Stochastic method based on copulas for predicting severe road traffic interactions

Chen, Zhankun; Yastremska-Kravchenko, Oksana; Laureshyn, Aliaksei; Johnsson, Carl; D'Agostino, Carmelo

Published in:
Analytic Methods in Accident Research

DOI:
10.1016/j.amar.2024.100347

2024

Document Version:
Publisher's PDF, also known as Version of record

Analytic Methods in Accident Research 44 (2024) 100347

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

journal homepage: www.elsevier.com/locate/amar

Stochastic method based on copulas for predicting severe road traffic interactions

Zhankun Chen, Oksana Yastremska-Kravchenko, Aliaksei Laureshyn, Carl Johnsson, Carmelo D'Agostino [*]

Department of Technology & Society, Lund University, Lund 221 00, Sweden

A R T I C L E I N F O

Keywords:
Traffic conflict measures
Copula
Extreme value theory
Severe interaction

A B S T R A C T

A major difficulty in assessing road traffic safety is the scarcity of historical accident data. xxThis is a common problem in contexts where a certain level of safety has been reached or where exposure is low, such as mixed traffic conditions with different levels of transport automation. Recent studies have demonstrated how severe interactions between road users and/or road users and infrastructure can be a direct measure of safety. However, limiting the investigation to only the most extreme events may lead to inconclusive results considering the lack of prediction robustness and the possible selection bias. In this context, extreme value theory (EVT) is commonly used to extrapolate crashes from road traffic interactions, even combining several indicators. The present work extends the EVT paradigm by proposing a method based on copula functions and EVT, which enables a more specific and continuous evaluation of interaction severity. Compared with pure EVT, this new approach extends the boundary to interactions of all severities while implicitly assuming that the relationship between safety-relevant events and road casualties is stochastic. This EVT-copula approach was also compared with bivariate peaks over threshold (BPOT). It was found that the two approaches yield similar prediction results for crash probabilities. Furthermore, the proposed approach applies to events not properly defined in BPOT and provides more accurate predictions for severe (and less severe) interactions compared with BPOT, when benchmarked against observations.

1. Introduction & background

Traffic conflict measures have emerged as an alternative to safety analysis based on historical accident records, either by exploiting the statistical association between conflicts and crashes (Wu and Jovanis, 2012) or using conflicts as crash precursors. The latter has the advantage of elucidating the development of a crash at a microscopic level and no accident data is required for model development (Arun et al., 2021b).

The reasons why traffic conflicts have become so popular are mainly related to known (by now well-described) issues associated with direct safety assessment, including under-reporting, low counts, over-dispersion (and statistical issues related to that), long periods for data collection, and lack of descriptive details available in a standard crash report (Lord and Mannering, 2010). Notably, traffic conflicts provides quick feedback where the data collection time is relatively short and can be applied to without aggregation of

* Corresponding author.
E-mail addresses: zhankun.chen@tft.lth.se (Z. Chen), oksana.yastremska-kravchenko@tft.lth.se (O. Yastremska-Kravchenko), aliaksei.laureshyn@tft.lth.se (A. Laureshyn), carl.johnsson@tft.lth.se (C. Johnsson), carmelo.dagostino@tft.lth.se (C. D'Agostino).

https://doi.org/10.1016/j.amar.2024.100347
Received 20 March 2024; Received in revised form 11 July 2024; Accepted 13 July 2024
Available online 17 July 2024

Table 1
Theoretical comparison between the bivariate POT approach and the proposed copula approach.

Method	Sample		Parametrization			Probability of interaction
Copula approach	Event	Interactions with limited proximity	Marginal distributions	Proximity	$1-\left(1+\gamma\frac{\widetilde{x}-T(u)}{\sigma}\right)^{-1/\gamma}$	$P(X\leq x, Y>y)=\overline{C}\left(1-F_{T(X)}(\widetilde{x}), 1-F_2(y)\right)\cdot P(\Omega_0)$ y can take any value greater than 0, hence covering all the severity.
				Severity	Depends on the conditional distribution	
	Probability space on which the copula is defined	$(\Omega_0\times\Omega, \mathscr{F}_0\otimes\mathscr{F}, Q)$	Dependence structure	$C(u,v) = exp\left(-\left((-\ln(u_1))^{\theta}+(-\ln(u_2))^{\theta}\right)^{1/\theta}\right)$		
Bivariate POT approach	Event	Interactions with limited proximity and severity	Marginal distributions	Proximity	$1-P(X\leq u)\cdot\left(1+\gamma\frac{\widetilde{x}-T(u)}{\sigma}\right)^{-1/\gamma}$	$P(X\leq x, Y>y)=\overline{C}\left(1-F_{T(X)}(\widetilde{x}), 1-F_2(y)\right)$ y can only take value greater than a severity threshold v, hence covering only the severe interaction.
				Severity	$1-P(Y>v)\cdot\left(1+\gamma\frac{y-v}{\sigma}\right)^{-1/\gamma}$	
	Probability space on which the copula is defined	$(\Omega\times\Omega, \mathscr{F}\otimes\mathscr{F}, P)$	Dependence structure	$C(u,v) = exp\left(-\left((-\ln(u_1))^{\theta}+(-\ln(u_2))^{\theta}\right)^{1/\theta}\right)$		

multiple sites. Furthermore, traffic conflicts reveal how crashes develop, and more importantly, they support a proactive response.

Two fundamental assumptions are made when modeling traffic conflicts as crash precursors: the continuity assumption and the hierarchy assumption (Glauz and Migletz, 1980). The continuity assumption states that traffic conflicts and crashes are governed by the same source of randomness, whereas the hierarchy assumption states that the conflicts can be measured numerically, which enables comparisons among different levels of conflict severity. Among the available statistical approaches for exploring the conflict-crash relationship, extreme value theory (EVT)-based models have received significant attention owing to their simplicity and flexibility in implementation (Zheng et al., 2021; Ali et al., 2023a).

The advantages of addressing the conflict-crash relation with EVT are numerous. First, crashes are generally defined in a straightforward way, stating that the proximity between two road users becomes zero, which is measured using one (Songchitruksa and Tarko, 2006; Zheng et al., 2014; Farah and Azevedo, 2017; Åsljung et al., 2017; Borsos et al., 2020; Ali et al., 2023b) or more crash proximal indicators (Jonasson and Rootzén, 2014; Zheng et al., 2018; Wang et al., 2019; Fu and Sayed, 2021). Second, interaction severity can be modeled alongside crash proximity to capture the complex dynamics of an interaction, which can be further used to evaluate the risk of injury-related accidents (Arun et al., 2021a; Borsos, 2021; Arun et al., 2022; Howlader et al., 2024; Tahir and Haque, 2024). Moreover, EVT can be extended to include multisite heterogeneities, such as behavioral and geometric factors (Zheng and Sayed, 2019a; Cavadas et al., 2020; Fu et al., 2020; Ali et al., 2022), which help to predict crash risk in real-time. It is also worth mentioning that including more indicators using multivariate EVT is conceptually beneficial because more information is aggregated in the model, and crash estimates are more aligned with historical accident records (Ali et al., 2023a).

However, previous reviews and discussions (Wang et al., 2021; Zheng et al., 2021; Ali et al., 2023a) have also pointed out several drawbacks concerning the recent advancement in multivariate EVT methodology. The main issues are related to the heavy data requirements and rigid dichotomous classification of events. As a result, cases exceeding the thresholds are regarded as conflicts, but the threshold selection is purely mathematical. In multivariate peaks over threshold (POT), these two problems make it difficult to maintain the bias-variance balance because less data is used to estimate more parameters. Consequently, lower threshold values are typically selected to ensure estimator convergence, but lowering the threshold too much runs the risk of violating the generalized Pareto (GP) distribution assumption in the marginal distributions, which introduces further bias in the resulting probability (see Table 1).

The basic assumption of the present work is that extreme values are not required for all the indicators in safety analysis, for example, in assessing the crash severity. This assumption is based on the definition of crash severity defined using the human-body tolerance of the energy released, which is specific to individual road users (Augenstein et al., 2003; Bahouth et al., 2014). Therefore, in terms of the energy involved, even non-extreme crashes can result in severe injuries. For example, when vulnerable road users are involved, lower energy does not necessarily mean lower severity, and specific interaction angles, even at low speed, can lead to severe consequences. However, safety relevant (i.e., have enough energy to potentially cause a severe crash) events are not necessarily identified as extremes by the BPOT. Expanding the range of severity levels has the clear advantage of providing a more robust crash frequency prediction.

For the reasons mentioned above, this study proposes an efficient method to simultaneously model crash proximity and interaction severity, while the range of interaction severity is extended to include more than just extreme values. The proposed method combines a copula function with the extreme-value-distributed margin of crash proximity. This research differs from previous studies that have applied copulas to traffic conflict measures (Fu and Sayed, 2021; Arun et al., 2022) because the severity dimension (described by interaction severity traffic conflict measures) is not treated as an "extreme value" herein. Meanwhile, the proximal dimension still preserves the extreme value-type distribution. Consequently, the full range of severities that occur during critical situations is obtained. This is advantageous because more data is used more efficiently than in bivariate extreme value models, and the whole range of severity is included, which overcomes the limitation of only selecting extreme events in multivariate EVT. As a result, it is possible to model crashes at any observed (or unobserved but possible) severity. The proposed method based on the copula and EVT has been applied to a dataset comprising interactions involving left turn maneuvers at a signalized intersection with a permitted left turn. The results have also been compared with a state-of-the-art bivariate threshold excess model using the same dataset.

