Accident Analysis and Prevention 193 (2023) 107306

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Crash frequency prediction based on extreme value theory using roadside lidar-based vehicle trajectory data

Nischal Bhattarai [a,*], Yibin Zhang [a], Hongchao Liu [a], Hao Xu [b]

[a] Department of Civil, Environmental and Construction Engineering, Texas Tech University, Lubbock, TX 79409, USA
[b] Department of Civil and Environmental Engineering, University of Nevada Reno, Nevada 89557, USA

A B S T R A C T

Crash prediction models (CPMs) are mostly developed using statistical or data-driven methods that rely on observed crashes. However, the historical crash records can be unreliable due to availability and data quality issues. Near-crashes based CPMs offer a proactive approach to predict crash frequencies prior to the occurrence of crashes. Surrogate safety measures can be used to identify near-crashes from road user trajectories. Roadside LiDAR offers an innovative approach to collect vehicle trajectory data at a microscopic resolution with high accuracy providing detailed information of all road user movements. This study presents a methodology to identify near-crashes from Roadside LiDAR based vehicle trajectory data using the surrogate indicators: TTC (Time to Collision), PET (Post Encroachment Time), ACT (Anticipated Collision Time) and MaxD (Maximum Deceleration). Additionally, time-based, and evasive-action-based surrogate measures are combined as different pairs to obtain crash probabilities using extreme value theory (EVT). The study results show that the bivariate EVT model displays a better fit to conflict extremes, predicting crash frequencies better than the univariate model. Likewise, while the bivariate model with ACT and MaxD pair performed the best in terms of accuracy, the TTC and MaxD pair was able to reflect the relative threat levels at the study intersections. Overall, the methodology lays ground for using roadside lidar based trajectory data for proactive safety analysis of signalized intersections.

1. Introduction

Crash frequency prediction lies at the core of traffic safety analysis, aiding traffic safety professionals to evaluate the safety levels of different traffic facilities and to implement suitable countermeasures when necessary. Most traditional crash prediction models employ a reactive approach based on historical crash records in order to predict crash frequencies within a roadway facility (Pei et al., 2011; Lee et al., 2017). There are several challenges associated with crash data such as underreporting, rare occurrences, erroneous reporting, and lengthy data collection period, which affect the reliability of this approach (Mannering et al., 2016). The use of "near-crash" events that occur prior to the occurrence of crashes has the potential to provide a viable alternative to the crash-data-based approach, while also making provisions for proactive safety analysis of a traffic facility.

Surrogate safety measures (SSMs) are the most popular methods used to detect near-crash events by measuring the space/time proximity of road users or detecting their evasive actions. Developed in the early 1970s (Hayward, 1971), SSMs enabled traffic safety researchers to adopt a proactive approach to road safety analysis without relying on historical crash data. Additionally, SSMs have been established as a promising method of evaluating the safety of new roadway facilities that have not encountered any crashes (Vasconcelos et al., 2014). Over the past 50 years, numerous surrogate safety indicators have been developed (Mahmud et al., 2017; Wang et al., 2021). Among their many applications, there are studies of driver behavior (Wang et al., 2019a; van Haperen et al., 2018), safety assessments (Essa and Sayed, 2018; Farah and Azevedo, 2017), before-and-after safety analysis (Tageldin et al., 2018), and real time safety prediction (Machiani and Abbas, 2016). A great deal of interest has been attracted to the development of SSMs in the recent years due to the causal link between near crashes identified by SSMs and crash frequencies estimated by field observations (Yang et al., 2021).

