures (SSM) to identify and quantify traffic conflicts or risk events, which are statistically related to crashes [7]. The SSM can be classified into sub-categories,

such as the time-based, distance-based and deceleration-based SSM, based on their measurement units. Frequently used time-based SSM includes time-to-collision (TTC) [8], time headway (THW) [9] and post-encroachment time (PET) [10], etc. For example, the distance-based SSM includes a proportion of stopping distance (PSD) [11] and potential index of collision with urgent deceleration (PICUD) [12]. Deceleration rate to avoid the crash (DRAC) [13] and the ratio of DRAC to maximum deceleration [6] are typical deceleration-based SSMs. For a given traffic scenario, the SSM is usually calculated and compared with the pre-determined thresholds to identify and determine the severity of the traffic conflict. These thresholds can have a great impact on the evaluation results and can be obtained from statistical percentiles [14], fitted distributions [4] or cluster analysis [15]. As the SSMs generally have their applicable conditions and limitations [16], they can be fused as a composite to identify high-risk events and achieve more reliable safety evaluation [17, 18].

Data driven techniques, such as machine learning and deep learning, have also been developed to predict driving risk status. On the basis of the STISIM simulation driving data, Zhou et al. divided the real-time driving sequence signal into the safe and dangerous modes by establishing a conditional random field (CRF) model [19]. Ning et al. constructed the expected score function of the driving risk through time series difference learning (TDL) and used this function to evaluate the degree of driving risk in real time [20]. Lee et al. trained the multi-layer perceptron neural network (MPNN) based on the observed sample braking levels from the NGSIM and utilised the developed network model in a real time driving risk estimation [21]. On the basis of the Matlab simulation, Fu et al. constructed a risk status classification algorithm based on the neural network model, considering both driving safety and comfort [22]. However, although these methods can produce high-accuracy prediction results, they are sensitive to the transferability of the data and can hardly provide insights into the failure mechanism of the high risk events. Statistical modelling methods, on the other hand, are more model-driven and have better model interpretability, which have been widely employed to analyse the mechanism of risk development in driving. Using vehicle trajectory data from HighD, Yu et al. developed various lo-gistic regression models to examine the relationship of high risk event occurrence and the traffic flow characteristics in front of the subject vehicle [23]. Also with HighD data, Hu et al. employed a binary logistic model to quantify the relationship between the occurrence of traffic conflict and lane-level traffic states [24]. Chen et al. explored the vehicle operation level contributing factors to lane-changing risks via random parameter logit models using HighD data [15]. Similarly, Chen et al. developed mixed binary logit models to investigate risk during the lane-changing process from the perspective of vehicle group using the NGSIM data [25].

It can be seen from above that the existing research studies generally assess driving safety from either the vehicle operation level or the road section traffic flow level. However, few have explored the multi-level risk relationship within the safety system, which is feasible with the increasing availability of large-scale vehicle trajectory data. From the spatial perspective, the risk influential space scope for a subject vehicle can be divided into three levels, including the subject vehicle itself, the local area of the subject vehicle covering its surrounding vehicles and the road section in which the subject vehicle is located. A thorough understanding of the interrelationship of these multi-level risks can fill in the gap and shed light on the comprehensive formation of risk event prediction. This paper explores the potential relationship among these risks of different levels based on vehicle trajectory data using the structural equation model (SEM), which is a quantitative statistical technique that can be designed to confirm conceptual hypotheses [26] (details of which will be further discussed in the next section). The remainder of the paper is organised into the following sections: Section 2 describes the proposed multi-level risk framework and SEM modelling methodology. Section 3 provides the data preparation details of the model validation. The modelling results and discussions are presented in Section 4 and Section 5 concludes the paper.

2. METHODOLOGY

In this section, the framework of multi-level risks and representing indicators is firstly proposed for safety assessment (as shown in *Figure 1*), and the SEM modelling technique to analyse their latent relationships is then presented. Specifically, in the
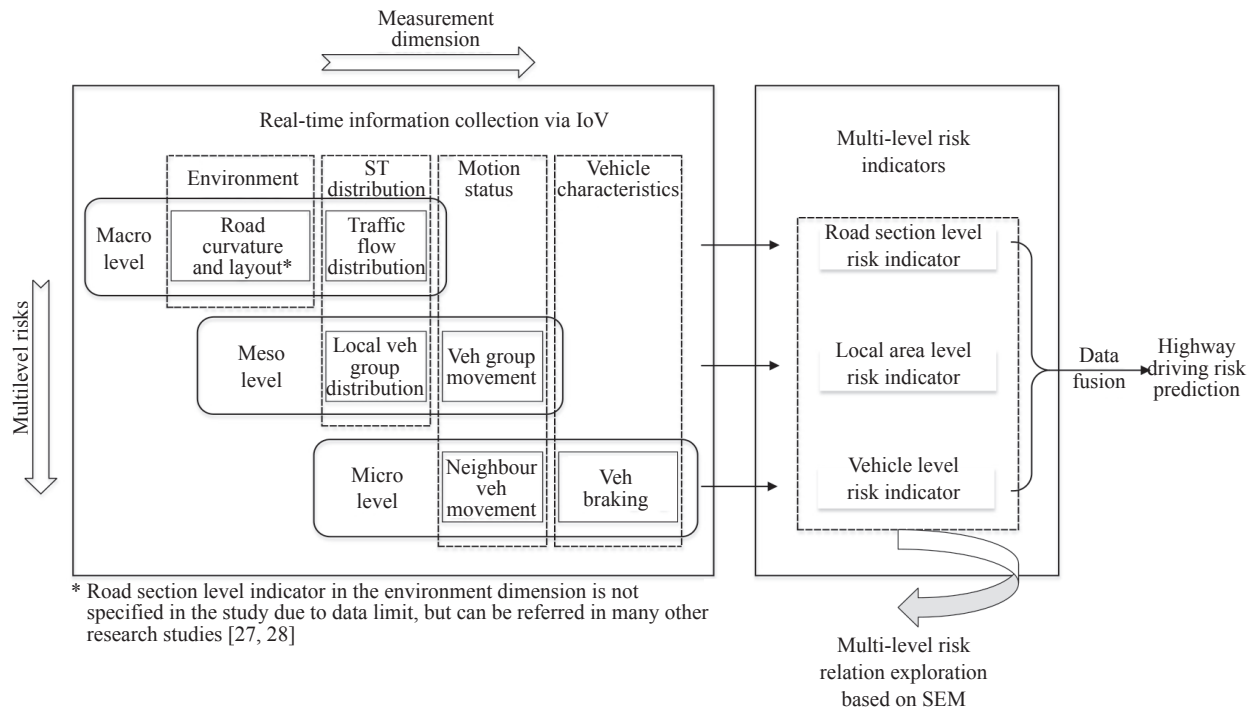
Figure 1 – Framework of multi-level risks and representing indicators for safety assessment

study, the multi-level risks refer to the quantified driving risk for a subject vehicle from the perspective of space scopes, including road section (macro) level, local area (meso) level and vehicle (micro) level risks. Each level risk can then be described with a set of observable indicators from different measurement dimensions, i.e. from spatial and temporal distribution (ST distribution) of the traffic flow to individual motion status and characteristics of the vehicle. All required data for indicator calculation are assumed to be able to be collected in real time via the Internet of Vehicles (IoV).

