Modeling conflict risk with real-time traffic data for road safety assessment: a copula-based joint approach

Yeping Hu1, Ye Li1,*, Chen Yuan1,2 and Helai Huang1

1School of Traffic and Transportation Engineering, Central South University, Changsha 410075, Hunan, China; 2Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong 999077, China.

*Corresponding author. E-mail: yelicsu@csu.edu.cn

Abstract
This study proposes a conflict-based traffic safety assessment method by associating conflict frequency and severity with short-term traffic characteristics. Instead of analysing historical crash data, this study employs microscopic trajectory data to quantify the relationship between conflict risk and traffic characteristics. The time-to-collision (TTC) index is used to detect conflicts, and a severity index (SI) is proposed on the basis of time-integrated TTC. With SI, the k-means algorithm is applied to classify the conflict severity level. Then the severity of regional conflict risk is split to three levels. Zero truncated Poisson regression and ordered logit regression methods are employed to estimate the effects of short-term traffic characteristics on conflict frequency and severity, respectively. Furthermore, the copula-based joint modelling method is applied to explore the potential non-linear dependency of conflict risk outcomes. A total of 18 copula models are tested to select the optimal ones. The HighD dataset from Germany is utilized to examine the proposed framework. Both between-lane and within-lane factors are considered. Results show that the correlations between traffic characteristics and conflict risk are significant, and the dependency of conflict outcomes varies among different severity levels. The difference of speed variation between lanes significantly influences the conflict frequency and severity simultaneously. Findings indicate that the proposed method is practicable to assess real-time traffic safety within a specific region by using short-term (30-second time interval) traffic characteristics. This study also contributes to develop targeted proactive safety strategies by evaluating road safety based on conflict risk, and considering different severity levels.

Keywords: Conflict frequency, conflict severity, real-time safety, copula model

1. Introduction
Since crash events are public health problems, there are considerable studies exploring the crash mechanism and how to improve road safety. Most previous safety analysis is based on crash data, which have indicated that traffic characteristics are significant influential factors of road crashes [1]. With the development of intelligent transportation systems, the demand for a proactive safety technique is urgent. Recent studies have proposed that traffic conflict is an effective surrogate safety measure [2]. Some limitations of historical crash data (e.g. under-reporting, long-term collection) can be overcome in the safety analysis by virtue of the traffic conflict technique (TCT) [3, 4]. Conflict-based research has the potential to detect unsafe traffic dynamics before the crash occurrence, which is also beneficial for proactive safety management.

However, current conflict-based analysis mostly considered the microscopic perspective of traffic safety. Individual conflict occurrence was often emphasized, and mainly investigated by utilizing trajectory data and micro-simulation data [5, 6]. There is a lack of relevant work applying TCT to assess real-time road safety toward some macroscopic aspects. It is known that a conflict happens more frequently compared to a crash, so conflicts collection can be achieved in shorter period. Particularly, crash data are difficult to be collected at one newly operated road segment. It may be practicable to estimate the collision risk by quantifying the relationship between traffic characteristics and conflict risk. Additionally, the time-to-collision (TTC) index has been widely used to identify traffic conflicts [7]. Less but more severe conflicts could be detected with a smaller TTC threshold. In fact, frequency and severity are both important aspects, which need to be together considered in conflict-based analysis. Similar to the previous crash-based studies [8–10], some effective statistical methods can be employed to examine that interaction relationship. Previous joint modeling analysis of crash frequency and severity has been a topic worth discussing. The method can provide a more comprehensive and systematic understanding of influential mechanism. In terms of conflict-based research, it is a promising approach in safety analysis, but existing relative work is limited.

Therefore, this study aims to design an approach of real-time conflict risk assessment from a new perspective. The investigation of relationship between conflict and real-time traffic characteristics can fill the current research gap. Alternative to individual conflict occurrence, this study focuses on lane-level conflict risk within specific time and space. The influential mechanism of short-term traffic characteristics and road safety will be better known, if that relationship is validated. The objectives are: (i) associating the vehicular trajectories with regional conflict risk; (ii) exploring the effects of short-term traffic characteristics on conflict risk; and (iii) finding the potential dependence of conflict frequency and severity. This study provides a new idea of...
evaluating the real-time road safety, and also contributes to the design of proactive countermeasures.