2. Theory

The theory behind the analysis carried out in the paper is based on the following technical tools, which are described in detail later:

1) The mathematical definition of interaction via traffic conflict measures is the basis for the construction of any probabilistic model.
2) A dependence structure that associates two or more random variables allows the simultaneous modeling of several traffic conflict measures.
3) A partition of probability space based on the extreme selection method is used to formulate a new probability space restricted to only the near interaction.

2.1. Definitions of interaction

Random elements in an interaction

Traffic conflict measures (or indicators) provide different aspects for the quantitative assessments of a road traffic interaction, which cannot be measured directly. Let $(\Omega, \mathscr{F}, P)$ be the probability space in which Ω represents the space of all interactions (outcome space), and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is the measurable space on $\mathbb{R}$ (state space). Then, a traffic conflict measure $X : \Omega \mapsto I \subset \mathbb{R}$ is a random variable that maps an arbitrary interaction to a concrete interval on the real line.

An arbitrary subset $\bar{I} \in \mathscr{B}(\mathbb{R})$ corresponds to an interval of indicator values. The measurability of X yields $X^{-1}(\bar{I}) \in \mathscr{F}, \forall \bar{I} \in \mathscr{B}(\mathbb{R})$, meaning that the subset of interactions that takes indicator values in $\bar{I}$ are denoted as $X^{-1}(\bar{I})$. Applying probability measures on $X^{-1}(\bar{I})$ gives the probability of those events.

$$\boldsymbol{P}(\{\boldsymbol{\omega} \in \Omega : \boldsymbol{X}(\boldsymbol{\omega}) \in \bar{\boldsymbol{I}}\}) = \boldsymbol{P}\left(\boldsymbol{X}^{-1}(\bar{\boldsymbol{I}})\right) \tag{1}$$

When more than one traffic conflict measures are involved in the definition of interaction, the probability space is expanded accordingly. Suppose p indicators $X_1, \cdots, X_p$ are used, then the previous construction on a single indicator can be extended to the random vector $(X_1, \cdots, X_p)$.

$$\left(X_1, \cdots, X_p\right) : (\Omega^p, \mathscr{F}, P) \mapsto (\boldsymbol{I} \subseteq \mathbb{R}^p, \mathscr{B}(\mathbb{R}^p)), \mathscr{F} = \mathscr{F}\underbrace{\otimes \cdots \otimes}_{p\text{times}}\mathscr{F}$$

Mathematically, applying different conflict indicators to the same interaction is observed as mappings from the same outcome space to different intervals on the real line. For an arbitrary subset of interaction $\overline{\Omega} := \overline{\Omega_1} \times \cdots \times \overline{\Omega_p}$, suppose $X_1(\overline{\Omega_1}) = \overline{I_1}, \cdots, X_p(\overline{\Omega_p}) = \overline{I_p}$, then $\overline{\Omega}$ can be defined as the intersection of the pre-images of the indicators $\overline{\Omega} = \cap_{i=1}^{n} X_i^{-1}(\overline{I_i})$. The measurability of $(X_1, \cdots, X_p)$ states that for each $\bar{\boldsymbol{I}} := \prod_{i=1}^{p} \bar{I_i} \subseteq \mathbb{R}^p$, $\cap_{i=1}^{p} X_i^{-1}(I_i) \in \mathscr{F}$. Hence, interactions measured using multiple indicators can be represented by $\cap_{i=1}^{p} X_i^{-1}(I_i)$. Applying probability measures to such events results in the following:

$$\boldsymbol{P}(\overline{\Omega}) = \boldsymbol{P}\left(\cap_{i=1}^{n} \boldsymbol{X}_i^{-1}(\overline{\boldsymbol{I}_i})\right) = \boldsymbol{P}(\boldsymbol{X}_1 \in \overline{\boldsymbol{I}_1}, \cdots, \boldsymbol{X}_n \in \overline{\boldsymbol{I}_n}) \tag{2}$$

To further generalize the definition of interaction severity, suppose X_1 is a crash proximal indicator and X_2 is an interaction severity indicator, then each interaction is equipped with a coordinate whose first entry and second entry are the values of X_1 and X_2, respectively. An interaction is then defined by the random vector (X_1, X_2) in probability space $(\Omega \times \Omega, \mathscr{F} \otimes \mathscr{F}, P)$.

2.2. The copula and multivariate extreme value distribution (MEVD)

A copula is a joint distribution function with uniformly distributed margins. It is a convenient tool for modeling the multivariate dependence structure because it accommodates random variables that are characterized by different parametric families of distribution functions (Genest and Favre, 2007).

The distributional equivalence between a copula and any arbitrary joint probability distribution was established by Sklar (1959). Sklar's result can be described as follows:

if $(X_1, \cdots, X_d) \sim F$ be random vector on probability space $(\Omega, \mathscr{F}, P)$, $F_1, \cdots, F_d$ are marginal distributions of F, then there exists a d-dimensional copula:

$C(F_1(x_1), \cdots, F_d(x_d)) = F(x_1, \cdots, x_d) \forall (x_1, \cdots, x_d)$ in the continuity of F.

Conversely, if C is a copula and there is a function $F(x_1, \cdots, x_d) = C(F_1(x_1), \cdots, F_d(x_d))$, then F is a joint distribution function whose margins are $F_1, \cdots, F_d$.

Similarly, there exists a copula distribution that equates in distribution to a survival function, such that $\overline{C}(1 - F_1(x), 1 - F_2(y)) = P(X > x, Y > y)$. $\overline{C}$ is called the survival copula. The relation between survival copula and the copula is (Nelsen, 2006):

$$C(u, v) = \overline{C}(1 - u, 1 - v) + u + v - 1 \tag{3}$$

Multivariate extreme value distributions (MEVD) with uniform margins are also considered extreme value copulas. It is important to note that extreme value copulas refer to marginal distributions being modeled by the extreme value-type distribution rather than the inherent dependence structures of the copulas. The dependence structure inherent in extreme value copulas can also be used beyond extreme values, as stated by Gudendorf and Segers (2010), they "*…not only arise naturally from the extreme values, it is also a convenient tool to model data with positive dependence.*"

A bivariate extreme value distribution with uniform margins (i.e., copula) takes the following form:

$$\boldsymbol{C}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{exp}\left(\boldsymbol{log}(\boldsymbol{uv}) \cdot \boldsymbol{A}\left(\frac{\boldsymbol{log}(\boldsymbol{u})}{\boldsymbol{log}(\boldsymbol{uv})}\right)\right), \boldsymbol{u}, \boldsymbol{v} \in [0, 1] \tag{4}$$

Eq. (4) is used to approximate the samples selected through bivariate component-wise block maxima.

2.3. Multivariate threshold excess model

The conditional upper-tile distribution is a standard result from Pickands (1975), such that, if $X_1, \cdots, X_n$ are random samples from a distribution function F,

$$\lim_{u \to x^F} \boldsymbol{P}(\boldsymbol{X} > \boldsymbol{x} | \boldsymbol{X} > \boldsymbol{u}) \to 1 - \boldsymbol{H}(\boldsymbol{x}) = \left(1 + \gamma \frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma}\right)^{-1/\gamma}, x^F = \inf\{x \in \mathbb{R} : F(x) = 1\} \tag{5}$$

$H(x)$ is a generalized pareto (GP) distribution, where u is the pre-specified threshold, and σ, γ are the parameters to be estimated. As an immediate consequence of Eq. (5), the unconditional upper-tail distribution $\boldsymbol{P}(X \leq x)$, when $X > u$, coincides with:

$$\widetilde{\boldsymbol{H}}(\boldsymbol{x}) = 1 - \boldsymbol{P}(\boldsymbol{X} \leq \boldsymbol{u}) \left(1 + \gamma \frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma}\right)^{-1/\gamma} \tag{6}$$

The principles of excess models based on multivariate thresholds were initially explained by section 3 of Smith (1994). First, the component-wise exceedances are transformed to the desired marginal distributions using Eq. (6). Second, the method of censored likelihood is used to determine the sample contribution to the estimation. It was stated that the joint distribution in the upper-tail region coincides with the multivariate extreme value distribution, such that:

$$P(X \leq x, y \leq y) \approx C(u, v) \text{ if } x > u_1, y > u_2 \tag{7}$$

where $C(u, v)$ is given by Eq. (4), $u = \widetilde{H}(x), v = \widetilde{H}(y)$ are transformed to uniform margins using Eq. (6), and u_1, u_2 are the thresholds.

In the context of modeling traffic conflict, the proximal indicator is first transformed by a decreasing transformation T before applying the POT method. Let $\widetilde{X} := T(X)$, then the lower tail of the crash proximal indicator is expressed as the upper tail after the transformation.

$$P(X \leq x | X < u) = P(T(X) \geq T(x) | T(X) > T(u)) \sim \overline{H}(T(x)) = \left(1 + \gamma \frac{\widetilde{x} - T(u)}{\sigma}\right)^{-1/\gamma} \tag{8}$$

2.4. Motivation of the new approach

The following approach relies on the hierarchical assumption of traffic conflicts; that is, the severe interactions (including crashes) are the subset of the near interactions. Near interactions are separated from far interactions by a threshold. Following the definition in Section 2.1, let u_1 be the separating threshold.

Near interaction:.$\Omega_0 = X^{-1}((\infty, u])$

Far interactions:.$\Omega \setminus \Omega_0$

The probability of a near interaction can be rewritten as:

$$P(X \leq x, Y > y) = P(X \leq x, Y > y | \Omega_0) \cdot P(\Omega_0) + P(X \leq x, Y > y | \Omega_0) \cdot P(\Omega \setminus \Omega_0)$$

$P(X \leq x, Y > y | \Omega_0) \cdot P(\Omega \setminus \Omega_0)$ is equal to 0.