In order to establish the relationship among the multi-level risks, conflict driving samples will be obtained based on the commonly used surrogate safety measure TTC, which will be detailed in Section 3. For each conflict driving sample, a set of multi-level indicators will be selected and calculated to reflect the multi-level risks of the sampled subject vehicle. Considering the wider variety in surrogate safety measures in longitudinal scenarios than in lateral scenarios (which directly affects the selection of risk indicators for the vehicle level), only longitudinal conflict cases are considered in the study. The details of multi-level risks and representing indicators, as well as their interrelationship exploration methodology SEM, will be presented in the following sections. It should be noted that the main purpose of this paper is to explore and validate

the interrelationship of multi-level risks based on multi-level indicators, whereas effective fusion of different level indicators for risk prediction would be left to future work.

2.1 Road section level risk

Traditional road level indicators generally include the average (avg) speed, standard deviation (S.D.) of speed, variation coefficient (V.C.) of speed and the longitudinal distance between neighbouring vehicles within the traffic flow. Specifically, the average value of vehicle speed directly reflects the operation efficiency of traffic flow; dispersion coefficient of traffic flow, including S.D. and V.C. of vehicle speed, have been proved to reflect the safety state of traffic flow to some extent [23]; the longitudinal distance between vehicles, including their average and minimum distance, measures the spatial proximity of neighbouring vehicles and partly reflects the stability of traffic flow.

In this study, the above statistical indicators are calculated based on different measurement scales (see *Table 1*), including (1) vehicles in front of the subject vehicle (up to the end of the road segment), and abbreviated as FS hereinafter; (2) vehicles on the subject lane and in front of the subject vehicle, and abbreviated as FL; (3) vehicles in the whole road segment (from the start to the end of the road segment the subject vehicle is driving in; the road

Table 1 – Road section level indicators for consideration

Scale	Indicator variable	Definition
FS	avg_xVelo_FS, std_xVelo_FS, vc_ xVelo_FS, avg_dist_FS, min_dist_FS, FI_FS	avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for FS
FL	avg_xVelo_FL, std_xVelo_FL, vc_ xVelo_FL, avg_dist_FL, min_dist_FL, FI_FL	avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for FL
WS	avg_xVelo_WS, std_xVelo_WS, vc_ xVelo_WS, avg_dist_WS, min_dist_WS, FI_WS	avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for WS
WL	avg_xVelo_WL, std_xVelo_WL, vc_ xVelo_WL, avg_dist_WL, min_dist_WL, FI_WL	avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for WL

segment is predefined by the dataset which will be discussed in more detail in Section 3) and abbreviated as WS; (4) vehicles on the subject lane in the whole road segment and abbreviated as WL. For FS and WS measurement scale, in addition to vehicles on all lanes, indicators are also calculated among lanes to obtain difference measures between adjacent lanes, which have been shown to have impact on real-time crashes [24].

The statistical indicators above generally reflect the overall driving characteristics of the vehicles in the section. However, the complexity in interrelationship within the traffic flow is hard to capture. In light of this, a road section level indicator, traffic flow instability index, is proposed based on the theory of complex networks [29] as follows.

Construction of traffic flow graph of the road section

A global coordinate system is first established taking the upper left corner of the road section as the origin, the vehicle driving direction as the positive x-axis and its 90° clockwise rotation direction as the positive y-axis (see *Figure 2*). Take all vehicles within the measurement scale as the nodes of the

graph, denoted as $\mathbf{C} = \{c_k\}_{k=1}^{K}$, where K designates the total number of vehicles (nodes). Denote the position of the mass centre of vehicle c_k as (x_k, y_k), its longitudinal and lateral speed as v_{xk} and v_{yk}, its longitudinal and lateral acceleration rate as a_{xk} and a_{yk}, and its lane number as LL_k. In the study, the inner lane is indexed as 1 and the indices of other lanes are increased by one per lane from the inside to the outside. Any two vehicle nodes are assumed to be linked with an undirected edge only if there exist potential longitudinal and lateral interactions between them. Mathematically, an edge between nodes c_i and c_j exists if one of the following rules is met:

1) $LL_i = LL_j : \{c_k \in \mathbf{C} | (LL_k = LL_i) \& (\min(x_i, x_j) < x_k < \max(x_i, x_j))\} = \emptyset$

2) $|LL_i - LL_j| = 1 : ① |x_i - x_j| < d; ② (LL_i - LL_j) v_{yi} < 0 \lor (LL_i - LL_j) a_{yi} < 0 \lor (LL_i - LL_j) v_{yj} > 0 \lor (LL_i - LL_j) a_{yj} > 0$

Rule 1 means there exists an edge between two vehicle nodes if they are driving on the same lane and are located next to each other. Rule 2 accounts for the adjacent vehicles which are driving on the neighbouring lanes within a certain distance d
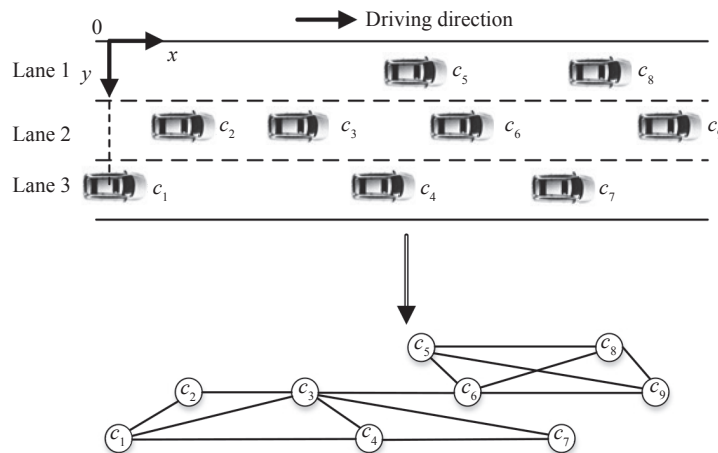

Figure 2 – Diagram of the traffic flow network graph

(which is set to 150m in the study) and have lateral velocities or accelerations in opposing directions. As can be noted in *Figure 2*, although the adjacent-lane vehicle pairs c_3-c_5 and c_7-c_9 are located within the required distance, they do not meet the condition on lateral movement direction, and as a result no connection is established within these pairs.