2. Literature review

Traditionally, researchers have explored the influential factors of crash frequency or severity at a macroscopic level [8, 9]. Analysing crash frequency is based on some count-data models methods (Poisson, negative binomial, etc.), and Lord and Manning [10] provided a systematic review of that. When the dependent variables are count, discrete, and non-zero, Poisson and Zero-truncated Poisson (ZTP) models were compared, and ZTP model was demonstrated to own a considerable advantage [11]. Crash severity is a measure of the societal impact and harm to the society, and the modelling process is mostly based on logistic regression. For example, Jung et al. [12] combined vehicle-to-vehicle crash frequency and severity estimations to examine factor impacts on highway safety in rainy weather, using a negative binomial regression and multinomial logit regression for crash frequency and severity modeling, respectively. As crash data are often classified according to the injury severity or collision type, some researchers attempted to explore the interdependence and develop joint-modelling methods to overcome the limitations of previous analysis [13–15]. The copula-based approach is verified as an efficient way to realize bivariate joint-modelling.

The major strength of using copulas is that the process of estimating marginal distributions is separate from the dependence structure estimation. Zou et al. [16] considered the underreporting in wildlife-vehicle collision (WVC) data, and proposed a copula regression model linking WVC and the underreporting outcome, which combined the logistic and negative binomial regression together. Stipancic et al. [17] used incorporating GPS-derived surrogate safety measures (SSMs) as predictive variables, then developed a full Bayes spatial negative binomial model for crash frequency and a fractional multinomial logit model for crash severity. Yang et al. [18] proposed a multivariate copula-based framework to model crash count and conflict risk measures jointly, which promoted the understanding of the exposure and traffic risk factors and their heterogenous impact on crash count and conflict risk measures. However, copula-based modelling of regional conflict risk has not been conducted in safety literature. Previous studies demonstrated that there was a strong relationship between crashes and near crashes. The use of near crashes as a crash surrogate for risk assessment was beneficial [19, 20]. Nevertheless, the conflict-based analysis mostly focused on vehicular safety. Recently, some researchers have attached more attention to associating conflicts with the evaluation of segment-level safety. El-Basyouny and Sayed [21] proposed a two-phase model, where a conflict-based negative binomial (NB) safety performance function (SPF) was employed to predict collisions in the second phase. Arun et al. [22] estimated the frequency of severe and non-severe crashes by jointly modelling the indicators of crash frequency and bivariate peak-over-threshold models for both TTC and Delta-V were estimated. As for conflict severity, different thresholds of conflict indicators were proposed to measure the severity of one conflict event [23]. Thus, how to measure conflict severity within a specific unit needs to be clarified. Conflict counts and severity are considered simultaneously in this study, which also promotes understanding of the systematic effects of real-time traffic characteristics.

3. Methodology

The proposed analysis methods include two stages, where the conflict frequency and severity are quantified independently in the first stage, and the interdependence between the two outcome variables is explored in the second stage. The overall framework of this study, including data processing and data analysis, is presented in Fig. 1.

3.1 Conflict risk indicators

In order to establish the interrelationship between traffic characteristics and conflicts, the aforementioned TTC index was utilized in this study TTC is one of the temporally proximal based indicators that extensively applied in road safety analysis [24, 25]. Fig. 2(a) presents the principle of TTC calculation. By obtaining position and speed information from microscopic vehicle trajectories, the TTC value of a target vehicle i at instant t can be calculated according to Eq. (1).

\[
TTC_i(t) = \begin{cases} 
\frac{x_i(t) - x_i(t-1)}{v_i(t)} & \text{if } v_i(t) > v_{i-1}(t), \\
\infty & \text{if } v_i(t) \leq v_{i-1}(t).
\end{cases}
\]

where \(x_i(t)\) and \(v_i(t)\) are the position and velocity of targeted (following) vehicle \(i\) at timestamp \(t\), and \(L_{i-1}\) is the length of preceding vehicle \(i - 1\).

Generally, a threshold of TTC will be used to determine whether the condition is safe [26–28], and 4 s was used in this study. If the TTC value is lower than the critical threshold, it is considered that a probable conflict exists. Thus, the conflict count data can be collected, which represent the first outcome variable. To measure the severity level of regional conflict risk, the time-integrated TTC (ITT) index is further utilized. A high ITT value means a large exposition time to duration-weighted unsafe TTC values, which is negative for road safety [29]. ITT suggests a risk accumulation in both risk severity and exposure duration, and the risk value can be defined by the average value of the risk integration, as presented
in Fig 2(b). The individual TTC for subject i in discrete time can be calculated with Eq. (2), and Eq. (3) denotes the calculation of an average risk value.