2.5. Probability of near interactions/crashes of different severity

We restrict the probability space to only the near interaction $(\Omega_0 \times \Omega, \mathscr{F}_0 \otimes \mathscr{F}, Q)$, where $Q(A) = P(A|\Omega_0)$ is the conditional probability. Let F_1 be the distribution of crash proximal indicator, F_2 be the distribution of interaction severity indicator, and C be the joint distribution of two indicators (all with respect to probability measure Q). Then, the probability of an interaction with specified severity y can be written as:

$$P(X \leq x, Y > y) = P\left(\widetilde{X} > \widetilde{x}, Y > y\right) = Q\left(\widetilde{X} > \widetilde{x}, Y > y\right) \cdot P(\Omega_0)$$

For $Q\left(\widetilde{X} > \widetilde{x}, Y > y\right)$, the final formula for evaluating interaction at the severity associated with y is then given by Eq. (3):

$$P(X \leq x, Y > y) = \overline{C}(1 - F_{T(X)}(\widetilde{x}), 1 - F_2(y)) \cdot P(\Omega_0) \tag{9}$$

The formal construction of measure Q is provided in Appendix A. We also note that there is a possibility for inherent negative rank correlation in the data. Some copulas (e.g., the Gumbel copula) do not support data with a negative rank correlation. A remedy for this is also provided in Appendix A.

3. Method

The following sub-sections summarize all the study details: dataset *(3.1)* – provides a description of the data this study is based on, objective measures *(3.2)* – presents explanations regarding the choice of traffic conflict measures, analysis *(3.3)* – describes the implementation procedure of the novel approach used in this study.

3.1. Data-processing

This study conducts an extended analysis using a dataset from previous research (Laureshyn et al., 2017). The data includes 89 h of video recordings from 2010, recorded between 07:00 and 19:00 with a framerate of 21 frames per second and a resolution of 640 × 480 pixels. The present study focuses only on interactions between cars turning left and cars driving straight through the intersection (Fig. 1).

This study used the object detector YOLO8 (Jocher et al., 2023), along with the object tracking algorithm strong-SORT (Du et al., 2023). These provide the trajectories of classified objects (cars, trucks, buses, etc.) traveling through the camera field of view, with coordinates measured in meters (Fig. 2).

Finally, the trajectories have been smoothed using a Savitzky–Golay filter in MATLAB. This step removes any suspiciously short trajectories that appear in the case of a vehicle being detected multiple times, as well as trajectories that appeared outside the roads because of detection errors. Altogether, 25,126 encounters were recorded over two weeks from the real-life traffic flow. Based on these data, various objective parameters were calculated and extracted for further analysis in this study. Only the interactions that showed the collision course were kept for further data analysis. The selection process involves two steps. First, a boxed area is drawn around the intersection point of the two trajectories. Second, for each vehicle, a time interval that describes how long the target vehicle remains in the boxed area is computed based on motion prediction in every common frame. We say that the collision course exists if the time intervals overlap more than once during the common frames. Overall, 4772 interactions have been selected for analysis.

3.2. Essentials for defining a safety-relevant event (interaction) for EVT application

There are different definitions of conflicts in the classical framework of traffic conflict (Older and Spicer, 1976; Amundsen and Hyden, 1977; Allen et al., 1978; Hydén, 1987; Saunier and Laureshyn, 2021), and they are not consistent (Chin and Quek, 1997; Yastremska-Kravchenko et al., 2022). The validity of various traffic conflict measures can differ depending on the types of interactions (Johnsson et al., 2021a,b). The definitions can be grouped into four types (Arun et al., 2021b), namely proximity-based (Hayward (1972); Cafiso et al., (2018)), severity-based (Hydén (1987); Laureshyn et al., (2010)), evasion-based (Perkins and Harris (1968); Tarko (2018)), and counterfactual-probability-based (Davis et al., (2011)).

Another issue is that the output from EVT models depends on the selection of the conflicts. This is not trivial because different indicators are used to evaluate the events at different time instances. An example is the use of post encroachment time and Delta-V; the former is measured when the interaction is resolved, and the latter is measured during the culmination phase. Hence, the same event can be identified as a conflict or a normal interaction depending on the set of indicators, which often measure values in a different time or space position (Zheng and Sayed, 2019b). Unlike conventional statistical models where the parameters depend on the overall distribution of a sample, EVT models are governed by outliers. Consequently, if the definition of conflict is flawed, more data may not lead to improved model performance (Ali et al., 2023a).

The main problem that remains unsolved is that some traffic conflict measures are more representative than others and no universal indicator exists that is suitable for all situations (Arun et al., 2021c). Therefore, we consider that the measurement within a conflict should be elementary and intuitive, particularly regarding the spatial proximity and the energy involved, which are preferably measured at the same point in time. The situation-based factors can be pre-specified in the selection of safety-relevant events.

The adopted definition of conflicts is motivated by the framework postulated by Tarko (2021), who divided the encounters into safety-relevant or other events and treated conflicts (in the traditional sense) as subsets of safety-relevant events. Safety-relevant events are conceptually involved in the "pre-conflict" stage, where encounters have the potential for being conflict candidates. Although it is difficult to identify the sufficient conditions for a failure to occur without insight vehicle information, we can formulate

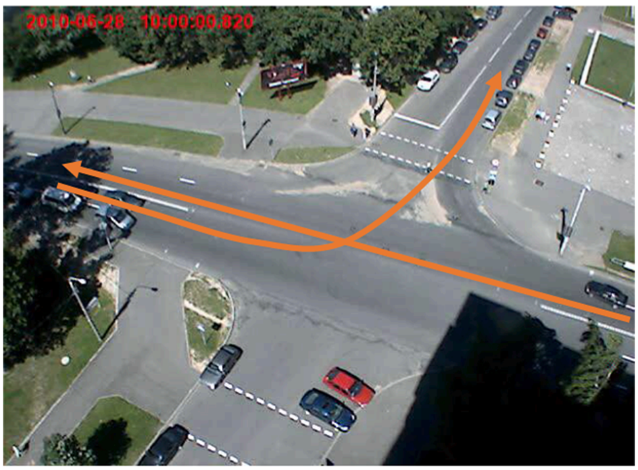

Fig. 1. Camera view of the target interaction.

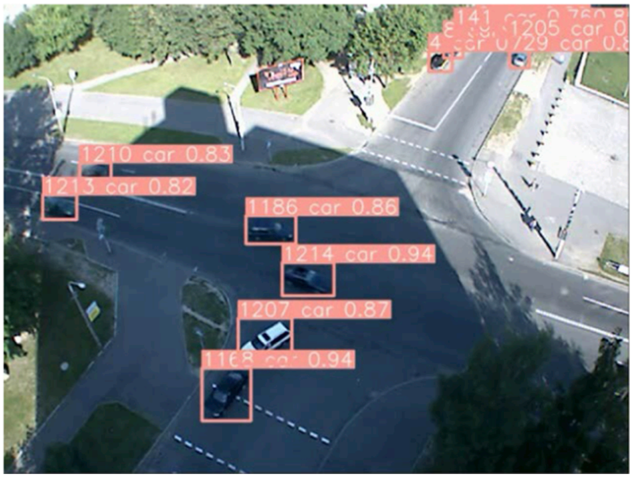

Fig. 2. Camera view with the cars being detected and tracked.

the necessary condition(s) using logic. One simple requirement for safety-relevant events is the presence of a collision course when the proximity between two vehicles is low. Therefore, we define an event to be safety-relevant if it has a non-zero probability of contact (i.e., the crash proximity is low, and the vehicles are on a collision course). The mathematical definition is described in Section 2.2. Among the safety-relevant events, severe interactions[1] are the ones with sufficient energy to cause injuries or fatalities in a potential crash.

3.2.1. *Measure of proximity*

Distance is used to measure the proximity between two vehicles in an interaction. This is chosen to ensure an assumption-free indicator, which is unlike the time to collision or other indicators that require the prediction of trajectories and speed after the evasive action. The identification of evasive action is still not perfect in video analysis and requires more research, which is outside the scope of the present paper. The use of assumption-free indicators is motivated by Fay Patterson (2021) and has been modified to limit the analysis to only events where collision courses have been detected. The measurement is extracted when the distance between two vehicles is minimized during the interaction (i.e., in the time interval from the beginning to the moment when the interaction is resolved). The minimum distance is denoted as MD.

3.2.2. *Measure of severity*

The relationship between speed and critical events in traffic is widely recognized and understood. Speed has often been used as a proxy for consequences, for example, in the Swedish traffic conflict technique (Hydén, 1987; Laureshyn and Varhelyi, 2020). Several studies have also demonstrated the increasing risk of injuries and fatalities associated with increased impact speed (Wramborg, 2005; Rosen et al., 2011; Żuchowski, 2016). However, many factors, including the mass of involved road users or the angle of collision, influence the relationship between speed and possible consequences/energy involved in the crash (Shelby, 2011). Instead of actual speed, many studies have suggested using the expected change in velocity between the hypothetical pre-collision and post-collision trajectories of each road user, represented by the Delta-V (Δv) indicator (Shelby, 2011; Bagdadi, 2013; Laureshyn et al., 2017). Delta-V contains the velocity (vector quantity containing direction) and information regarding both the mass of the road users involved and their approaching angle (Eq. (10).

$$\Delta \boldsymbol{v}_1 = \frac{\boldsymbol{m}_2}{\boldsymbol{m}_1 + \boldsymbol{m}_2} \times \sqrt{\boldsymbol{v}_1^2 + \boldsymbol{v}_2^2 - 2\boldsymbol{v}_1\boldsymbol{v}_2\mathbf{cos}\boldsymbol{\alpha}};\, \Delta \boldsymbol{v}_2 = \frac{\boldsymbol{m}_1}{\boldsymbol{m}_1 + \boldsymbol{m}_2} \times \sqrt{\boldsymbol{v}_1^2 + \boldsymbol{v}_2^2 - 2\boldsymbol{v}_1\boldsymbol{v}_2\mathbf{cos}\boldsymbol{\alpha}} \tag{10}$$

where $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are vehicle masses, $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ are their speeds, and $\boldsymbol{\alpha}$ refers to the approach angle. Herein, the masses of the vehicle are assigned a fixed value depending on whether the vehicle is a passenger car, minivan, or truck.