Each edge connecting two nodes is assigned a given weight to reflect the intensity of interaction between the two vehicles, which is defined as the speed difference between the two vehicles within a unit distance:

$$w_{ij} = \frac{|v_{xi} - v_{xj}|}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}} \quad (1)$$

Calculation of the traffic flow instability index

Based on the defined graph, unit vertex strength of each vehicle can be obtained as follows:

$$\overline{W}_i = \sum_{j \in N_i} w_{ij}/d_i \quad (2)$$

where N_i designates the set of indices of adjacent nodes of the node c_i. $\sum_{j \in N_i} w_{ij}$ represents the sum of weights on the edges connecting the node, also referred to as vertex strength of node c_i. d_i designates the degree of node c_i or the number of connections node c_i has with other nodes.

Taking the squared degree of each vehicle node as the weight, the weighted average of unit vertex strength of all vehicle nodes $\{c_k\}_{k=1}^K$ in the graph can be calculated as below and is used as an index of traffic flow instability (FI) within the measurement scale.

$$FI = \sum_{i=1}^K \left(\frac{d_i^2}{\sum\limits_{k=1}^K d_k^2} \overline{W}_i \right) \quad (3)$$

2.2 Local area level risk

The local area of the subject vehicle takes into account its surrounding vehicles within a certain longitudinal distance, including its front, backward, left front, left backward, right front and right backward vehicles, i.e. c_3, c_4, …, c_9 for subject vehicle c_6 in *Figure 3*. Similar to statistical indicators for different measurement scales defined in Section 2.1, indicators including avg, S.D. and V.C. of speed, avg dist. and min. dist. are calculated for the vehicles within the local area of the subject vehicle.

In addition to the statistics of the driving characteristics of surrounding vehicles, a local traffic instability index is also defined considering the spatial distribution of these vehicles. Based on the two-dimensional Gaussian distribution probability density function, the spatial proximity between the surrounding vehicle and the subject vehicle can be measured as [30]:

$$f(X_j; \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left\{ -\frac{1}{2}(X_j - \boldsymbol{\mu})^T \boldsymbol{\Lambda}^{-1}(X_j - \boldsymbol{\mu}) \right\} \quad (4)$$

$$X_j = \left(\frac{S_x^{0j}}{(|v_{xj}| + |v_{x0}|)/2}, \frac{S_y^{0j}}{(|v_{yj}| + |v_{y0}|)/2} \right)^T \quad (5)$$

where X_j measures the relative longitudinal and lateral time distance between the subject vehicle c_0 and its surrounding vehicle c_j. $S_x^{0j} = |x_j - x_0| - (L_j + L_0)/2$ and $S_y^{0j} = |y_j - y_0| - (U_j + U_0)/2$ measures the spatial longitudinal and lateral distance between the two vehicles, respectively. L_0 and L_j are the lengths of c_0 and c_j, respectively, U_0 and U_j are the widths of c_0 and c_j, respectively. $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda} = \text{diag}(\sigma_x^2, \sigma_y^2)$ designates the mean and covariance of the two-dimensional Gaussian distribution and is set to $\boldsymbol{\mu} = (0,0)^T$ and $\boldsymbol{\Lambda} = \text{diag}(1,1)$ in the study.

Based on the vehicle proximity function defined above and the theory of information entropy [31], the local traffic instability index (LTI) can then be defined considering the speed difference between the subject vehicle and all of its surrounding vehicles as well as the uncertainty in their spatial distribution:

$$LTI = -\sum_{j=1}^{n_0} |v_{x0} - v_{xj}|^2 \left[f(X_j; \boldsymbol{\mu}; \boldsymbol{\Lambda}) \log(f(X_j; \boldsymbol{\mu}; \boldsymbol{\Lambda})) \right] \quad (6)$$

where n_0 represents the number of surrounding vehicles of the subject vehicle c_0.
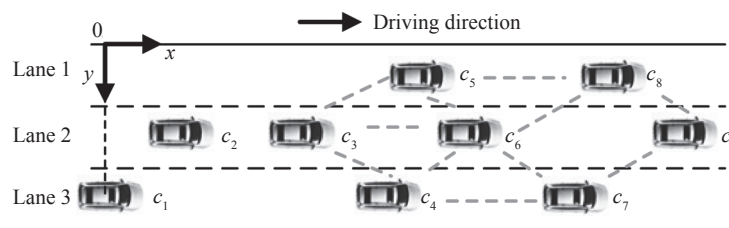

Figure 3 – Diagram of the area of local traffic for veh c_6

Table 2 – Local area level indicators for consideration

Indicator variable	Definition
avg_xVelo_GP, std_xVelo_GP, vc_ xVelo_GP, avg_dist_GP, min_dist_GP	avg, S.D., V.C. of speed, avg dist., and min. dist. between vehs for the local area of the subject vehicle
LTI	Local traffic instability index

2.3 Vehicle level risk

As mentioned above, there has been a diversity of surrogate safety measures to evaluate the safety status of driving vehicles. In the study, vehicle level indicators are selected from deceleration based, distance based and time based surrogate measures. As the study only considers the longitudinal scenario, these vehicle level indicators are calculated for the pair of the subject vehicle c_0 and its leading vehicle c_j only.

Deceleration based

Deceleration based measures focus on the emergency of the evasive action (braking) the vehicle needs to take to prevent the occurrence of crash. Deceleration Rate to Avoid the Crash (DRAC) calculates the required deceleration rate for the subject vehicle to avoid crash with its leading vehicle that retains the same speed and trajectory [13]:

$$DRAC_x^{0j} = \frac{(v_{x0} - v_{xj})^2}{2S_x^{0j}} \qquad (7)$$

It can be noted that DRAC takes a valid value only when the traveling speed of the subject vehicle exceeds that of its leading vehicle (which can lead to a potential collision).

Maximum deceleration index (MDI) [6] can then be calculated as follows:

$$MDI = \frac{DRAC_x^{0j}}{MADR_x^0} \qquad (8)$$

where $MADR_x^0$ represents the braking capability (the maximum deceleration rate) of the subject vehicle. It should be noted that the value of the MADR are set according to different types of vehicles to account for potential vehicle size effects in risk analysis, and is set to 3.4m/s² for cars and 2.4 m/s² for trucks [15].

Distance based

Proportion of Stopping Distance (PSD) calculates the ratio between the remaining distance for the two vehicles to collide if the leading vehicle suddenly stops (a hypothetically extreme condition) and the minimum stopping distance of the subject vehicle [11].

$$PSD = \frac{S_x^{0j}}{v_{x0}^2/2MADR_x^0} \qquad (9)$$

Potential Index of Collision with Urgent Deceleration (PICUD) is also considered as a vehicle level risk indicator, which characterises the final distance between the subject vehicle and its leading vehicle if both of them decelerate under maximum braking force [12].

$$PICUD = S_x^{0j} + \frac{v_{xj}^2}{2MADR_x^j} - \frac{v_{x0}^2}{2MADR_x^0} - t_h v_{x0} \qquad (10)$$

where t_h is the driver reaction time of the subject vehicle and is set to 1s in the study.