\[ TIT_i = \sum_{t=1}^{T} (TTC_i - TTC(\Delta t)) \cdot \tau_s, \quad 0 \leq TTC_i(\Delta t) \leq TTC^* \]

\[ \text{Risk Value} = \frac{TIT_i}{\Delta t} \]

where TTC\(^*\) is the TTC threshold (i.e. 4 s), \( \tau_s \) is the time scan interval and \( \Delta t \) is the duration of unsafe TTC values. For one spatio-temporal unit, there may be a total of \( N \) conflicts \((N = 1, 2, 3, \ldots)\). We just focus on the most severe event with the largest risk value. Then, the severity index (SI) was proposed to show the severity of one observation, which can be calculated as follows:

\[ SI = \max \text{(Risk Value)} = \max \left( \frac{TIT_i}{\Delta t} \right), \quad j = 1, 2, 3, \ldots N \]

Compared to other clustering techniques like k-median and hierarchical, k-means is used more widely \([6,30,31]\) and has a good interpretability. The k-means method focuses on the smallest difference within each cluster and largest difference between clusters. With SI, the severity level can be classified simply by k-means algorithm as expected.

### 3.2 Modeling the effects of traffic characteristics on conflict risk

#### 3.2.1 Conflict frequency model

The ZTP regression method was used for modeling, which is suitable for modeling non-zero count data \([11]\). The modeling process includes variable correlation test and examining the variables’ significance, and then insignificant variables are subsequently removed from the model. The probability function of ZTP distribution is as follows:

\[ p(y_1|y_1 > 0) = \frac{p(y_1)}{1 - p(y_1)} = \frac{e^{-\lambda y_1}}{y_1!(1-e^{-\lambda})} = \frac{\lambda^{y_1}}{y_1!(e^{-\lambda}-1)}, \quad y_1 = 1, 2, 3, \ldots \]

\[ E(y_1) = \frac{\lambda e^\lambda}{e^\lambda - 1} \]

\[ Var(y_1) = \frac{\lambda e^\lambda}{e^\lambda - 1} \left( 1 - \frac{\lambda}{e^\lambda - 1} \right) \]

\[ g(\lambda) = \log(\lambda) = X \beta_i + \epsilon_i \]

where \( y_1 \) is considered as a count variable in a specific time interval, \( X \) and \( \beta_i \) are the design matrix and the vector of regression coefficients, respectively; \( g(\lambda) \) is the link function of ZTP; \( \epsilon_i \) is random error, which follows the standard logistic distribution.

#### 3.2.2 Conflict severity model

For severity estimation, the typical ordered logit (OL) model was employed in this study, which is formulated as follows:

\[ y_2^* = X \beta_2 + \epsilon_2 \]

where \( y_2^* \) is a continuous latent variable measuring the severity of observed regional conflict risk; \( X \) means the same input explanatory variables as that in ZTP model; \( \beta_2 \) is the corresponding coefficient estimates, and \( \epsilon_2 \) is random error.

\[ y_2 = \begin{cases} 
1, & \text{if } y_2^* \leq \mu_1 \\
2, & \text{if } \mu_1 < y_2^* \leq \mu_2 \\
3, & \text{if } y_2^* \geq \mu_2 
\end{cases} \]

where \( \mu_1 \) and \( \mu_2 \) are the threshold values to be estimated, and the discrete variable \( y \) is assumed to be a mapping of \( y_2^* \). The probability of \( y_2 = s \) \((s = 1, 2, 3)\) equals the probability of that the \( y_2^* \) value is within corresponding threshold range. Therefore, the probability of each severity level can be calculated as follows:

\[ p(y_2 = s | \text{conflict}) = F(\mu_{s-1} - \beta, X) - F(\mu_{s-2} - \beta, X), \quad s = 1, 2, 3 \]

where \( F(\lambda) \) denotes the cumulative distribution function (standard logistic) of the random error term. \( F(\mu_{s-1} - \beta, X) \) is assumed to be 0 and 1, when \( s = 0 \) and \( s = 3 \), respectively.

### 3.3 Joint modelling of conflict risk level and count data

The occurrence of conflicts and the probability of different conflict risk levels are affected by the explanatory variables. Conflict severity is replaced by three binary variables here, and the modelling method is binary logit regression. Hence, there are three joint modelling processes. To capture the interdependence, a bivariate copula approach is employed to describe the correlation. The joint cumulative density function (cdf) can be defined as follows:

\[ F(y, z) = C_{	heta}(u = F_1(y), v = F_2(z)) \]

where \( C_{\theta} \) is a cdf function that joins \( F_1(y) \) and \( F_2(z) \) into a joint distribution; \( F_1(y) \) denotes the marginal distribution (binary logit model) of each severity level variable; \( F_2(z) \) is the marginal distribution (ZTP model) of the count variable.