3.3. *Implementation procedure*

The data analysis was performed using R software (R Development Core Team, 2023) and several R packages. Namely, “ggplot2” (Wickham, 2016) and “plotly” (Sievert, 2020) were used to enhance the aesthetics of the plots, “extRemes” (Eric and Richard, 2016) was used for univariate extreme value analysis, and “copula” (Jun Yan et al., 2020) was used for parameter estimation of the copulas,

[1] In the paper, the terms “conflicts” and “severe interactions” are interchangeable. However, we want to highlight the fact that they may not follow the strict conflict definition (e.g., the Swedish or Dutch traffic conflict techniques).

computing copula probabilities, and simulating copulas.

We use the inference for margins and a semi-parametric estimator based on rank correlation to estimate the copula parameters. Prior to the analysis, the proximal indicator must be transformed decreasingly. The simplest decreasing transformation (i.e., negation) was used in this study. The transformed distance is denoted as $\left(\widetilde{X_1}, \cdots, \widetilde{X_n}\right) := (T_1(X_1), \cdots, T_1(X_n))$. The implementation is outlined in six steps .

1. *Apply the POT model to the margin of proximity*: apply univariate POT to determine the upper-tail distribution of $\widetilde{\mathrm{X}}$. The upper-tail marginal distribution of transformed distance is approximated using Eq. (5).
2. *Transform the marginal distribution into a standard uniform distribution*: first, we restrict the outcome space to Ω_0 by excluding observations whose distances do not exceed the chosen threshold. The corresponding Δv values are kept unchanged. The uniform observations $(u_1, v_1), \cdots, (u_m, v_m)$ are obtained using $(u_i, v_i) = \left(F_1(\widetilde{x_i}), F_2(y_i)\right)$, in which m is the number of observations that exceed the threshold, and F_2 is the distribution function for the interaction severity indicator on Ω_0. We use the empirical distribution $F_n(y) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{Y_i \le y\}$. Accordingly, the uniform observation is equal to $(u_i, v_i) = \left(F_1(\widetilde{x_i}), F_n(y_i)\right)$.
3. *Fit the copula to the transformed data*: the uniform observations $(u_1, v_1), \cdots, (u_m, v_m)$ are used to fit different copula functions. The rank approximate estimator based on Kendall's τ (See Appendix B) is used here.
4. *Selection of the models*: compare the non-parametric and parametric estimates from different models using the goodness of fit.
5. *Compute the probability*: the probability of an interaction is evaluated by plugging in different values of x, y into Eq. (9).
6. *Relative validation*: the proposed approach is validated against BPOT. In the tail region, where no observations exist, we compare the magnitudes of the probabilities obtained from both approaches. In the region where observations exist, we compare the probabilities from the two approaches with the observation.

A flow chart (Fig. 3) describes the entire procedure, from data selection to the analysis and validation with reference to the Method section of the manuscript.

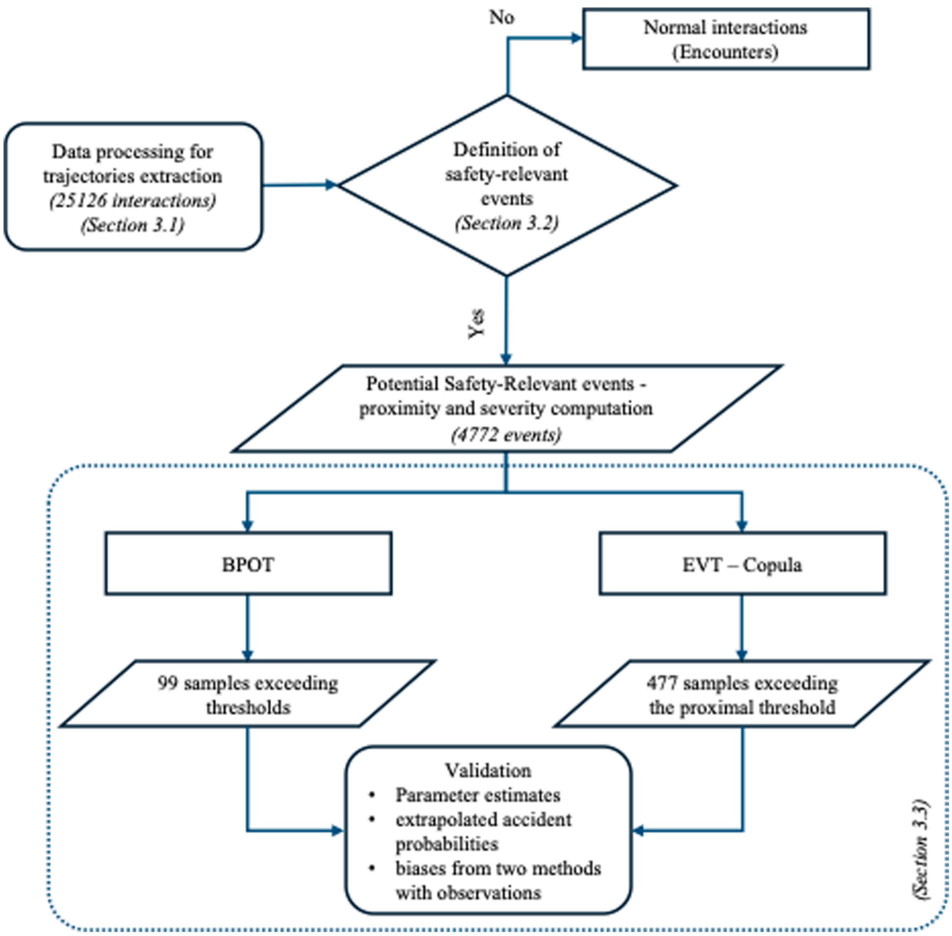

Fig. 3. Flow chart of the implementation procedure.

4. Results

The near interactions (Ω_0) are plotted in Fig. 4.

The threshold that separates near interactions from the rest is set based on the 0.9 empirical percentile of the MD, which is chosen according to the threshold selection criteria from the univariate POT model. The conditional marginal distribution (Eq. (4)) for MD is estimated by maximizing the likelihood as follows:

$$F_{T(X)}(\widetilde{x}) = 1 - \left(1 - 0.493 \cdot \frac{\widetilde{x} + 4.84}{2.60}\right)^{1/0.493}$$

where $\widetilde{x} = -x$ is the transformed distance, $u = -4.84$ is the threshold, $\sigma = 2.60$, and $\gamma = -0.493$.

The Δv values of the red samples in Fig. 2 were transformed into uniform margins using the empirical distribution. Then, three one-parameter copulas (Gumbel, normal, and Clayton) representing different dependence structures were fitted to the uniform data. The copula expressions, as well as their estimation results, are presented in Table 2.

The last two columns of Table 2 display the goodness of fit results based on the Cramer-von Mises statistic (Kojadinovic et al., 2011). The p-values indicate that the normal copula and Clayton copula should be rejected, and the Gumbel copula is the fittest candidate for modeling near interaction in this dataset.

The results of the proposed model are presented in Table 3, as well as the results estimated using BPOT. The parameters σ, γ represent the scale and shape parameters of the GP distributions, and θ is the dependence parameter of the Gumbel copula. In the proposed approach, no GP distribution, nor other parametric distribution, was used in fitting the distribution of Δv. Thus, σ_2, γ_2 were not available. The probability space has a different interpretation in the proposed approach; it represents the dependence between the lower proximity tail and the whole span of interaction severity, but in BPOT, it represents the dependence between the low proximity tail and the high severity tail. In both cases, the dependences were very weak (independence when $\theta = 1$). The last four columns of Table 3 show the estimated probabilities for different interactions, and the observed probabilities are included in the parenthesis (if possible). The thresholds for defining different interaction severity are consulted based on the logistic model for accident severity outcomes reported by Augenstein et al., (2003). When an accident happens, the MAIS3 + injury probability for near-side passengers is 0.67 ($\Delta v = 16m/s$), 0.25 ($\Delta v = 10m/s$), and 0.05 ($\Delta v = 3m/s$).

The proposed approach is less biased than the BPOT when the interaction severity is not an extreme value (Figs. 5 and 6). The bias is measured by determining the distance between the observations and predictions (Fig. 6). Moreover, the last column of Table 3 represents a less severe crash (the threshold is derived from the MAIS3 +); this cannot be predicted by BPOT, but the proposed approach is still applicable. Fig. 5 displays the continuous, conditional survival function of Δv when near interactions (MD of less than 1 m) are selected. The results show that no interactions with Δv less than 8 m/s can be assessed using BPOT. For interactions with $\Delta v > 8m/s$, the BPOT prediction contains larger bias than the proposed approach when Δv is less extreme (Figs. 5 and 6), and the biases become similar when Δv is more extreme. No conclusions about the biases can be drawn in the regions where the observations are not available, but the probabilities of those events were of the same magnitude for the two methods. We conclude that the proposed approach is less biased (produces smaller residuals) than BPOT in the analysis of conflicts with non-extreme severity.