Time based

TTC is a typical time based indicator in measuring risk trend. However, it is not included in the vehicle level indicator as it has been employed to define and obtain conflict driving samples in the study as mentioned before. Instead, another typical time-based safety indicator Time Headway (THW) is selected, which characterises the remaining time for the two vehicles to collide if the leading vehicle suddenly stops (a hypothetically extreme condition) while the subject continues driving at current speed along the current trajectory [9].

$$THW = \frac{S_x^{0j}}{v_{x0}} \qquad (11)$$

2.4 Risk modelling with multi-level indicators based on SEM

Given the framework of multi-level risks and representing indicators described above, the multi-level indicators can be thought to be the observed reflections of the latent multi-level risks, and the relationship among these multi-level risks are of interest to our study that needs further validation. Such research question generally fits into the scope of the confirmatory factor analysis (CFA), which belongs to the family of structural equation model (SEM) analysis [26]. In a CFA model, multiple items are created for each theory-derived latent factor constructs and the correlations among latent factors can be assessed by their covariance matrix.

Table 3 – Road section level indicators for consideration

Measurement category	Indicator variable	Definition
Deceleration-based	DRAC	Deceleration rate to avoid the crash
	MDI	Maximum deceleration index
Distance-based	PSD	Proportion of stopping distance
	PICUD	Potential indicator of collision with urgent deceleration
Time-based	THW	Time Headway

CFA helps bridging the gap between theory and observation by investigating the relations among a priori specified, theory driven latent and observed variables, and has become an important analysis tool for many social and behavioural science applications [26, 32].

As justified above, in this study, the multi-level risks, including road section level risk, local area level risk and vehicle level risk, are modelled as latent variables (also called factors in SEM) which are assumed to be correlated (see *Figure 4*). Multi-level indicators specified in Section 2.1 to 2.3 are treated as observed variables, also referred to as fixed covariates in statistics, which form the multi-level risk latent variables. The latent variables are linked to observed variables using the measurement equation as follows:

$$I_{ij} = \lambda_{ij} R_i + e_{ij} \tag{12}$$

where R_i, $i=1,2,3$, refers to the road section level risk (R_1), local area level risk (R_2) and vehicle level risk (R_3) respectively. $I_{ij}, j=1,2,\ldots$, designates the jth observed variable (i.e. the specified indicators) of the ith risk latent variable. λ_{ij} represents factor loading, which measures how much of the variability in the jth observed indicator is explained by the ith level risk status; it can also be interpreted as the amount of unit change in the observed in-

dicator I_{ij} when the latent risk status R_i undertakes one unit change. e_{ij} designates the error term for the jth observed indicator of the ith level risk. The regression weight of each error term is set to 1. Accordingly, the set of multi-level indicators can then be formulated as follows:

$$I = \Lambda R + \varepsilon$$

where $I = \begin{pmatrix} I_{11} \\ I_{12} \\ \vdots \\ I_{21} \\ I_{22} \\ \vdots \\ I_{31} \\ I_{32} \\ \vdots \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & \lambda_{21} & 0 \\ 0 & \lambda_{22} & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \lambda_{31} \\ 0 & 0 & \lambda_{32} \\ \vdots & \vdots & \vdots \end{pmatrix}, R = \begin{pmatrix} R_1 \\ R_2 \\ R_3 \end{pmatrix}, \varepsilon = \begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{21} \\ e_{22} \\ \vdots \\ e_{31} \\ e_{32} \\ \vdots \end{pmatrix} \tag{13}$$

Specifically, the multi-level risk latent variables are assumed to be normally distributed and their scale are standardised and follow the restrictions as below:

$$E(R) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; COV(R) = \Psi = \begin{pmatrix} 1 & & \\ \Psi_{21} & 1 & \\ \Psi_{31} & \Psi_{32} & 1 \end{pmatrix} \tag{14}$$

To achieve proper identification of the formulated model, no cross loadings are assumed and for each of the latent variable (factor), one of its factor loading is fixed to 1.
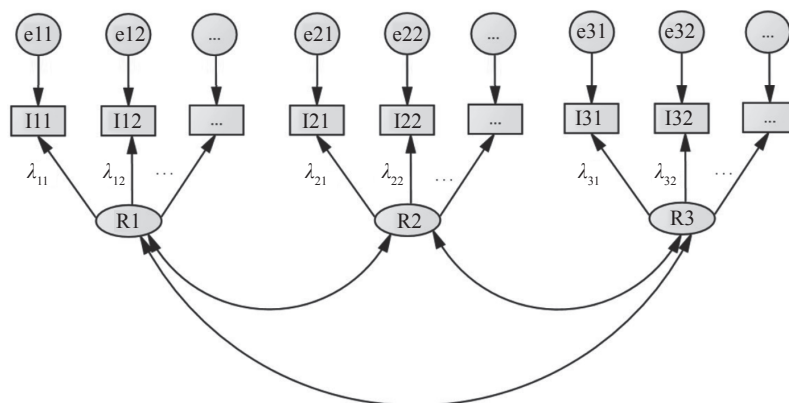

Figure 4 – Conceptual multi-level driving risk model

The performance of the established CFA model can be assessed by three parts, including model fitness, convergent and discriminant validity. The model fitness can be judged by GFI, RMSE and other typical indicators that have acceptable value ranges [33]. If the overall fitting of the model is acceptable, the internal validity of the factor constructs can then be further examined, including convergent and discriminant validity [34]. Convergent validity measures how well a set of indicators converge in explaining their common latent factor (usually judged by checking if the observed indicators in the same factor construct share a high proportion of variation), while discriminant validity, on the other hand, measures how divergent the indicators are from other indicators that assess different constructs [33]. Average variance extracted (AVE) and composite reliability (CR) for each latent construct are usually computed to assess convergent validity, and an AVE over 0.5 and CR over 0.7 indicates a good convergent validity of the construct. Discriminant validity of a given construct can then be evaluated by comparing the square root of its AVE with its correlation with other latent constructs, and a lower correlation compared with square root of AVE indicates a better discriminant validity. More details of calculation of the model fitness indicators and construct validity indicators can be referred to other resources [33] due to space limit of the paper.