Different families of bivariate copulas were compared, including Gaussian, Clayton, Frank, Gumbel, Ali-Mikhail-Haq (AMH) and Farlie-Gumbel-Morgenstern (FGM). Characteristics of those copulas have been introduced in previous studies \([16]\), which are displayed in Table 1.
4.1 Study area

The vehicle trajectory data were collected from the HighD dataset, which provides post-processed trajectories of 110,000 vehicles from 60 sub-datasets (4K resolution, 25 Hz) at six different locations on freeways near Cologne, Germany, during 2017 and 2018 [33]. All data were collected during sunny and windless weather from 8 AM to 5 PM to maximize the quality of data. The average length of the road segments in each location is about 420 m. The detailed information included the recorded frame (25 frames per second), vehicle ID, vehicle type, longitudinal (x) and lateral (y) positions, velocities, driving direction, surrounding vehicles’ ID, TTC, lane position. More details can be seen on the HighD website [34].

A 30-second time interval was used to collect short-term vehicle information at the lane level without any on-ramps and off-ramps. Fig. 3 displays the schematic diagram of data extraction. There are two freeway types included in the HighD dataset, so five types of the lane layouts were defined. For instance, as shown in Fig. 3, Inner_1 variable denotes the inner lane of a six-lane two-way freeway, with 11 as the corresponding type ID. The Middle_1 (12), Outer_1 (21), Inner_2 (13), Inner_2 (22) and Outer_2 (22) variables are similarly defined. The selected variables include traffic characteristics (e.g. volume, speed) and other factors. Lane-based data were collected from 325 lanes in the HighD data set during a 30-second time interval.

3.4 Model evaluation indices

There are many methods for comparing models and selecting the best model, among which the most popular ones are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [32]. The smaller amounts of AIC and BIC indices indicate better fitting of the models. These information criteria are in the following forms:

\[
AIC = -2 \log L(\hat{\theta}) + 2k \quad (13)
\]

\[
BIC = -2 \log L(\hat{\theta}) + \log(n)k \quad (14)
\]

where \(p\), \(k\) and \(L\) are the number of parameters in the model, the number of observations (samples) and the likelihood of the models, respectively.

Kendall’s tau (\(\tau\)) is also utilized to characterize and compare the dependency structure in different copula models. A zero-value Kendall’s \(\tau\) stands for independence (i.e. no dependency). It is an effective indicator that satisfies the properties for assessing the dependency between random variables and detailed information can be referred to the previous study [15].

4.2 Data description

The TTC threshold of 4 s was used in this study, thus the conflict events will be detected if \(0 < \text{TTC} < 4\) s. Given the focus of this study, a total of 1,749 observations involving conflicts were analysed. It should be noted that observations without conflict occurrence were not included. Table 2 displays the amount of origin data and the filtered data collected from each type of lane. Also, the quantity of corresponding data points is presented. For the origin data, there are 47 video recordings of six-lane two-way roads, and the number of data points corresponding to the three types

<table>
<thead>
<tr>
<th>Model</th>
<th>Copula (C(u, v; \theta))</th>
<th>Parameter range of (\theta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>(\phi_u(\phi^{-1}(u), \phi^{-1}(v); \theta))</td>
<td>(\theta \in (-1, 1)), (\theta = 0) is independence</td>
</tr>
<tr>
<td>Clayton</td>
<td>(\psi_u^{-1}(u)\phi_u^{-1}(v) - \psi_u^{-1}(u)\phi_u^{-1}(v) + k)</td>
<td>(\theta \in (0, \infty)), (\theta \to 0) is independence</td>
</tr>
<tr>
<td>Frank</td>
<td>(-\frac{1}{\theta} \ln(1 + u^{-\theta} - 1)^{-1} \phi_u^{-1}(v))</td>
<td>(\theta \in (-\infty, \infty)), (\theta \to 0) is independence</td>
</tr>
<tr>
<td>Gumbel</td>
<td>(\exp(-[(\ln u)^{\phi} + (\ln v)^{\phi}]^{1/\phi}))</td>
<td>(\theta \in (1, \infty)), (\theta = 1) is independence</td>
</tr>
<tr>
<td>AMH</td>
<td>(u^{\phi}[(1-u)/\phi - 1) \phi^{-1}(v)])</td>
<td>(\theta \in (-1, 1)), (\theta \to 0) is independence</td>
</tr>
<tr>
<td>FGM</td>
<td>(\psi_u[1 + \theta(1-u)(1-v)])</td>
<td>(\theta \in (-1, 1)), (\theta = 0) is independence</td>
</tr>
</tbody>
</table>

Note: \(\phi\) represents the standard cdf of bivariate normal distribution and \(\phi^{-1}\) denotes the inverse cdf of the standard univariate normal distribution.
Table 2. Overview of data distribution at different lanes.