Conflict identification was traditionally conducted using manual observations in the field, however, the subjectivity and unreliability of this technique led to other approaches. Naturalistic driving data collected from on-board sensors of probe vehicles (Klauer et al., 2006; Guo et al., 2010), and driving-simulator experiments (Zhang and Yan, 2023) are some common techniques of conflict identification from road user level observations. Besides, Surrogate safety assessment model (SSAM) is an automated tool that has been used with the microscopic traffic simulation models to identify conflicts (Gettman et al., 2008). More recently, the advent of advanced sensor technologies has laid grounds for the facility-level conflict observations from the vehicle

* Corresponding author.
E-mail addresses: Nischal.Bhattarai@ttu.edu (N. Bhattarai), Yibin.Zhang@ttu.edu (Y. Zhang), Hongchao.Liu@ttu.edu (H. Liu), Haox@unr.edu (H. Xu).

https://doi.org/10.1016/j.aap.2023.107306
Received 24 April 2023; Received in revised form 6 September 2023; Accepted 13 September 2023
Available online 26 September 2023

trajectory data collected at specific sites. Inductive loop detectors (Dimitriou et al., 2018), UAV or infrastructure-based video cameras (Krajewski et al., 2018; Khan et al., 2018), GPS enabled smart phones (Sun et al., 2013), microwave radars (Göhring et al., 2011) etc. are some of the most common sensors used for collecting trajectory data. Roadside Light Detection and Ranging (LiDAR) has recently emerged as an innovative approach to collect vehicle trajectory data of the road users (Wu et al., 2019).

A LiDAR sensor can track and report the precise location and speed of objects within their 360 degrees scanning range. While the application of LiDAR sensors on autonomous and semi-autonomous vehicles is quite common (Campbell et al., 2018), they can also be deployed at different traffic facilities to record real-time high-resolution traffic data (HRTD) of vehicles, pedestrians, and other road users (Wu et al., 2020a). Several recent advances in automotive technology have enabled the development of connected and autonomous vehicles (CVs/AVs) with the aim of improving traffic safety and mobility. Although CV deployment is increasing, it is expected to take 25–30 years to reach 95% penetration rate (Wong et al., 2019). This will result in a very long-term mixed vehicle environment (CVs and non-CVs). The inability of non-CVs to generate real-time data will create a data gap obstructing the development of a fully connected traffic environment. Roadside LiDAR based systems were developed to fill the data gap by generating real-time HRTDs of all non-connected road users (vehicles and pedestrians) to facilitate the development of connected environments. LiDAR sensors have features such as fast data processing speed (Sun et al., 2018), and good performance in bad weather and low light conditions (Wu et al., 2020b), that gives them advantage over video cameras, the most common method of obtaining vehicle trajectory data. Additionally, with the continued development of Lidar technology and the widespread use of such sensors, the cost has significantly decreased over the past few years (Zhang et al., 2020). Therefore, roadside LiDAR sensors will likely be implemented in smart cities in the near future for use in intelligent transportation systems.

The majority of studies on roadside LiDAR have primarily focused on developing algorithms to extract high-resolution traffic data (HRTD) from point clouds in real-time. However, a few studies have utilized roadside LiDAR for traffic safety analysis (Zhang et al., 2023; Zhao et al., 2022). For example, (Wu et al., 2018; Lv et al., 2019) proposed near-crash identification methods for vehicle-pedestrian interactions using lidar-based trajectory data. Similarly, (Bhattarai et al., 2023) conducted a spatial analysis using roadside lidar trajectories to identify rear-end conflict hotspots for signalized intersections. These studies employed surrogate safety indicators to detect near-crashes, which aligned with manually observed conflicts, highlighting the potential of high-resolution trajectory data. Recognizing the promise of lidar sensors as reliable tool for facility-based traffic safety evaluations, our study employs roadside lidar trajectory data to identify conflicts, distinguish conflict types, and evaluate surrogate measures across various conflict scenarios.