3. DATA PREPARATION

The data in the study was obtained from HighD dataset, which provides high-resolution trajectory data of more than 110,500 vehicles collected at six different locations on freeways near Cologne, Germany, from 8 a.m. to 5 p.m. during 2017 and 2018 with sunny and windless weather, using a drone [35]. The road section at each location is straight in alignment, approximately 420m in length, and without any on-ramps and off-ramps. The road sections include both two-lane and three-lane segments, but only three-lane highway segments were included in the analysis to standardise the extraction of traffic flow data. The trajectory data of each vehicle contains information including vehicle type, longitudinal (x) and lateral (y) positions (with the coordinate system presented in *Figure 5*), driving direction and lane ID, ID of surrounding vehicles and TTC, etc., which covers the data needed to calculate the

multi-level indicators proposed in Section 2. More details of the dataset can be found at the HighD website [36].
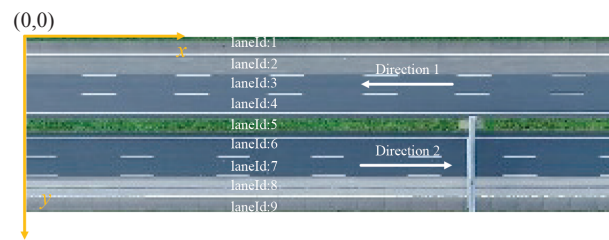

Figure 5 – HighD coordinate system

As discussed in the previous section, the interrelationship of multi-level indicators in longitudinal conflict cases are of interest to our study, and the longitudinal conflict samples were extracted based on the following steps:

- Longitudinal driving sample extraction: Lane changing segments of the subject vehicle were firstly removed from the original trajectory data. In the study, the starting point of a lane change segment is defined as 5 seconds prior to the moment when the boundary of the subject vehicle begins to enter the target lane, and the ending point is 5 seconds after the moment when the whole body of the subject vehicle enters the target lane. For the remaining trajectory data, only the observation segments that have existing preceding vehicle (with valid TTC value in highD) were kept for the following conflict and non-conflict vehicle sample extraction.

- Conflict sample extraction: Based on the longitudinal driving data obtained from step 1), conflict and non-conflict samples were classified using TTC. A TTC threshold of 4 seconds was selected in this study to obtain sufficient conflict cases [24], and the moment when TTC becomes lower than 4 seconds is defined as the observation moment (t_0) of the conflict driving sample.

- Non-conflict sample extraction: Observation segments within 20 seconds before and after the t_0 moment of conflict samples were removed first considering the potential risk impact within these data. For each continuous longitudinal driving segment in the remaining trajectory observations, the moment when TTC reaches the minimum value is defined as the observation moment (t_0) of the non-conflict driving sample.

A total of 356 conflict samples and 5,397 non-conflict cases were finally obtained following the steps above.

4. RESULTS AND DISCUSSION

4.1 Validation of the multi-level indicator based latent risk constructs

The confirmatory factor analysis of latent risk constructs was carried out with AMOS [37] by using the conflict samples obtained in Section 3. Given the large number of indicators with limited sample size, some of the items (indicators) in each factor construct should be removed from the measurement model in order to fulfil the requirement of model fit-ness and construct validity. In the study, the process of item removal was based on manual adjustment, which allows certain individual items to be retained/removed on the basis of previous research findings to achieve better model fitting, while maintaining reasonable explainability of the model. Specifically, in accordance with [33], an item would be considered for removal from the model if its standardised weight is estimated to be under 0.4 and at the same time its modification index (MI, which is available in the Amos report) is close to zero, which criteria usually suggest a small explanatory power of the item to both the factor it belongs to and other factor constructs. After such process of repeated trials, the final driving risk CFA model results are presented in *Figure 6* and *Tables 4–6*. To keep consistent the chang-
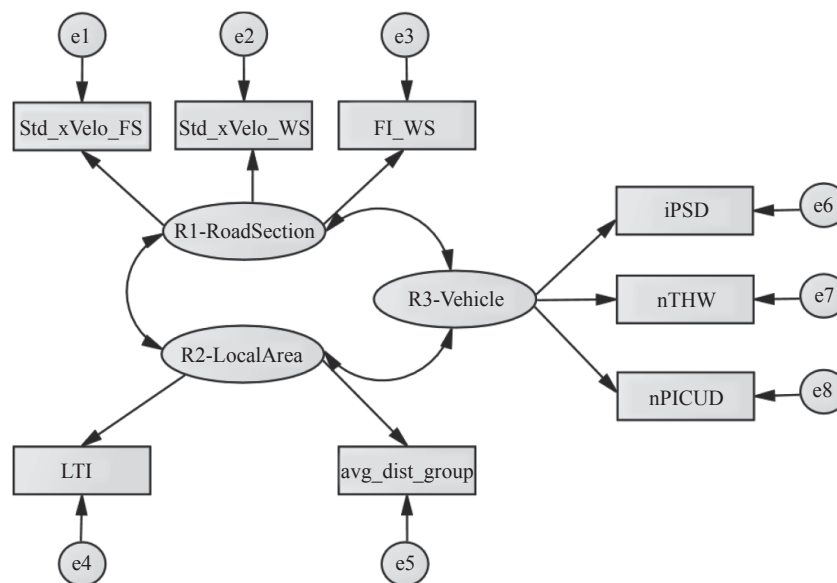

Figure 6 – The estimated driving risk CFA model

Table 4 – Descriptive statistics for all selected indicators in the driving risk model

Indicators	Description	Mean	Std.Dev.	Min	Max
std_xVelo_FS [m/s]	S.D. of speed for vehicles in front of the subject vehicle	3.745	1.766	0.042	11.004
std_xVelo_WS [m/s]	S.D. of speed for vehicles in the whole road segment	4.118	1.610	1.438	11.208
FI_WS	Traffic flow instability index	0.110	0.054	0.033	0.398
LTI	Local traffic instability index	7.687	9.524	0.097	63.838
avg_dist_group [m]	Average speed of vehicles within the local area of the subject vehicle	43.971	29.565	9.750	155.228
iPSD	Inverse of proportion of stopping distance	6.883	7.628	0.001	40.967
nTHW [s]	Negative time headway	-1.074	0.929	-8.160	-0.120
nPICUD [m]	Negative potential indicator of collision with urgent deceleration	15.085	25.638	-323.735	66.161

Table 5 – Modelling results of the driving risk model

Measurement equation factor loadings	Unstandardized weights			Standardized weights
	Estimate	Std.Error	P-Value	
Road section level				
Std_xVelo_FS←R_1	1*			0.884
Std_xVelo_WS←R_1	0.898	0.049	<0.001	0.871
FI_FS←R_1	0.031	0.002	<0.001	0.893
Local traffic level				
LTI←R_2	1*			0.779
avg_dist_group←R_2	44.428	3.680	<0.001	0.655
Vehicle level				
iPSD←R_3	1*			0.631
nTHW←R_3	0.129	0.010	<0.001	0.670
nPICUD←R_3	4.890	0.523	<0.001	0.919
Model fit indicators				
χ^2/df=4.361<5, GFI=0.957>0.900, RMSEA=0.097<0.100, SRMR=0.0510.050, CFI=0.967>0.900, NFI=0.958>0.900				

* Note that to identify the measurement equations, factor loadings corresponding to Std_xVelo_FS, LTI, and iPSD were set to one in the model.

Table 6 – Convergent and discriminant validity of the model

Factor	Convergent validity		Discriminant validity factor correlation estimates		
	AVE	*CR*	R_1	R_2	R_3
R_1	0.779	0.914	-	0.836*	0.412*
R_2	0.518	0.681	0.836*	-	0.736*
R_3	0.564	0.790	0.412*	0.736*	-
Square Root of *AVE*			0.883	0.720	0.751

* Statistically significant at the 0.001 level

ing trend in indicators and driving risk in the model (i.e. higher indicator value goes with higher driving risk), inverse PSD (iPSD), negative THW (nTHW) and negative PICUD (nPICUD) were utilised in the model. It should be noted that road section level indicators were calculated at 1 second before the observed moment (t_0) to account for the potential effect of flow propagation, while local traffic and vehicle level indicators were calculated at the observed moment (t_0) of the conflict sample as defined in Section 3.

