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 indicator; 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 logistic 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
In this 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 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...
Xiong X, He Y, Gao X, Zhao Y. A Multi-Level Risk Framework for Driving Safety Assessment Based on Vehicle Trajectory

962
Promet – Traffic&Transportation, Vol. 34, 2022, No. 6, 959-973

graph, denoted as \( C \),
\[ c_k^{1 \leq k \leq K} \]
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) \[
(LL_i \cdot LL_j) |\{(c_k \in C | (c_k \in LL_i) \& (c_k \in LL_j)) \cdot ((x_k, y_k) \cdot (x_i, y_i)) < \max (x_k, y_k)\cdot (x_i, y_i)\} = 0
\]

2) \[
(LL_i \cdot LL_j) |\{\zeta \in (LL_i \cdot LL_j) \cdot |\{(c_k \in LL_i) \& (c_k \in LL_j) \cdot (y_k, y_i) < 0 \lor (LL_i \cdot LL_j) \cdot (y_k, y_i) > 0 \lor (LL_i \cdot LL_j) \cdot y_i < 0\} = 0
\]

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 \( d \)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Indicator variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td>avg _xVelo_FS, std _xVelo_FS, vc _xVelo_FS, avg_dist_FS, min_dist_FS, FI_FS</td>
<td>avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for FS</td>
</tr>
<tr>
<td>FL</td>
<td>avg _xVelo_FL, std _xVelo_FL, vc _xVelo_FL, avg_dist_FL, min_dist_FL, FI_FL</td>
<td>avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for FL</td>
</tr>
<tr>
<td>WS</td>
<td>avg _xVelo_WS, std _xVelo_WS, vc _xVelo_WS, avg_dist_WS, min_dist_WS, FI_WS</td>
<td>avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for WS</td>
</tr>
<tr>
<td>WL</td>
<td>avg _xVelo_WL, std _xVelo_WL, vc _xVelo_WL, avg_dist_WL, min_dist_WL, FI_WL</td>
<td>avg, S.D., V.C. of speed, avg dist., min. dist. between vehs, and flow instability index for WL</td>
</tr>
</tbody>
</table>

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 \( C = \{c_k \}^{K}_{k=1} \), 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) \[
(LL_i \cdot LL_j) |\{(c_k \in C | (c_k \in LL_i) \& (c_k \in LL_j)) \cdot ((x_k, y_k) \cdot (x_i, y_i)) < \max (x_k, y_k)\cdot (x_i, y_i)\} = 0
\]

2) \[
(LL_i \cdot LL_j) |\{\zeta \in (LL_i \cdot LL_j) \cdot |\{(c_k \in LL_i) \& (c_k \in LL_j) \cdot (y_k, y_i) < 0 \lor (LL_i \cdot LL_j) \cdot (y_k, y_i) > 0 \lor (LL_i \cdot LL_j) \cdot y_i < 0\} = 0
\]

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 \( d \)
(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 \(c_5-c_7\) and \(c_8-c_{10}\) 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:

\[
w_{ij} = \frac{|v_{x_i} - v_{x_j}|}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}
\]

(1)

**Calculation of the traffic flow instability index**

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

\[
\omega_i = \frac{\sum_{j \in N_i} w_{ij}}{d_i}
\]

(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.

\[
FI = \frac{\sum_{i=1}^{K} \frac{d_i^2}{\sum_{j=1}^{K} d_j^2} \omega_i}{K}
\]

(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. \(c_2, c_4, ..., c_9\) 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]:

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

(4)

\[
X_j = \begin{bmatrix} S_0^x \ y_j \ v_j^x \ v_j^y \end{bmatrix}^T
\]

(5)

where \(X_j\) measures the relative longitudinal and lateral time distance between the subject vehicle \(c_6\) and its surrounding vehicle \(c_j\). \(S_0^x = |x_j - x_0| - (L_0 + L_j)/2\) and \(S_0^y = |y_j - y_0| - (U_0 + U_j)/2\) 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. \(\mu = \text{diag}(\sigma_x, \sigma_y)\) designates the mean and covariance of the two-dimensional Gaussian distribution and is set to \(\mu = (0, 0)^T\) and \(\Lambda = \text{diag}(1, 1)\) 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:

\[
LTI = \sum_{j=0}^{n_s} \left| v_{x_j} - v_{x_0} \right| \log(f(X_j; \mu, \Lambda))
\]

(6)

where \(n_s\) 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

<table>
<thead>
<tr>
<th>Indicator variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg_xVelo_GP, std_xVelo_GP, vc_xVelo_GP, avg_dist_GP,</td>
<td>avg, S.D., V.C. of speed, avg dist., and min. dist. between (</td>
</tr>
<tr>
<td>min_dist_GP, LTI</td>
<td>Local traffic instability index</td>
</tr>
</tbody>
</table>