<table>
<thead>
<tr>
<th>Description</th>
<th>Lane type (type ID)</th>
<th>Data points</th>
<th>Origin data</th>
<th>Filtered data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner lane at six-lane two-way road</td>
<td>Inner_1 (11)</td>
<td>91</td>
<td>3,129</td>
<td>216</td>
</tr>
<tr>
<td>Middle lane at six-lane two-way road</td>
<td>Middle_1 (12)</td>
<td>91</td>
<td>3,129</td>
<td>951</td>
</tr>
<tr>
<td>Outer lane at six-lane two-way road</td>
<td>Outer_1 (13)</td>
<td>13</td>
<td>756</td>
<td>319</td>
</tr>
<tr>
<td>Inner lane at four-lane two-way road</td>
<td>Inner_2 (21)</td>
<td>13</td>
<td>756</td>
<td>61</td>
</tr>
<tr>
<td>Outer lane at four-lane two-way road</td>
<td>Outer_2 (21)</td>
<td>13</td>
<td>756</td>
<td>202</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>325</td>
<td>10,899</td>
<td>1,749</td>
</tr>
</tbody>
</table>

Fig. 4. SI range and conflict frequency distribution of each severity level.

is the same. There are 13 video recordings of four-lane two-way roads. Similarly, data points of two corresponding types are the same. For the filtered data, there are some differences.

With SI value, three severity levels (Details of k-means clustering results are illustrated in Appendix A.) were classified by the k-means clustering algorithm. The overall range of SI value is from 0.25 to 94.1, and the range of each severity level is marked in Fig. 4. Level 1 denotes the lowest severity level, while level 2 is medium and level 3 represents the highest. For level 1, the amount of samples is 997, while level 2 is 610 and level 3 is 142. Conflict frequency distribution at each severity level is also presented in Fig. 4. It indicates that the range of conflict frequency is from 1 to 13, and samples only including one conflict occurrence occupy a large proportion.

The description statistics of variables and details of conflict severity clusters are presented in Table 3. The between-lanes factors were considered. For two-way four-lane roads, the difference of between-lane traffic characteristics is easy to calculate. As for two-way six-lane roads, if the subject lane is the on the middle lane, the difference will use an average value that consider lanes on both sides. In the lane-based traffic data set, cross-section volume varies from 4 to 43 vehicles per 30 s, and vehicles’ average speed ranges from 2.72 m/s to 45.87 m/s. Coefficient variation of speed keeps as small scale, and the proportion of truck ranges from 0 to 0.93. Additionally, the data were collected on weekdays, while weekends were not included.

5. Results and discussions

5.1 Conflict risk independent modelling results

For the independent modelling process, the results are presented in Table 4. Significant variables are included in the final models. The relationship between influential factors and the two outcome variables can be quantified by independent ZTP and OL modelling, respectively. Based on the estimated coefficient, six variables were found to be significant in the ZTP model and five variables were significant in the OL model. Among them, Diff_StdS and Diff_CvS were both statistically significant in the two models, but the corresponding sign of coefficient was reversed. If the difference of speed’s standard deviation between subject lane and the adjacent lane increases, the conflict frequency on target lane tends to be higher, but the probability of high-severity conflict risk will decrease. On the contrary, the effect of Diff_CvS on conflict frequency was found to be negative while positive on severity. In the ZTP model, the coefficients of Volume and Cv_speed were positive, while the Prop_truck and Diff_AvgS were negative. For modelling conflict severity, the effects of Std_speed and Diff_CoU were positive, and lane type (Middle_1) was also statistically significant with a positive coefficient, while Avg_speed was negative. Variables including Day of week and Time of day were not significant, thus they were not in the final models.