From *Table 5*, it can be seen that all level indicators are statistically significant at the 0.001 level, and common model fit indicators including χ^2/df, GFI, RMSEA, Standardised RMR (SRMR), CFI and NFI can generally meet or are close to the suggestion value [33]. In addition, the standardised estimates of weights (or called factor loadings) of all indicators are larger than 0.60, suggesting these indicators have relatively strong correlation with their corresponding latent multi-level risks. Results above indicate that the established model fitting basically meets the requirement and can be used for the subsequent analysis.

Table 6 show that average variance extracted (AVE) of each latent factor is above 0.50 and their composite reliability (CR) is close to or over 0.70, indicating that the convergent validity of each factor construct is adequate [33], meaning the indicators within the same parent factor correlate well with each other, or each factor construct is well explained by its observed selected indicators. As to discriminant validity, the correlation estimates of all factor pairs range from 0.41 to 0.84 (statistically signifi-

cant at the 0.001 level), indicating that significant positive correlation can be observed between the latent risks, while no redundant risk construct is observed in the model (a latent factor correlation over 0.85 can raise redundancy concerns [33]). Specifically, except for the R_2 (local traffic level risk) column, the correlation estimates in the other two columns are lower than their corresponding square root of AVE of construct R_1 (road section level risk) and R_3 (vehicle level risk), respectively, indicating that there exists a certain distinction between the latent risks although they have significant inter-correlations. Moreover, the inter-correlations are relatively higher for R_2 and the other two latent risks, meaning it is difficult to completely separate the local traffic level risk from road section level risk or vehicle level risk, which is in accordance with what one would expect, as road section level traffic flow and vehicle level driving manoeuvre are supposed to interact with each other through the operating status of the local level vehicle groups.

Such modelling results above indicate that the theoretical concept of multi-level indicator based latent risk constructs is generally consistent with the observed results, which can enlighten on a more comprehensive framework of risk event prediction by including a set of multi-level indicators.

4.2 Feature importance of the multi-level indicators

To further examine the importance of the proposed multi-level indicators in risk event prediction, the technique of LightBGM feature importance scoring was employed. LightGBM, a decision tree-based learning algorithm under the gradient boosting framework, can deal with both the regression and classification problem [37], and can output the importance of features in solving the problem, which fits with the objective of this paper. Unlike most decision tree learning algorithms with the level-wise tree growth strategy, LightGBM grows trees through the leaf-wise strategy (i.e. only the leaf with the highest split gain is identified and split) and can generally achieve better model accuracy. Key parameters in the setting of LightGBM involve decision tree-based indicators, such as the number of leaf nodes (denoted as num_leaves, the main parameter that controls the complexity of the tree model) and the maximum depth of the tree (denoted as max_depth, an important parameter that handles the overfitting problem). Given the established

LightGBM model, the importance of a feature can be determined either by the frequency of the feature acting as split nodes with positive split gains, or by the total gains of splits that use the feature.

In our case, the conflict and non-conflict samples were labelled as risk and non-risk events (labelled as 1 and 0), respectively, the prediction of which can be well solved by a binary LightGBM classifier. To keep consistent with the previous CFA analysis, the 8 multi-level indicators in establishing the driving risk CFA model (*Table 4*) were fed into the binary LightBGM classifier as input features (all scaled to 0~1 by z-score normalisation) and the classifier output the predefined risk labels. Thus, different from the CFA analysis above, both conflict and non-conflict samples were utilised here to establish the LightBGM classifier, and the ratio of the number of risk samples and non-risk samples was set to 1:4 (i.e. non-risk samples are randomly under-sampled from the non-conflict samples obtained in Section 3) as previous research findings show that a better model fit can be obtained when the ratio of the case group to the control group is 1:4 [24].

To explore the potential differential importance of multi-level indicators across different prediction time, feature importance at 0 seconds, 1 second, 2 seconds and 3 seconds prediction horizons were calculated respectively for comparison. For example, for the 3 seconds prediction horizon, input features were calculated using the data observed 3 seconds earlier than the observation moment t_0 defined in Section 3. It should be noted, same as with the previous CFA analysis to reflect the flow propagation effect, that the baseline time for the calculation of road section level indicators are 1 second in advance of that for local traffic and vehicle level indicators. The parameters of LightBGM classifier were determined based on the K-fold cross validation, which can help avoid over-fitting of the model and yields more convincing results. Specifically, num_leaves and max_depth were set to 15 and 4, respectively, both L1 and L2 regularisation terms were set to 1.0 and other parameters were set to default values [37]. The performance of the established classifier was evaluated by typical indicators including precision, recall, their harmonic mean value F1-score and Area Under Curve (AUC) [38]. The value of these indicators ranges from 0 to 1, and a higher value generally means a better classification capability of the classifier.

Five-fold cross validation results of the established LightBGM are shown in *Figures 7 and 8.* *Figure 7* shows that AUC scores have a decreasing trend when the prediction time becomes earlier, which is as expected as more future uncertainties are involved with prolonged prediction horizon and prediction gets more complicated and difficult. For different prediction horizons, all AUC scores are above 0.9 level and F1 scores are close to 0.8 level, indicating that the established Light-BGM classifier has reasonable results and can be used for the subsequent feature importance analysis. The feature importance was then calculated as the total times the feature is utilised in the established classifier. The ranking of feature importance in *Figure 8* shows vehicle level indicators have a

dominant position in classifying driving risk when the prediction horizon is short, while local and road level indicators (especially for local level indicator avg_dist_group and road level indicator Std_xVelo_FS) generally take a higher importance rate when the prediction time gets earlier, which time period is of primary concern for many early warning applications. It suggests that such typical local and road indicators, such as the average gap distance between the subject vehicle and its surrounding vehicles, as well as the standard deviation of travelling speed of the front vehicles in the road section should be taken into account in order to obtain more timely prediction of the driving risk. As risk prediction is not the prime purpose of the paper, the manner of establishing a more accurate risk prediction model by effectively fusing the multi-level indicators will not be further discussed.