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]:

\[
DRAC_j = \frac{(v_{c0} - v_{cj})^2}{2S_j} \tag{7}
\]

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:

\[
MDI = \frac{DRAC_j}{MADR_j} \tag{8}
\]

where \( MADR_j \) 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 3.4 m/s\(^2\) for cars and 2.4 m/s\(^2\) 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].

\[
PSD = \frac{S_j}{v_{c0}^2 / 2MADR_j^2} \tag{9}
\]

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].

\[
PICUD = S_j^0 + \frac{v_{c0}^2}{2MADR_j^2} - \frac{v_{c0}^2}{2MADR_j^2} \cdot t_h \cdot v_{c0} \tag{10}
\]

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].

\[
THW = \frac{S_j^0}{v_{c0}} \tag{11}
\]

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.
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:

\[ I_{ij} = \lambda_{ij} R_i + e_{ij} \]  

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 \( j \)th observed variable (i.e. the specified indicators) of the \( i \)th risk latent variable. \( \lambda_{ij} \) represents factor loading, which measures how much of the variability in the \( j \)th observed indicator is explained by the \( i \)th level risk status; it can also be interpreted as the amount of unit change in the observed indicator \( I_{ij} \) when the latent risk status \( R_i \) undertakes one unit change. \( e_{ij} \) designates the error term for the \( j \)th observed indicator of the \( i \)th 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:

\[ I = \Lambda R + \epsilon \]

where

\[ I = \begin{bmatrix} I_{11} \\ I_{12} \\ \vdots \\ I_{21} \\ I_{22} \\ \vdots \\ I_{31} \\ I_{32} \\ \vdots \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ \vdots & \vdots & \vdots \\ \lambda_{21} & 0 & 0 \\ \lambda_{22} & 0 & 0 \\ \vdots & \vdots & \vdots \\ \lambda_{31} & 0 & 0 \\ \lambda_{32} & 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{21} \\ e_{22} \\ \vdots \\ e_{31} \\ e_{32} \\ \vdots \end{bmatrix} \]

(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:

\[ E(R) = 0, \quad Cov(R) = \Psi = \begin{bmatrix} 1 & \Psi_{31} \\ \Psi_{13} & 1 \end{bmatrix} \]

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.

Table 3 – Road section level indicators for consideration

<table>
<thead>
<tr>
<th>Measurement category</th>
<th>Indicator variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deceleration-based</td>
<td>DRAC</td>
<td>Deceleration rate to avoid the crash</td>
</tr>
<tr>
<td></td>
<td>MDI</td>
<td>Maximum deceleration index</td>
</tr>
<tr>
<td>Distance-based</td>
<td>PSD</td>
<td>Proportion of stopping distance</td>
</tr>
<tr>
<td></td>
<td>PICUD</td>
<td>Potential indicator of collision with urgent deceleration</td>
</tr>
<tr>
<td>Time-based</td>
<td>THW</td>
<td>Time Headway</td>
</tr>
</tbody>
</table>

![Figure 4 – Conceptual multi-level driving risk model](image-url)
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 standardize 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].

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.

![Figure 5 – HighD coordinate system](image-url)
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 fitness 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**

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

<table>
<thead>
<tr>
<th>Factor</th>
<th>Convergent validity</th>
<th>Discriminant validity factor correlation estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVE</td>
<td>CR</td>
</tr>
<tr>
<td>R&lt;sub&gt;1&lt;/sub&gt;</td>
<td>0.779</td>
<td>0.914</td>
</tr>
<tr>
<td>R&lt;sub&gt;2&lt;/sub&gt;</td>
<td>0.518</td>
<td>0.681</td>
</tr>
<tr>
<td>R&lt;sub&gt;3&lt;/sub&gt;</td>
<td>0.564</td>
<td>0.790</td>
</tr>
<tr>
<td>Square Root of AVE</td>
<td>0.883</td>
<td>0.720</td>
</tr>
</tbody>
</table>

* Statistically significant at the 0.001 level

\[ \chi^2/df=4.361<5, \ GFI=0.957>0.900, \ RMSEA=0.097<0.100, \ SRMR=0.0510.050, \ CFI=0.967>0.900, \ NFI=0.958>0.900 \]

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

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 \( \chi^2/df \), 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-
dles the overfitting problem). Given the established model) and the maximum depth of the tree (denoted as max_depth, an important parameter that han 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 LightGBM 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.
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](image_url)

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

![Figure 8](image_url)

**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 further explored by LightGBM, 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

This research was funded by the National Nature Science Foundation of China (52002154, 52072160, and U20A20331). The authors would like to thank the Institute for Automotive Engineering (ika), RWTH Aachen University for sharing their HighD dataset.

REFERENCES