In this study, the independent modelling process focused on quantifying two relationships. By knowing the respective effects
Table 3. Description statistics of variables.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Statistics of continuous variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>Outcome1</td>
<td>Number of conflicts, response variable</td>
<td>1</td>
</tr>
<tr>
<td>Volume</td>
<td>Number of vehicles detected at the midsection line (veh/30 s)</td>
<td>4</td>
</tr>
<tr>
<td>Avg_speed</td>
<td>Average speed of vehicles (m/s)</td>
<td>2.72</td>
</tr>
<tr>
<td>Std_speed</td>
<td>Standard deviation of vehicles’ speed (m/s)</td>
<td>0.52</td>
</tr>
<tr>
<td>Cv_speed</td>
<td>Coefficient variation of vehicles’ speed</td>
<td>0.02</td>
</tr>
<tr>
<td>Prop_truck</td>
<td>The proportion of trucks among all vehicles</td>
<td>0</td>
</tr>
<tr>
<td>Diff_V</td>
<td>The difference of volume between subject lane and the adjacent lane (veh/30 s)</td>
<td>−16</td>
</tr>
<tr>
<td>Diff_AvgS</td>
<td>The difference of Avg_speed between subject lane and the adjacent lane (m/s)</td>
<td>−19.09</td>
</tr>
<tr>
<td>Diff_StdS</td>
<td>The difference of Std_speed between subject lane and the adjacent lane (m/s)</td>
<td>−7.01</td>
</tr>
<tr>
<td>Diff_CvS</td>
<td>The difference of Cv_speed between subject lane and the adjacent lane</td>
<td>−0.47</td>
</tr>
<tr>
<td>Diff_Ptruck</td>
<td>The difference of Prop_truck between subject lane and the adjacent lane</td>
<td>−0.83</td>
</tr>
</tbody>
</table>

Table 4. Zero truncated Poisson regression and ordered logit regression results.

| Type                          | Variable     | Coef.  | Std. Err. | z value | Pr(>|z|) |
|-------------------------------|--------------|--------|-----------|---------|---------|
| Conflict risk outcome variable 1 (frequency modeling) | Intercept   | −1.257 | 0.064     | −19.49  | <0.001  |
|                               | Volume       | 0.519***| 0.052     | 9.97    | <0.001  |
|                               | Cv_speed     | 0.445***| 0.033     | 13.28   | <0.001  |
|                               | Prop_truck   | −0.193* | 0.079     | −2.43   | 0.015   |
|                               | Diff_AvgS    | −0.486***| 0.074    | −6.56   | <0.001  |
|                               | Diff_StdS    | 0.202**  | 0.066     | 3.07    | 0.002   |
|                               | Diff_CvS     | −0.174***| 0.045    | −3.85   | <0.001  |
| Conflict risk outcome variable 2 (severity modeling) | Avg_speed    | −0.285***| 0.059    | −4.85   | <0.001  |
|                               | Std_speed    | 0.155*   | 0.064     | 2.41    | 0.016   |
|                               | Diff_AvgS    | −0.267** | 0.076     | −3.44   | 0.001   |
|                               | Diff_CvS     | 0.278**  | 0.083     | 3.34    | 0.001   |
|                               | Lanetype_12  | 0.364*   | 0.160     | 2.28    | 0.023   |
|                               | Threshold1   | 0.494    | 0.147     |         |         |
|                               | Threshold2   | 2.719    | 0.165     |         |         |

Note: *, **, *** refer to p-value <0.05, <0.01, <0.001 significance level respectively.

of traffic characteristics on the two outcomes of conflict risk, we discovered that conflict frequency and severity are influenced by different factors, especially with some reverse effects. The results imply that the increase of traffic volume or coefficient variation of vehicles’ speed will increase the frequency, while not significant for severity. If the standard deviation of vehicles’ speed is higher, the probability of higher severity will increase. The difference of speed features between subject lane and adjacent lane were identified as significant predictors. Higher between-lane speed variation may be caused by the trucks’ involvement on the road, which tend to be slower than the rest of the traffic [35].

5.2 Copula-based joint modelling results
Considering the possible interdependence that may exist between conflict frequency and different severity levels, a copula method was then employed to explore that relation. Different families of copula structures were investigated, including Gaussian, Clayton, Frank, AMH and FGM. As aforementioned, the conflict severity outcome variable is split into three binary variables, then a total of 18 models were tested. Table 5 shows all the models’ performance. The goodness-of-fit measures of all the models differ slightly. The model structure with the lowest AIC and BIC was selected as the most suitable model. Therefore, the Gaussian
copula structure outperforms the other structures for joint modelling count data with level 1 and level 3, while Frank copula structure is the best for level 2.