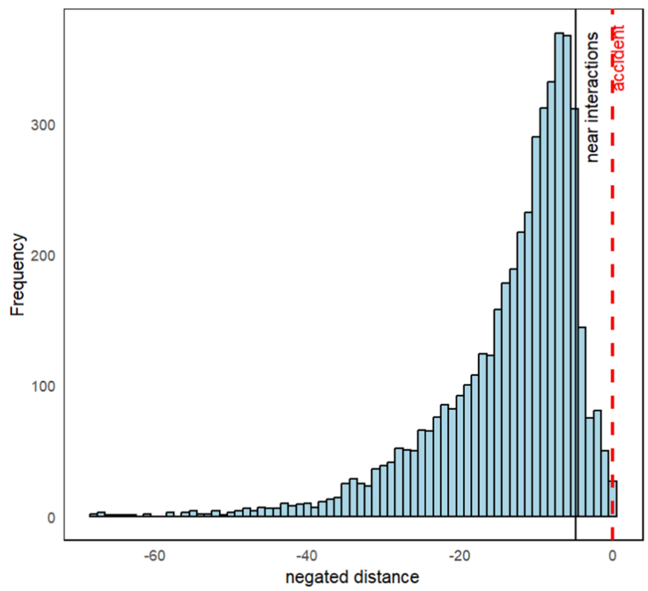

Fig. 4. Histogram of the negated minimum distance in meters (effectively the MD) when the threshold for near interaction is applied.

Table 2
Parameter estimates for different copula models. The estimates in the parenthesis were obtained when the input data was transformed by the empirical distribution.

Copula	Model form	Estimated parameter (empirical estimation)	Test statistics	p-value
Gumbel	$C(u,v) = exp\left(-\left((-\ln(u))^{\theta} + (-\ln(v))^{\theta}\right)^{1/\theta}\right)$	$\theta = 1.169$	0.033	0.248
Normal	$C(u,v) = F(\sigma_1 \cdot \Phi^{-1}(u), \sigma_2 \cdot \Phi^{-1}(v))$*	$\rho = 0.225$	0.047	0.001
Clayton	$C(u,v) = \max(u^{-\theta} + v^{-\theta} - 1, 0)^{-1/\theta}$	$\theta = 0.338$	0.06	0.002

*Φ is the c.d.f. of standard Gaussian, F is a bivariate normal distribution with a covariance matrix whose diagonal entries are σ_1, σ_2 and off-diagonal entries are ρ

5. Discussion

This study contributes to the development of the traffic conflict framework. Specifically, a model is proposed that includes events (whether observed or unobserved) across all severities. In the proposed model, a GP distribution was used to approximate the margin of crash proximity. Considering the presence of an extreme value-type margin, the distribution of severity conditioned on unobserved proximity can be obtained.

The distribution of severity conditioned on different levels of proximities can be interpreted as the drivers' behavioral patterns or other unknown infrastructural parameters, which are controlled by the copula. The three chosen copulas represent different dependence structures for describing the relationship between the distance and Δv. The Gumbel copula is a member of the extreme value copulas, which is used for data that are dependent on the upper tail. The normal copula is equal in distribution to the bivariate normal distribution, where the extremal upper and lower tails are independent. The Clayton copula is suitable for modeling the lower tail dependence. The interpretation of the copula in the present study is based on the association between distance and Δv. The Gumbel copula is a suitable candidate if we assume that high Δv values can be associated with low distance values.[2] However, in this study, the final copula choice is not motivated by theoretical assumptions; instead, it is data-driven, and we should not generalize the use of Gumbel copulas to other data sets. Moreover, the association between proximity and severity may be too complex to be captured using a single-parameter copula. This is because the presence of unobserved heterogeneities can cause asymmetry in the dependence structure, which is more compatible with multiple-parameter copula. This can be further investigated using more choices of copulas when more empirical data are available, or perhaps a breakthrough in theory that explains how other parameters affect safety-relevant interactions is required, to establish each copula's nature and enable the selection based on theoretical assumptions.

One potential application for the distribution of severity is the site-specific evaluation of crash injury. Previous studies have indicated that the boundaries separating exact injury levels based on the values of interaction severity indicator do not exist and even a low threshold value can result in a relatively high probability of serious injury in a crash (Augenstein et al., 2003; Bahouth et al., 2014). A site-specific injury probability can be evaluated by "averaging the outcome model over the local road users' behaviors (in terms of Δv at crashes)" (See Appendix C).

$$P(Injurylevels|crashes) = \int_y P(Injurylevels|\Delta v = y)dF_{\Delta v|accidents}(y)$$

This can be seen as a simple mixed logit model with one latent variable Δv. In general, it is possible to carry out site-specific evaluation using a mixture model when traffic conflict measures serve as explanatory variables. For instance, Howlader et al., (2024) adopted this approach to estimate site-specific and type-specific crash injury probability with $dF_{\Delta v|accidents}(y)$ derived from a non-stationary bivariate extreme value distribution.

From a traffic conflict point of view, both BPOT and the proposed method rely on the definition of safety-relevant events. As postulated by Tarko (2021), the distribution of all road events is a mixture of two distributions, where only one represents safety-relevant events. Though the identification of a boundary that completely separates the two distributions is hardly feasible without insight information, it was argued that the non-safety-irrelevant events can be ruled out if the interactions are restricted to very near interactions because the *unawareness of hazards may be presumed if a separation between two road users becomes unacceptably small*. The proposed approach is suitable for future investigation of this topic. We showed that it is both mathematically and realistically feasible to model the whole distribution of observed interaction severity that can be regarded as safety-relevant, given low proximity.

A significant advantage of the proposed approach is the mitigation of the data-hungry problem associated with BPOT when the data size is relatively moderate or low. Among the 4772 valid interactions, the number of near interactions (exceeding only one threshold) was 477, and the number of severe interactions (exceeding both thresholds) was even smaller (99 samples). To achieve a similar sample size, the requirements imposed on the threshold values must be lowered, and the underlying GP assumption is lost. The implication is that the BPOT model suffers from more biases if we are to match its variance with the variance level of the proposed approach. Although we did not explicitly compute the variance because the implementation mainly relies on non-parametric methods

[2] Owing to the decreasing transformation, the upper tail of the transformed distance corresponds to the lower tail of distance.

Table 3
Comparison between estimation results of the bivariate POT approach and the proposed copula approach.

Method	Parameters						Estimated Probability (observed probability)			
							Severe interaction $P(X \leq 1, Y > 16)$	Less severe interaction (Observed)$P(X \leq 3, Y > 10)$	Severe crash $P(X \leq 0, Y > 16)$	Non-severe crash $P(X \leq 0, Y > 3)$
Proposed copula approach	Threshold (−4.84, N/A)	σ_1 2.603	γ_1 −0.493	σ_2 N/A	γ_2 N/A	θ 1.169	$1.18 \cdot 10^{-4}$ (N/A)	$4.3 \cdot 10^{-3}$ $(4.2 \cdot 10^{-3})$	$6.45 \cdot 10^{-5}$ (N/A)	$6.3 \cdot 10^{-4}$(N/A)
Bivariate POT approach	Threshold (−4.84, 8)	σ_1 2.593	γ_1 −0.477	σ_2 1.769	γ_2 0.262	θ 1.075	$1.23 \cdot 10^{-4}$ (N/A)	$8.34 \cdot 10^{-3}$ $(4.2 \cdot 10^{-3})$	$3.73 \cdot 10^{-5}$ (N/A)	N/A (N/A)

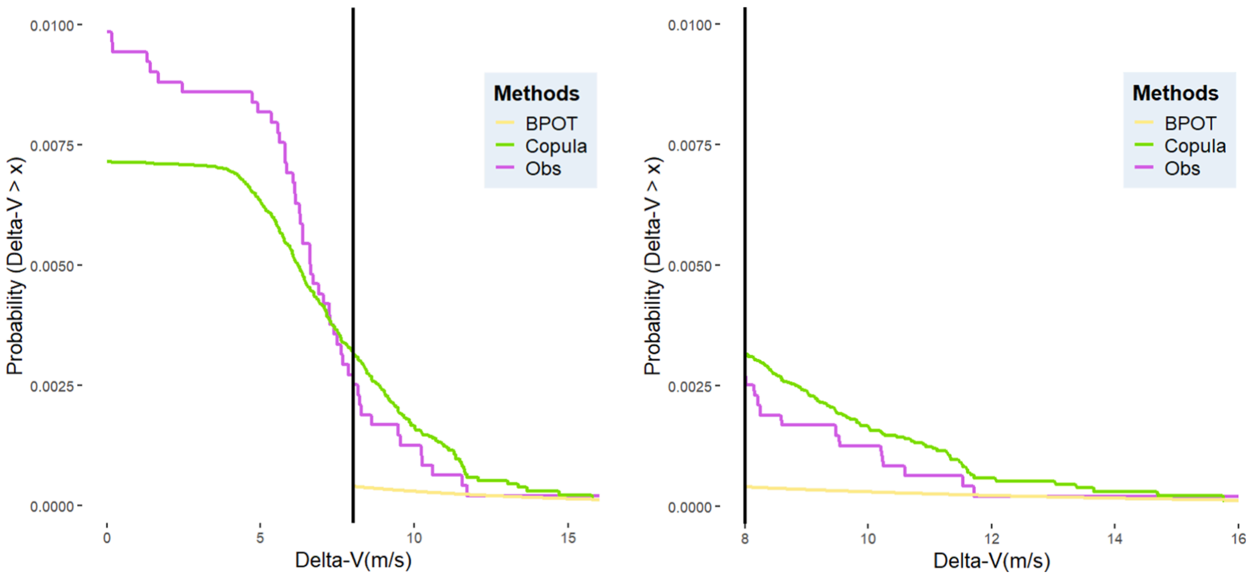

Fig. 5. The survival functions of $\Delta\boldsymbol{v}$ $\boldsymbol{P}(\Delta\boldsymbol{v} > \boldsymbol{x}|\boldsymbol{X} \leq 1)$. The black vertical line indicates the threshold for $\Delta\boldsymbol{v}$ in BPOT, below which the BPOT prediction is not available. The right plot is a zoomed-in view of the region where BPOT prediction is available.