5. CONCLUSIONS

A safety assessment framework of multi-level risks, i.e. road section level, local area level and vehicle level risks, represented by indicators in different measurement dimensions, was proposed based on vehicle trajectory data from HighD. The relationship among the multi-level risks and indicators was examined using the CFA technique in SEM, where the multi-level indicators were processed as the observed reflections of the interrelated latent
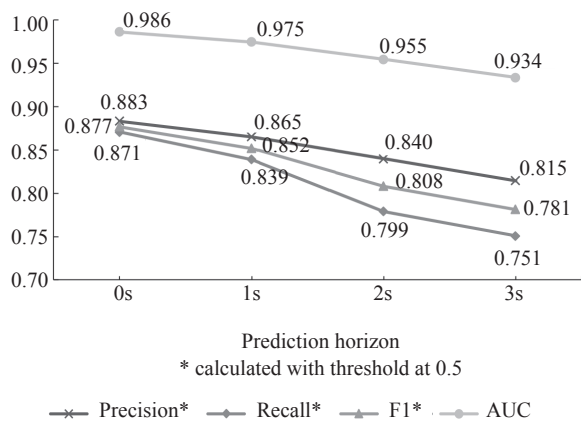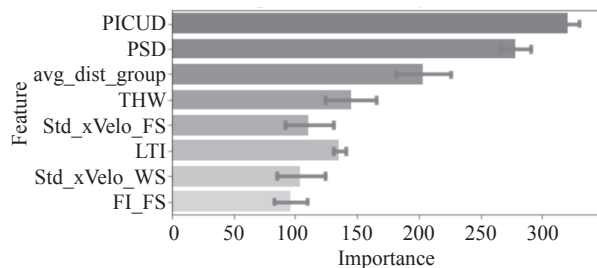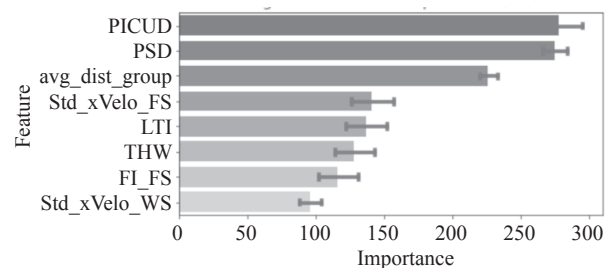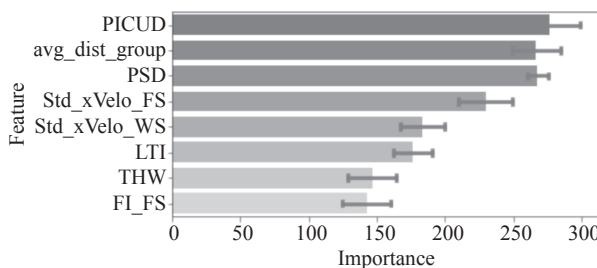

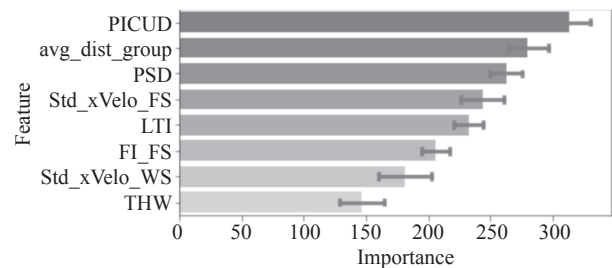*Figure 7 – LightGBM 5-fold cross validation scores*

a) LightGBM feature importance (0s)

b) LightGBM feature importance (1s)

c) LightGBM feature importance (2s)

d) LightGBM feature importance (3s)

Figure 8 – Feature importance by LightGBM (average over 5 folds)
Note: The horizontal line at the end of each bar represents the 95% confidence interval of the mean

multi-level risks. The final CFA model shows adequate model fitness, convergent and discriminant validity; its results show that significant positive correlation can be observed between the latent risks, and the distinction between the road section level risk and vehicle level risk is more pronounced than that between either of them and the local area level risk. The importance ranking of the proposed multi-level indicators was furthered explored by LightBGM, and the results indicate that local and road level indicators tend to be of more importance when the risk prediction time is sooner. To sum up, the proposed conceptual multi-level indicator based latent risk framework generally fits with the observed results, which suggests deficiencies in the existing driving safety assessment methodologies that are built solely based on single-level indicators, and emphasises the importance of including multi-level indicators for risk event prediction in the future.

There exist several limitations that need to be addressed in future research. Firstly, more temporal and spatial effect indicators should be examined for risk evaluation, such as differential effects of weekdays and weekends, basic and ramp segments of highway, as well as road geometric characteristics, which were not taken into account in the study due to the unavailability of these data. Secondly, currently only the interrelationship has been examined for the latent multi-level risks, while their potential causal relations can be further analysed with more conflict event samples. Also, given more sample data, considering the large number of multi-level indicators, exploratory factor analysis can be further employed to explore more latent factors within the current macro-meso-micro level risk framework. The methodology to effectively fuse different level indicators for risk prediction, such as conventional statistical models and machine-learning techniques, will also be explored in the future. The influence/performance of the recognised multi-level indicators and the obtained constructs in this work can also be compared across statistical models and machine-learning methods in risk event prediction, so as to obtain further insights on the proposed multi-level risk framework.

ACKNOWLEDGEMENTS

熊晓夏，博士[1]
E-mail: x_xiong623@163.com
何禹，硕士研究生[1]
E-mail: 2212104009@stmail.ujs.edu.cn
高翔，博士研究生[1]
E-mail: motgaoxiang@foxmail.com
赵叶灵，硕士[2]
E-mail: zhaoyeling2006@ujs.edu.cn
[1] 江苏大学汽车与交通工程学院，
   江苏省镇江市学府路301 号，212013
[2] 江苏大学财经学院，
   江苏省镇江市学府路301 号，212013

基于车辆轨迹的驾驶安全多层级风险评估框架

摘要

现有研究很少探讨高速公路交通安全系统中路段级别、局部区域级别和车辆级别风险之间的关系，而这对于构建有效的风险事件预测方法具有重要意义。本文在一组仔细挑选和设计的指标基础上，提出了一个多层级风险评估框架。基于车辆轨迹数据，采用结构方程模型（SEM）探索潜在的多层级风险与其可观测指标之间的相互关系。结果表明，潜在多层级风险变量之间具有显著的正相关关系，且每个潜变量都具有较好的聚合效度，但局部区域层级风险与路段层级风险或车辆层级风险之间的区别难以完全界定。基于 LightGBM 的特征重要性评分结果显示，当风险预测时间提早时，局部区域层级和和路段层级的指标将变得更为重要。本文提出的基于多层级指标的潜在风险框架总体上符合观测结果，并强调了在未来风险事件预测中包含多层级指标的重要性。

关键词
交通安全；多层级风险；安全指标；
结构方程模型；车辆轨迹

REFERENCES

[1] World Health Organization. *Road Traffic Injuries*. 2021. http://www.who.int/mediacentre/factsheets/fs358/en/.

[2] Shunying Z, et al. Review of research on traffic conflict techniques. *China Journal of Highway and Transport*. 2020;33(2): 15-33.

[3] Wang X, et al. Effect of daily car-following behaviors on urban roadway rear-end crashes and near-crashes: A naturalistic driving study. *Accident Analysis and Prevention*. 2022;164(November 2021): 106502. doi: 10.1016/j.aap.2021.106502.