These results indicate that the correlation between conflict frequency and severity varies among different levels. For level 1 and level 3, the dependence of two response variables reflects the Gaussian copula structure, while for level 2, the dependence is Frank copula’s characteristic. Actually, Gaussian and Frank copulas are referred to as ‘comprehensive copulas’ in terms of their ability to parameterize the full range of stochastic dependency by allowing positive and negative dependence with symmetry in both tails. Nonetheless, as compared to Gaussian copula, the Frank copula is characterized by stronger dependence in the middle of distribution and weaker dependence in distribution tails. Such correlations may occur due to some omitted or unobserved variables, measurement errors, or from simultaneous causality [36].

Table 6 shows the coefficient estimation results of the optimal copula models. The Kendall’s τ values can examine the level of dependence, which are also reported in Table 6. Results reveal that the dependency between the level 1 and count data is negative, and the dependency is positive in another two models. The variables, which are significant in the independent models, also show statistical significance in the copula models. It is clear that the impact of explanatory variables on the frequency follows the same trend in the independent ZTP and three copula models. However, the magnitude of those estimated coefficients and the effects of variables on severity levels differ. When focusing on whether it is low severity (level 1), the Diff.StdS and Diff.CvS have similar effects on the two outcome variables. If Diff.StdS increases, both the frequency and the probability of low-severity risk will increase. Meanwhile, both frequency and probability of low-severity risk tend to be lower with higher Diff.CvS.

The significant variables are less in the high-level model than lower one. Diff.CvS is not a significant variable for severity in the second model. Compared to the same significant variables in the first model, the sign of corresponding coefficients is reversed. Besides, the Lanetype_22 is found to be statistically significant in the second model. In the third joint model, there are just two variables found to influence the probability of high-severity risk significantly. Only the Diff.CvS influences both outcome variables simultaneously. The increase of Diff.CvS will cause a higher probability of high-severity risk, but decrease the frequency. Consistent with previous studies, the related variables of speed influence traffic safety greatly [37, 38]. In summary, the influential mechanism of conflict risk varies under each severity level. The proposed joint model provides a more reasonable means of conflict risk modelling.

### 6. Conclusions and recommendations

This paper focused on modelling conflict risk, including quantifying the effects of traffic characteristics on conflict and revealing the dependency between conflict frequency and severity levels. Trajectory data of a HighD dataset from Germany were utilized to collect lane-based data, and 30-second time interval was considered. In other words, this paper aimed at estimating conflict risk on a target lane within specific spatio-temporal unit. Time-to-collision (TTC) index was used to detect conflicts, and a severity index (SI) was proposed on the basis of time-integrated TTC (TIT) indicator. The severity of regional conflict risk could be defined with the SI, then three severity levels were classified by the k-means clustering method. Moreover, conflict severity was regarded as an ordinal variable, while each level was separately set as a binary variable in the joint modelling process. Subsequently, a zero

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Level 1 &amp; frequency (Gaussian copula)</th>
<th>Level 2 &amp; frequency (Frank copula)</th>
<th>Level 3 &amp; frequency (Gaussian copula)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary logit</td>
<td>ZTP</td>
<td>Binary logit</td>
</tr>
<tr>
<td></td>
<td>Coef. (Std. Err.)</td>
<td></td>
<td>Coef. (Std. Err.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.493 (0.147)</td>
<td>-2.148 (0.064)</td>
<td>-1.023 (0.157)</td>
</tr>
<tr>
<td>Volume</td>
<td>-0.137* (0.067)</td>
<td>0.508*** (0.051)</td>
<td>0.118* (0.056)</td>
</tr>
<tr>
<td>Avg_speed</td>
<td>0.258*** (0.060)</td>
<td>-0.131* (0.056)</td>
<td>0.200** (0.066)</td>
</tr>
<tr>
<td>Std_speed</td>
<td>-0.457*** (0.073)</td>
<td>-0.187* (0.079)</td>
<td>-0.465** (0.073)</td>
</tr>
<tr>
<td>Cy_speed</td>
<td>0.436*** (0.033)</td>
<td>0.118* (0.056)</td>
<td>0.443*** (0.033)</td>
</tr>
<tr>
<td>Prop.truck</td>
<td>-0.189* (0.079)</td>
<td>0.200** (0.066)</td>
<td>-0.131* (0.063)</td>
</tr>
<tr>
<td>Diff_AvgS</td>
<td>-0.091** (0.06)</td>
<td>0.322*** (0.078)</td>
<td>0.322*** (0.078)</td>
</tr>
<tr>
<td>Diff_CvS</td>
<td>-0.165** (0.045)</td>
<td>0.165** (0.045)</td>
<td>0.165** (0.045)</td>
</tr>
<tr>
<td>Lanetype_22</td>
<td>-0.354* (0.161)</td>
<td>0.518** (0.171)</td>
<td>-0.179*** (0.045)</td>
</tr>
<tr>
<td>Kendall’s τ</td>
<td>-0.234 (-0.290, -0.166)</td>
<td>0.152 (0.096, 0.215)</td>
<td>0.168 (0.098, 0.241)</td>
</tr>
</tbody>
</table>
A truncated Poisson (ZTP) model was established to quantify the relationship between traffic characteristics and conflict frequency and an ordered logit (OL) model was used to estimate effects of the same factors on conflict severity. Moreover, the copula-based joint modelling method is applied to explore the potential non-linear dependency of conflict risk outcomes, and different copula structures were tested. The joint modelling process was mainly to reveal the different correlations between conflict frequency and different severity levels, along with the different impacts of explanatory variables.