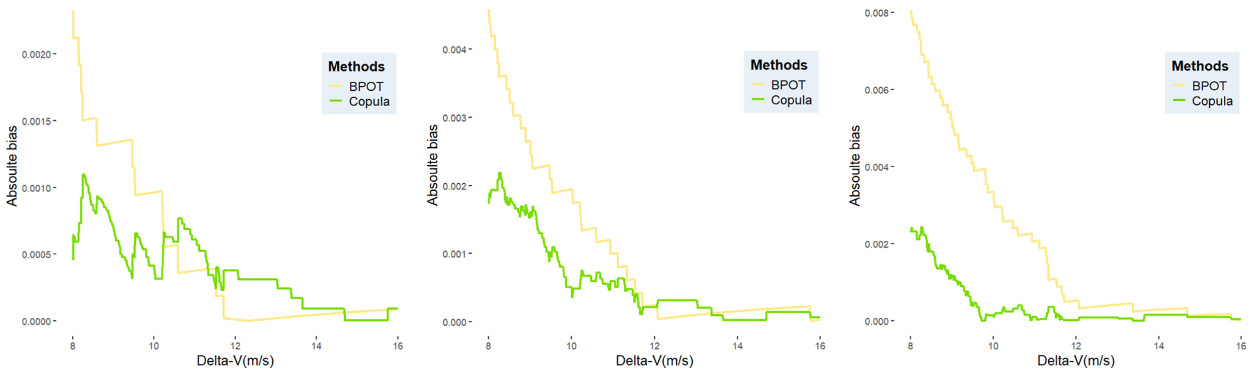

Fig. 6. Plots of the absolute bias between the survival functions of the observation $\boldsymbol{P}(\Delta\boldsymbol{v} > \boldsymbol{x}|\boldsymbol{X} \leq 1)$ (left), $\boldsymbol{P}(\Delta\boldsymbol{v} > \boldsymbol{x}|\boldsymbol{X} \leq 2)$ (middle), and $\boldsymbol{P}(\Delta\boldsymbol{v} > \boldsymbol{x}|\boldsymbol{X} \leq 3)$ (right).

(no covariance structure available), it may be insightful to validate the proposed method against BPOT based on variances. The present study did not aim to establish the absolute validity of the copula application. Instead, our objective was to demonstrate that the proposed approach not only replicates the capabilities of BPOT but also extends its boundary of defined events.

6. Conclusions

We propose a copula-based method for modeling the proximity and severity of an interaction. The proposed approach has been compared with the bivariate POT approach, revealing that our approach yields less biased predictions in overall severity and predictions of the same magnitude in interactions with high/extreme severity. The copula allows for the inclusion of the entire range of severities among events with low proximity (and non-zero probability of contact for the cases of potential collision course), even the unobserved events. Moreover, the proposed approach is applicable for events that are not properly defined in BPOT, and the prediction for less severe interaction is more aligned with observations than BPOT. Among the three dependence structures evaluated herein, the Gumbel copula was the fittest candidate for modeling near interactions.

CRediT authorship contribution statement

Zhankun Chen: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Oksana Yastremska-Kravchenko:** Writing – review & editing, Data curation. **Aliaksei Laureshyn:** Supervision, Investigation, Conceptualization. **Carl Johnsson:** Writing – original draft, Supervision, Formal analysis, Data curation, Conceptualization. **Carmelo D'Agostino:** Writing – review &

editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgment

This study is supported by the European Union (ERC, SUperSAFE, 101039222). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Appendix A. Technical steps

1. The probability measure for near interaction

$P(\cdot)$ is associated with the event space of all interactions. It is also possible to derive the probability measure $Q(\cdot)$ for only near interactions, namely $P(X \le x, Y > y|\Omega_0)$ can be replaced with $Q(X \le x, Y > y)$. This holds if Q is chosen as the conditional probability of P. The construction is done for arbitrary outcome space $\Omega_0 \subset \Omega$ and its corresponding sigma-algebra $\mathscr{F}_0 \subset \mathscr{F}$. The notations are adopted from Širjaev (2016).

Consider the probability spaces $(\Omega, \mathscr{F}, P)$ and $\mathscr{F}_0 \subseteq \mathscr{F}$, where ξ is an $\mathscr{F}$-measurable function. Then, $E(\xi|\mathscr{F}_0)$ is called the conditional expectation of ξ with respect to $\mathscr{F}_0$, if.

i. $E(\xi|\mathscr{F}_0)$ is $\mathscr{F}_0$-measurble.
ii. For every $\boldsymbol{A} \in \mathscr{F}_0$, $\int_{\boldsymbol{A}} \boldsymbol{\xi} \mathrm{d}\boldsymbol{P} = \int_{\boldsymbol{A}} \boldsymbol{E}(\boldsymbol{\xi}|\mathscr{F}_0) \cdot \mathrm{d}\boldsymbol{P}$

Define $Q(A) := \int_A E(\xi|\mathscr{F}_0) \cdot \mathrm{d}P$ and let $(\Omega_0, \mathscr{F}_0, Q)$ be a new probability space. The Radon–Nikodym derivative exists by definition and enables the change in probability space. Thus, on $(\Omega_0, \mathscr{F}_0, Q)$,

$$\forall A \in \mathscr{F}_0, Q(A) = \int_A \mathrm{d}Q = \int_A \frac{\mathrm{d}Q}{\mathrm{d}P} \cdot \mathrm{d}P = \int_{\boldsymbol{A}} \boldsymbol{E}(\boldsymbol{\xi}|\mathscr{F}_0) \cdot \mathrm{d}\boldsymbol{P}$$

$$\Leftrightarrow \frac{\mathrm{d}Q}{\mathrm{d}P} = \boldsymbol{E}(\boldsymbol{\xi}|\mathscr{F}_0)$$

where $\boldsymbol{E}(\boldsymbol{\xi}|\mathscr{F}_0) = \boldsymbol{E}(\xi)/\boldsymbol{P}(\Omega_0)$ (Theorem 1, chapter II §7, Širjaev, 2016)

Accordingly, the relation between the two probability spaces $(\Omega_0, \mathscr{F}_0, Q)$ and $(\Omega, \mathscr{F}, P)$ is established:

$$\forall A \in \mathscr{F}_0 \subset \mathscr{F},\ Q(A) = \int_A \xi \mathrm{d}P \Big/ P(\Omega_0) \tag{A1}$$

Next, we show that Eq. A(1) can be extended to the probability spaces $(\Omega^2, \mathscr{F} \otimes \mathscr{F}, P)$ and $(\Omega_0 \times \Omega, \mathscr{F}_0 \otimes \mathscr{F})$. Suppose (X, Y) is a random vector on $(\Omega^2, \mathscr{F} \otimes \mathscr{F}, P)$, we partition the outcome space Ω^2 based on the images of either entry of (X, Y) (X is chosen here). We partition $\mathbb{R}$ into finite intervals. $\mathbb{R} = \cup_{i=0}^m D_i$, in which $D_0 := (-\infty, d_0]$,$D_i := (d_{i-1}, d_i]$,$D_m := (d_m, \infty]$. Set $d_0 > 0$, and then the severe interactions are nested in D_0. Moreover, the outcome space can be partitioned using the pre-images of D_i,

$$\Omega = \cup_i \Omega_i =: \cup_i X^{-1}(D_i)$$

Define a projection $\pi_0 : (\Omega^2, \mathscr{F} \otimes \mathscr{F}) \mapsto (\Omega_0 \times \Omega, \mathscr{F}_0 \otimes \mathscr{F})$ such that:

$$\pi_0(\omega) = \omega \cdot \mathbb{I}\{(X, Y)(\omega) \in D_0 \times \mathbb{R}\}$$

Choose $\xi(\omega) = \pi_0(\omega)$, then $\forall A \in \mathscr{F}_0 \otimes \mathscr{F}$.

$$Q(A) = \int_A \pi_0 \mathrm{d}P \Big/ P(\Omega_0) = P(A) \Big/ P\left(X^{-1}(D_0)\right)$$

When $A = (-\infty, x] \times [y, \infty)$ and $x \le d_0$:

$$P(X \le x, Y > y) = Q(X \le x, Y > y) \cdot P\left(X^{-1}(D_0)\right)$$

2. Adjustment when the sample correlation (τ) is negative

Some copula families do not support data where the crash proximal indicator values and interaction severity indicator values have non-positive rank correlations. The problem can be solved by transforming Δv decreasingly. The resulting probability is unchanged owing to the invariance property of copula under monotone transformation. First, we denote the transformed Δv as $\left(\widetilde{Y_1}, \cdots, \widetilde{Y_n}\right) := (-Y_1, \cdots, -Y_n)$.

Proposition: Let (X_1, X_2) be a random vector that has a joint distribution $H_X(x_1, x_2) = C_X(F_1(x_1), F_2(x_2))$ and marginal distributions F_1, F_2. Suppose $T_1, T_2 : \mathbb{R} \mapsto \mathbb{R}$ are monotone decreasing transformations where $Y_1 = T_1(X_1)$, $Y_2 = T_2(X_2)$, and$H_Y(y_1, y_2) = C_Y(F_{T_1}(y_1), F_{T_2}(y_2))$, then $C_Y(u, v) = \overline{C}(1 - u, 1 - v)$.

proof: See (Nelsen, 2006).