[4] Chen S, et al. Risky driving behavior recognition based on vehicle trajectory. *International Journal of Environmental Research and Public Health*. 2021;18(23). doi: 10.3390/ijerph182312373.

[5] Bastos JT, et al. Naturalistic driving study in Brazil: An analysis of mobile phone use behavior while driving. *International Journal of Environmental Research and Public Health.* 2020;17(17): 1-14. doi: 10.3390/ijerph17176412.

[6] Mahmud SMS et al. Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. *IATSS Research.* 2017;41(4): 153-163. doi: 10.1016/j.iatssr.2017.02.001.

[7] Wang C, et al. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. *Accident Analysis and Prevention.* 2021;157(May): 106157. doi: 10.1016/j.aap.2021.106157.

[8] Hayward JC. Near-miss determination through use of a scale of danger. *Highway Research Record.* 1972;384: 24-34.

[9] Michael PG, Leeming FC, Dwyer WO. Headway on urban streets: Observational data and an intervention to decrease tailgating. *Transportation Research Part F: Traffic Psychology and Behaviour.* 2000;3(2): 55-64. doi: 10.1016/S1369-8478(00)00015-2.

[10] Allen BL, Shin BT, Cooper PJ. Analysis of traffic conflicts and collisions. *Transportation Research Record.* 1978;(667): 67-74.

[11] Astarita V, et al. A new microsimulation model for the evaluation of traffic safety performances. *Trasporti Europei.* 2012;(51): 16.

[12] Van Beinum A, et al. Critical assessment of methodologies for operations and safety evaluations of freeway turbulence. *Transportation Research Record.* 2016;(2556): 39-48.

[13] Almqvist S, Hyden C, Risser R. Use of speed limiters in cars for increased safety and a better environment. *Transportation Research Record.* 1991;(1318): 34-39.

[14] Park H, et al. Development of a lane change risk index using vehicle trajectory data. *Accident Analysis and Prevention.* 2018;110(October 2017): 1-8. doi: 10.1016/j.aap.2017.10.015.

[15] Chen Q, et al. Modeling accident risks in different lane-changing behavioral patterns. *Analytic Methods in Accident Research.* 2021;30. doi: 10.1016/j.amar.2021.100159.

[16] Jiang R, et al. Determining an improved traffic conflict indicator for highway safety estimation based on vehicle trajectory data. *Sustainability.* 2021;13(16). doi: 10.3390/su13169278.

[17] Yang D, et al. Fusing crash data and surrogate safety measures for safety assessment: Development of a structural equation model with conditional autoregressive spatial effect and random parameters. *Accident Analysis and Prevention.* 2021;152. doi: 10.1016/j.aap.2021.105971.

[18] Orsini F, et al. Large-scale road safety evaluation using extreme value theory. *IET Intelligent Transport Systems.* 2020;14(9): 1004-1012. doi: 10.1049/iet-its.2019.0633.

[19] Yue Z, et al. Detecting unsafe driving patterns using discriminative learning. *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, ICME 2007.* 2007: p. 1431-1434. doi: 10.1109/ICME.2007.4284929.

[20] Ning H, et al. A general framework to detect unsafe system states from multisensor data stream. *IEEE Transactions on Intelligent Transportation Systems.* 2010;11(1): 4-15. doi: 10.1109/TITS.2009.2026446.

[21] Lee D, Yeo H. Real-time rear-end collision-warning system using a multilayer perceptron neural network. *IEEE Transactions on Intelligent Transportation Systems.* 2016;17(11): 3087-3097. doi: 10.1109/TITS.2016.2537878.

[22] Fu Y, et al. Graded warning for rear-end collision: An artificial intelligence-aided algorithm. *IEEE Transactions on Intelligent Transportation Systems.* 2020;21(2): 565-579. doi: 10.1109/TITS.2019.2897687.

[23] Yu R, Han L, Zhang H. Trajectory data based freeway high-risk events prediction and its influencing factors analyses. *Accident Analysis & Prevention.* 2021;154: 106085. doi: 10.1016/J.AAP.2021.106085.

[24] Hu Y, et al. A high-resolution trajectory data driven method for real-time evaluation of traffic safety. *Accident Analysis & Prevention.* 2021;165. doi: 10.1016/j.aap.2021.106503.

[25] Chen Q, et al. Using vehicular trajectory data to explore risky factors and unobserved heterogeneity during lane-changing. *Accident Analysis & Prevention.* 2021;151(July 2020). doi: 10.1016/j.aap.2020.105871.

[26] Bollen KA. *Structural Equations with Latent Variables.* 2014. doi: 10.1002/9781118619179.

[27] Liang M, et al. Learning lane graph representations for motion forecasting. In:Vedaldi A, Bischof H, Brox T, Frahm JM. (eds) *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12347.* Springer, Cham; 2020; p. 541-556. doi: 10.1007/978-3-030-58536-5_32.

[28] Li Y, Chen Y. Driver vision based perception-response time prediction and assistance model on mountain highway curve. *International Journal of Environmental Research and Public Health.* 2016;14(1): 31. doi: 10.3390/IJERPH14010031.

[29] Kemper A. Complex networks theory. *Contributions to Management Science.* 2010. p. 135-157. doi: 10.1007/978-3-7908-2367-7_10.

[30] Hu H, et al. Cost-sensitive semi-supervised deep learning to assess driving risk by application of naturalistic vehicle trajectories. *Expert Systems with Applications.* 2021;178(April): 115041. doi: 10.1016/j.eswa.2021.115041.

[31] Shannon CE. A Mathematical theory of communication. *Bell System Technical Journal.* 1948;27(3): 379-423. doi: 10.1002/J.1538-7305.1948.TB01338.X.

[32] Javid MA, et al. Structural equation modeling of drivers' speeding behavior in Lahore: Importance of attitudes, personality traits, behavioral control, and traffic awareness. *Iranian Journal of Science and Technology - Transactions of Civil Engineering.* 2022;46(2): 1607-1619. doi: 10.1007/S40996-021-00672-1/FIGURES/1.

[33] Hau J, Wen Z, Cheng Z. *Structural equation models and their applications* (Rev. ed.). Beijing: Educational Science Publishing House; 2021. p. 29-60.

[34] Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables. *Journal of Marketing Research.* 1981;XVIII(February): 39-50.

[35] Krajewski R, et al. The highD dataset: A drone dataset

of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC.* 2018;2018-November: 2118-2125. doi: 10.1109/ITSC.2018.8569552.

[36] The highway drone (highD) dataset. RWTH Aachen University; 2018. https://www.highd-dataset.com.

[37] Thakkar JJ. Applications of structural equation modelling with AMOS 21, IBM SPSS. *Studies in Systems, Decision and Control.* 2020;285: 35-89. doi: 10.1007/978-981-15-3793-6_4/FIGURES/55.

[38] Morgan BJT, McNicol D, Freeman PR. A primer of signal detection theory. *The Statistician.* 1976;25(3): 231. doi: 10.2307/2987842.