The main findings can be summarized as follows:

(1) The severity of regional conflict risk can be identified and classified by using the proposed severity index (SI). The effects of traffic characteristics on conflict severity can be well estimated by the ordered logit model. The standard deviation of vehicles’ speed influences the severity positively on the subject lane, while the difference of between-lane speed’s standard deviation affects it negatively.

(2) The truncated Poisson model can well quantify the relationship between traffic characteristics and conflict frequency. Traffic volume owns the largest positive effect on conflict frequency. The speed’s coefficient variation and the difference of between-lane speed’s standard deviation also influence the frequency positively.

(3) Correlation varies under different severity levels. For levels 1 and 3, Gaussian copula reveals the dependency with frequency, while for level 2, the optimal joint model is Frank copula. The impact of explanatory variables on the frequency follows the same trend in the independent ZTP and three copula models. Different effects of variables on each severity level can be revealed in the joint models.

Since the relationship may differ under different severity levels, it is essential to know the dependency effects across conflict count components by each severity level. The copula-based method is demonstrated to be an effective way of reflecting the dependency among conflict counts and three severity levels. In this study, the independent model is mainly for understanding the effects of traffic characteristics on two conflict outcomes, respectively. Then the potential dependency among each case is considered in the joint model. This is a first attempt to estimate regional conflict risk considering the different traffic interactions among low-, medium- and high-severity levels. For practical applications, the copula-based approach can provide a new perspective of real-time traffic safety evaluation. With the calibrated joint model, the conflict risk can be assessed by short-term traffic characteristics that collected from detectors. For different severity levels, the traffic factors significantly affecting conflict risk can be different. Then, targeted countermeasures can be implemented by controlling the key factors’ effects at each level. The case that with larger amount of conflict counts under high severity level is undesired. Thus, developing a related proactive safety strategy to reduce the risk is urgent.

Limitations of the present study offer some directions for the future research. First, the study area was limited to freeways during clear and windless weather in Germany, and the length of each spatial slice was limited to 420 m due to the limitation of the HighD data set. Besides, the issue was simplified in this study that each severity level was set as three binary variables, and estimating three models may not be enough to exactly explore the interdependence between conflict frequency and severity. Finally, it would be valuable to consider a more feasible indicator to measure the severity of regional risk, and validate the transferability and temporal stability of the current findings.

Supplementary data

Supplementary data is available at Transportation Safety and Environment online.

Acknowledgement

This research was sponsored by the National Natural Science Foundation of China (Grant Nos. 71901223, 71971222); the Natural Science Foundation of Hunan Province (Grant No. 2021JJ40746), and the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 1053320214771). The authors are also grateful for the assistance of Dr. Pengpeng Xu.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


Appendix A.

As the clustering object, SI is a one-dimensional numerical value. To classify the severity of conflict risk, k-means method is employed based on SPSS software. Fig. A1(a) displays the sum of the squared errors (SSE) varying among different number of clusters. When the number of clusters is less than 3, SSE declines greatly, and the change of SSE slows down when the number is above 3. Besides, because the silhouette coefficient is a key index to describe the difference inside and outside the cluster, its change curve is presented in Fig. A1(b). A larger value of silhouette coefficient represents a better clustering effect. Although the coefficient reaches the highest point when the number of clusters is 2, it changes rapidly later, thus two clusters may not be appropriate. The overall results show that the optimum number of clusters is three. Therefore, three severity levels of conflict risk are further analysed in this study.

Fig. A1. Clustering effect of k-means method: (a) change of SSE with different clusters; (b) Silhouetter coefficient diagram.