Using the above proposition, we can establish a connection between the dependence structure when both margins are transformed and the dependence structure with the original data set, which is not possible if only one margin is transformed.

On the probability space $(\Omega_0 \times \Omega, \mathscr{F}_0 \otimes \mathscr{F}, Q)$, applying the proposition to $F_X(x) - C(F_X(x), F_Y(y))$, we obtain:

$$F_X(x) - C(F_X(x), F_Y(y)) = F_X(0) - (\overline{C}(F_X(x), F_Y(y)) + F_X(x) + F_Y(y) - 1)$$

$$= \overline{F_Y}(y) - C^T(1 - F_X(x), 1 - F_Y(y))$$

$$= F_{T_2}(T_2(y)) - C^T(F_{T_1}(T_1(x)), F_{T_2}(T_2(y)))$$

Then, the probability of an interaction becomes:

$$\boldsymbol{P(X \le x, Y > y) = P(X \le x, Y \le y) - F_X(0) - C(F_X(0), F_Y(y))}$$

$$= \left(\boldsymbol{F_{T_2(Y)}(T_2(y)) - C^T\left(F_{T_1(X)}(T_1(x)), F_{T_2(Y)}(T_2(y))\right)}\right) \cdot \boldsymbol{P}(\Omega_0)$$

$$\boldsymbol{P(X \le x, Y \le y) = C(F_X(x), F_Y(y))}$$

$$= \left(\boldsymbol{C^T\left(F_{T_1(X)}(T_1(x)), F_{T_2(Y)}(T_2(y))\right) + 1 - F_{T_1(X)}(T_1(x)) - F_{T_2(Y)}(T_2(y))}\right) \cdot \boldsymbol{P}(\Omega_0)$$

where T_1, T_2 are decreasing transformations for the margins, $F_{T_1(X)}(x) = 1 - \left(1 + \gamma \frac{T_1(x) - T_1(u)}{\sigma}\right)^{-1/\gamma}$, $F_{T_2(Y)}(T_2(y))$ can either be an empirical distribution or a parametric distribution, and C^T is the copula associated with the joint distribution $(T_1(X), T_2(Y))$.

Appendix B. Rank approximate z (RaZ) estimator of copula

For one-parameter copulas with parameter θ, the rank correlation τ is defined as

$$\tau = 4 \cdot E(C_\theta(U, V)) - 1 = 4 \int \int_{[0,1]^2} C_\theta(u, v) \mathrm{d}C_\theta(u, v) - 1 =: g(\theta)$$

$g : \mathbb{R} \mapsto [-1, 1]$. Moreover, if the function g is bijective such that $g^{-1}(\tau)$ is continuously differentiable with respect to τ, then the RaZ estimator $\widehat{\theta_n} = g^{-1}(\widehat{\tau})$ is a weakly consistent estimator of θ, where $\widehat{\tau}$ is the sample estimate of Kendall's τ (Theorem 1, Tsukahara, 2005).

$$\widehat{\tau} = \frac{2n}{n-1} \sum_{i=2}^{n} \sum_{j=2}^{n} \left(C_n(u_i, v_j) \cdot C_n(u_{i-1}, v_{j-1}) - C_n(u_i, v_{j-1}) \cdot C_n(u_{i-1}, v_j)\right)$$

where $C_n(u, v) = n^{-1} \sum_{i=1}^{n} \mathbb{I}\{U_{(i)} \le u, V_{(i)} \le v\}$ is the empirical copula, and $U_{(i)}, V_{(i)}$ are equal to the rank of the i^{th} observation divided by n.

Appendix C. Risk of road crash injury

The condition distribution of Δv is derived based on the results from Appendix A2, where both the proximal and consequential margins are transformed decreasingly.

$$P(Y \le y | X \le x) = \frac{P(X \le x, Y \le y)}{P(X \le x)} = \frac{Q(X \le x, Y \le y) P(\Omega_0)}{Q(X \le x) \cdot P(\Omega_0)}$$

$$= \frac{C^T\left(F_{T_1(X)}(T_1(x)), F_{T_2(Y)}(T_2(y))\right) - F_{T_1(X)}(T_1(x)) - F_{T_2(Y)}(T_2(y)) + 1}{\left(1 + \gamma\frac{T_1(x) - T(u)}{\sigma}\right)^{-1/\gamma}}$$

The conditional density is obtained by differentiating the above expression with respect to y, such that:

$$\frac{d}{dy}\left(\frac{C^T\left(F_{T_1(X)}(T_1(x)), F_{T_2(Y)}(T_2(y))\right) - F_{T_1(X)}(T_1(x)) - F_{T_2(Y)}(T_2(y)) + 1}{\left(1 + \gamma\frac{T_1(x) - T(u)}{\sigma}\right)^{-1/\gamma}}\right)$$

$$= \frac{\frac{dT_2}{dy}\frac{d\widetilde{T_2^{-1}}}{dy}f_Y(y)\cdot\left(1 - \frac{\partial C}{\partial v}\right)\Big|_{\left(F_{T_1(X)}(T_1(x)), F_{T_2(Y)}(T_2(y))\right)}}{\left(1 + \gamma\frac{T_1(x) - T_1(u)}{\sigma}\right)^{-1/\gamma}}$$

The mixed logit model based on traffic conflict indicators is expressed as:

$$P(Injurylevels|crashes) = \int_y P(Injurylevels|\Delta v = y)dF_{\Delta v|accidents}(y)$$

where,

$$dF_{\Delta v|crashes}(y) = d\Big/ds(P(Y \le s|X \le 0))\big|_y \cdot dy = \frac{\frac{dT_2}{dy}\frac{d\widetilde{T_2^{-1}}}{dy}f_Y(y)\cdot\left(1 - \frac{\partial C}{\partial v}\right)\Big|_{\left(F_{T_1(X)}(T_1(x)), F_{T_2(Y)}(T_2(y))\right)}}{\left(1 + \gamma\frac{T_1(x) - T_1(u)}{\sigma}\right)^{-1/\gamma}}\cdot dy$$

In the equation above, $f_Y(y)$ is a parametric density for Δv (untransformed) associated with measure Q. $\partial C/\partial v|_{\left(F_{T_1(X)}(T_1(0)), F_{T_2(Y)}(T_2(y))\right)} = P(X > 0|Y = y)$ is the probability of a non-crash interaction conditioned at $\Delta v = y$. When the probability is equal to one, no mass is used in the computation of injury. The integral can be expressed as the expectation and computed by Monte Carlo integration:

$$P(Injurylevels|crashes) = K^{-1}\cdot E_Q(h(Y))$$

$$K = \left(1 + \gamma\frac{T_1(0) - T_1(u)}{\sigma}\right)^{-1/\gamma}$$

$$h(y) = \frac{1}{1 + exp(-\beta_0 - \beta_1\cdot y)}\bullet\frac{dT_2^{-1}}{dy}\bullet\left(\frac{\partial C}{\partial v} - 1\right)\Big|_{\left(F_{T_1(X)}(T_1(0)), F_{T_2(Y)}(T_2(y))\right)}$$

$\frac{1}{1+exp(-\beta_0-\beta_1\cdot y)} = P(Injurylevels|\Delta v = y)$ comes from a logit probability, and the conditional copula distribution $\partial C/\partial v|_{\left(F_{T_1(X)}(T_1(0)), F_{T_2(Y)}(T_2(y))\right)}$ is evaluated numerically. Finally, the sequential Monte Carlo simulation requires the distribution of Δv at near misses $Q(Y \le y)$ to be parametric.

References

Ali, Y., Haque, M., Zheng, Z., 2022. An extreme value theory approach to estimate crash risk during mandatory lane-changing in a connected environment. Analytic Methods in Accident Research 33, 100193.

Ali, Y., Haque, M., Mannering, F., 2023a. Assessing traffic conflict/crash relationships with extreme value theory: recent developments and future directions for connected and autonomous vehicle and highway safety research. Analytic Methods in Accident Research 39, 100276.

Ali, Y., Haque, M., Mannering, F., 2023b. A Bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics. Analytic Methods in Accident Research 38, 100264.

Allen, B., Shin, B., Cooper, P., 1978. Analysis of traffic conflicts and collisions. Transportation Research Record 667, 67–74.

Amundsen, F., Hyden, C., 1977. Proceedings of first workshop on traffic conflicts. Internation Co-operation on Theories and Concepts of Traffic safety (ICTCT), Oslo, Norway.

Arun, A., Haque, M., Bhaskar, A., Washington, S., Sayed, T., 2021a. A bivariate extreme value model for estimating crash frequency by severity using traffic conflicts. Analytic Methods in Accident Research 32, 100180.

Arun, A., Haque, M., Bhaskar, A., Washington, S., Sayed, T., 2021b. A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accident; Analysis and Prevention 153, 106016.

Arun, A., Haque, M., Washington, S., Sayed, T., Mannering, F., 2021c. A systematic review of traffic conflict-based safety measures with a focus on application context. Analytic Methods in Accident Research 32, 100185.

Arun, A., Haque, M., Washington, S., Sayed, T., Mannering, F., 2022. How many are enough?: Investigating the effectiveness of multiple conflict indicators for crash frequency-by-severity estimation by automated traffic conflict analysis. Transportation Research Part C 138, 103653.
Åsljung, D., Nilsson, J., Fredriksson, J., 2017. Using extreme value theory for vehicle level safety validation and implications for autonomous vehicles. IEEE Transactions on Intelligent Vehicles 2 (4), 288–297.
Augenstein, J., Perdeck, E., Stratton, J., Digges, K., Bahouth, G., 2003. Characteristics of crashes that increase the risk of serious injuries. Annual Proceedings for the Association for the Advancement of Automotive Medicine 47, 561–576.
Bagdadi, O., 2013. Estimation of the severity of safety critical events. Accident; Analysis and Prevention 50, 167–174.
Bahouth, G., Graygo, J., Digges, K., Schulman, C., Baur, P., 2014. The benefits and tradeoffs for varied high-severity injury risk thresholds for advanced automatic crash notification systems. Traffic Injury Prevention 15 (1), 134–140.
Borsos, A., 2021. Application of bivariate extreme value models to describe the joint behavior of temporal and speed related surrogate measures of safety. Accident; Analysis and Prevention 159, 106274.
Borsos, A., Farah, H., Laureshyn, A., Hagenzieker, M., 2020. Are collision and crossing course surrogate safety indicators transferable? A probability based approach using extreme value theory. Accident; Analysis and Prevention 143, 105517.
Cafiso, S., D'Agostino, C., Kieć, M., Bak, R., 2018. Safety assessment of passing relief lanes using microsimulation-based conflicts analysis. Accident; Analysis and Prevention 116, 94–102.
Cavadas, J., Azevedo, C., Farah, H., Ferreira, A., 2020. Road safety of passing maneuvers: a bivariate extreme value theory approach under non-stationary conditions. Accident; Analysis and Prevention 134, 105315.
Chin, H., Quek, S., 1997. Measurement of traffic conflicts. Safety Science 26 (3), 169–185.
Davis, G., Hourdos, J., Xiong, H., Chatterjee, I., 2011. Outline for a causal model of traffic conflicts and crashes. Accident; Analysis and Prevention 43 (6), 1907–1919.
Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H., 2023. StrongSORT: make DeepSORT great again. IEEE Transactions on Multimedia 25, 8725–8737.
Eric, G., Richard, W., 2016. ExtRemes 2.0: an extreme value analysis package in R. Journal of Statistical Software 72 (1), 1–39.
Farah, H., Azevedo, C., 2017. Safety analysis of passing maneuvers using extreme value theory. IATSS Research 41 (1), 12–21.
Fu, C., Sayed, T., Zheng, L., 2020. Multivariate bayesian hierarchical modeling of the non-stationary traffic conflict extremes for crash estimation. Analytic Methods in Accident Research 28, 100135.
Fu, C., Sayed, T., 2021. Multivariate bayesian hierarchical gaussian copula modeling of the non-stationary traffic conflict extremes for crash estimation. Analytic Methods in Accident Research 29, 100154.
Genest, C., Favre, A., 2007. Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering 12 (4), 347–368.
Glauz, W., Migletz, D., 1980. Application of traffic conflict analysis at intersections. National Cooperative Highway Research Program Reort 219. Transportation Research Board. Washington, DC.
Gudendorf, G., Segers, J., 2010. Extreme-value copulas. Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009, Springer, New York, NY.
Hayward, J., 1972. Near-miss determination through use of a scale of danger. Highway Research Record 384, 22–34.
Howlader, M., Mannering, F., Haque, M., 2024. Estimating crash risk and injury severity considering multiple traffic conflict and crash types: a bivariate extreme value approach. Analytic Methods in Accident Research 42, 100331.
Hydén, C., 1987. The development of a method for traffic safety evaluation: the Swedish traffic conflict technique. Department of Traffic Planning and Engineering. Lund University. PhD disertation. Lund, Sweden.
Jocher, G., Chaurasia, A., Qiu, J., 2023.YOLO by Ultralytics (Version 8.0.0). https://github.com/ultralytics/ultralytics.
Johnsson, C., Laureshyn, A., D'Agostino, C., 2021a. A relative approach to the validation of surrogate measures of safety. Accident; Analysis and Prevention 161, 106350.
Johnsson, C., Laureshyn, A., D'Agostino, C., 2021b. Validation of surrogate measures of safety with a focus on bicyclist–motor vehicle interactions. Accident; Analysis and Prevention 153, 106037.
Jonasson, J., Rootzén, H., 2014. Internal validation of near-crashes in naturalistic driving studies: a continuous and multivariate approach. Accident; Analysis and Prevention 62, 102–109.
Kojadinovic, I., Yan, J., Holmes, M., 2011. Fast large sample goodness-of-fit tests for copulas. Statistica Sinica 21 (2), 841–871.
Laureshyn, A., Svensson, Å., Hydén, C., 2010. Evaluation of traffic safety, based on micro-level behavioural data: theoretical framework and first implementation. Accident; Analysis and Prevention 42 (6), 1637–1646.
Laureshyn, A., De Ceunynck, T., Karlsson, C., Svensson, Å., Daniels, S., 2017. In search of the severity dimension of traffic events: extended Delta-V as a traffic conflict indicator. Accident; Analysis and Prevention 98, 46–56.
Laureshyn, A., Varhelyi, A., 2020. The Swedish traffic conflict technique: observer's manual. Transport and Roads, Department of Technology and Society, Faculty of Engineering, LTH. Lund University, Lund, Sweden.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.
Nelsen, R., 2006. An introduction to copulas. Springer, New York, NY.
Older, J., Spicer, B., 1976. Traffic conflicts—a development in accident research. Human Factors 18 (4), 335–350.
Patterson, F., 2021. The adverse effects of paradigm and pragmatism on road safety with case studies in traffic conflicts technique and cyclist safety at roundabouts. The University of Adelaide. PhD disertation, Adelaide, Australia.
Perkins, S., Harris, J., 1968. Traffic conflict characteristics-accident potential at intersections. Highway Research Record(225), 35–43.
Pickands, J., 1975. Statistical inference using extreme order statistics. The Annals of Statistics 3 (1), 119–131.
R Development Core Team, 2023. R: a language and environment for statistical computing. https://www.R-project.org/.
Rosen, E., Stigson, H., Sander, U., 2011. Literature review of pedestrian fatality risk as a function of car impact speed. Accident; Analysis and Prevention 43 (1), 25–33.
Saunier, N., Laureshyn, A., 2021. Surrogate measures of safety. International Encyclopedia of Transportation. In: Vickerman, R. (Ed.), Elsevier. UK, Oxford, pp. 662–667.
Shelby, S., 2011. Delta-V as a measure of traffic conflict severity. the 3rd International Conference on Road Safety and Simulation, September 14-16, Indianapolis, USA.
Sievert, C., 2020. Interactive web-based Data visualization with R, plotly, and shiny. Chapman and Hall/CRC, Boca Raton, FL.
Širjaev, A., 2016. Probability. Springer, New York, NY.
Sklar, M., 1959. Fonctions de répartition à n dimensions et leurs marges. Annales de l'ISUP 8 (3), 229–231.
Smith, R., 1994. Multivariate threshold methods. Extreme Value Theory and Applications: Proceedings of the Conference on Extreme Value Theory and Applications, Volume 1 Gaithersburg Maryland 1993, Springer, New York, NY.
Songchitruksa, P., Tarko, A., 2006. The extreme value theory approach to safety estimation. Accident; Analysis and Prevention 38 (4), 811–822.
Tahir, H., Haque, M., 2024. A non-stationary bivariate extreme value model to estimate real-time pedestrian crash risk by severity at signalized intersections using artificial intelligence-based video analytics. Analytic Methods in Accident Research 43, 100339.
Tarko, A., 2018. Estimating the expected number of crashes with traffic conflicts and the lomax distribution – a theoretical and numerical exploration. Accident; Analysis and Prevention 113, 63–73.
Tarko, A., 2021. A unifying view on traffic conflicts and their connection with crashes. Accident; Analysis and Prevention 158, 106187.
Tsukahara, H., 2005. Semiparametric estimation in copula models. Canadian Journal of Statistics 33 (3), 357–375.
Wang, C., Xie, Y., Huang, H., Liu, P., 2021. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. Accident Analysis and Prevention 157, 106157.

Wang, C., Xu, C., Dai, Y., 2019. A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. Accident; Analysis and Prevention 123, 365–373.

Wickham, H., 2016. ggplot2. Elegant Graphics for Data Analysis. Springer, New York, NY.

Wramborg, P., 2005. A new approach to a safe and sustainable road structure and street design for urban areas. Proceedings of the Road Safety on Four Continents Conference, Warsaw, Poland.

Wu, K., Jovanis, P., 2012. Crashes and crash-surrogate events: exploratory modeling with naturalistic driving data. Accident; Analysis and Prevention 45, 507–516.

Yastremska-Kravchenko, O., Laureshyn, A., D'Agostino, C., Varhelyi, A., 2022. What constitutes traffic event severity in terms of human danger perception? Transportation Research Part F 90, 22–34.

Zheng, L., Ismail, K., Meng, X., 2014. Freeway safety estimation using extreme value theory approaches: a comparative study. Accident; Analysis and Prevention 62, 32–41.

Zheng, L., Ismail, K., Sayed, T., Fatema, T., 2018. Bivariate extreme value modeling for road safety estimation. Accident; Analysis and Prevention 120, 83–91.

Zheng, L., Sayed, T., 2019a. Bayesian hierarchical modeling of traffic conflict extremes for crash estimation: a non-stationary peak over threshold approach. Analytic Methods in Accident Research 24, 100106.

Zheng, L., Sayed, T., 2019b. Comparison of traffic conflict indicators for crash estimation using peak over threshold approach. Transportation Research Record 2673 (5), 493–502.

Zheng, L., Sayed, T., Mannering, F., 2021. Modeling traffic conflicts for use in road safety analysis: a review of analytic methods and future directions. Analytic Methods in Accident Research 29, 100142.

Żuchowski, A., 2016. Analysis of the influence of the impact speed on the risk of injury of the driver and front passenger of a passenger car. Eksploatacja i Niezawodność 18 (3), 436–444.