Major challenges with surrogate safety based near crash identification lie in the selection of suitable thresholds and validating the crash-conflict relationship (Arun et al., 2021a). Songchitruksa and Tarko (2006) introduced Extreme Value Theory (EVT) in traffic safety analysis to cope with these issues, which has subsequently been applied in many other studies. (Zheng et al., 2014a; Farah and Azevedo, 2017; Cavadas et al., 2020). EVT is a statistical approach commonly used in finance and insurance for risk management (Embrechts et al., 1999). The idea is to estimate the probabilities of unobserved events (crashes) based on the tail-end distribution of observed events (near-crashes). Threshold selection in EVT-based modeling is performed using empirical measures based on the distribution of data, and likewise, crash frequencies at a traffic facility can be predicted based on crash probabilities to check for the reliability of observed near-crashes. For instance, Wang, et al. (2019b) employed bivariate extreme value theory to develop crash prediction models at signalized intersections using near crashes identified from video-based vehicle trajectory data. Among other applications of EVT in traffic safety analysis are roundabout safety studies (Orsini et al., 2019), before and after safety analyses (Zheng et al., 2018), and calibration of microscopic simulation models (Wang et al., 2018). The findings from these studies highlight that EVT is an effective approach for developing crash prediction models (CPMs) based on the near-crash data.

The development of EVT-based CPMs was formerly done using univariate approaches, which were later replaced with bivariate approaches that provide the advantage of incorporating two surrogate indicators representing different aspects of a crash (Zheng and Sayed, 2019). Recently, Borsos (2021) developed bivariate EVT models with temporal and speed related surrogate indicators to analyze the severity levels of left-turn related conflicts at signalized intersections. In this study, we combine proximity-based and evasive-action-based surrogate measures using the bivariate extreme value approach to develop CPMs for signalized intersections. The most popular one-dimensional time-proximity based surrogate measures TTC, and PET are compared with the recently developed ACT measure (Venthuruthiyil and Chunchu, 2022), which uses a two-dimensional approach. MaxD is used to detect evasive actions and is paired with all three proximity measures separately.

Overall, considering the abundant information that can be extracted from roadside LiDAR based point cloud data, this study aims to extend its applications towards proactive safety analysis. The second objective is to extract near-crash events from high-resolution vehicle trajectory data and assess their reliability for the purpose of developing crash frequency prediction models.

The paper is organized as follows: The Methodology section presents a series of data processing steps adopted to extract near-crashes from LiDAR data and summarizes the theory of extreme value analysis. A summary of the study's findings is presented in the Results section, which are further interpreted and analyzed in the Discussion section. Finally, this paper concludes with a summary of the research reported.

2. Methodology

In this study, a three staged methodology is employed to predict crashes at signalized intersections: (i) extraction of vehicle trajectories from point cloud data collected using Lidar sensors; (ii) development of suitable surrogate safety indices and their application to detect conflicts; (iii) development of extreme value models to predict the probability of crashes (extreme events) from the identified conflicts. Fig. 1 shows the workflow summarizing the framework of the methodology. Detailed explanation of the procedures is provided in the following section.

2.1. Data collection and trajectory extraction

Five signalized intersections (four-legged) in Lubbock, Texas were selected in this study. Texas Department of Transportation (TxDOT) has identified these locations as crash hotspots within the city limits. A VLP-32 Velodyne Lidar sensor set to a 10 Hz rotational frequency was mounted on top of a tripod stand and placed at a corner of each intersection to record the data (Fig. 2(a)). With the VLP-32 LiDAR, a 3-dimensional point cloud can be created through the emission of 32 laser channels, which are mounted in a compact housing. The housing spins rapidly to scan the surrounding environment with a range of 100 m (328ft). The rotational frequency of the LiDAR was set to 10 Hz, being able to collect 10 frames of data with 600,000 3D points per second. It can cover a 360° horizontal field of view and a 30° vertical field of view with ±15° up and down.1.5 h of data was collected for each intersection, during evening commute periods (4 pm-5:30 pm) on 5 separate weekdays of October 2022. The time window was selected so as to represent the peak-traffic conditions based on the examination of 24-hour traffic volumes of all the intersections.

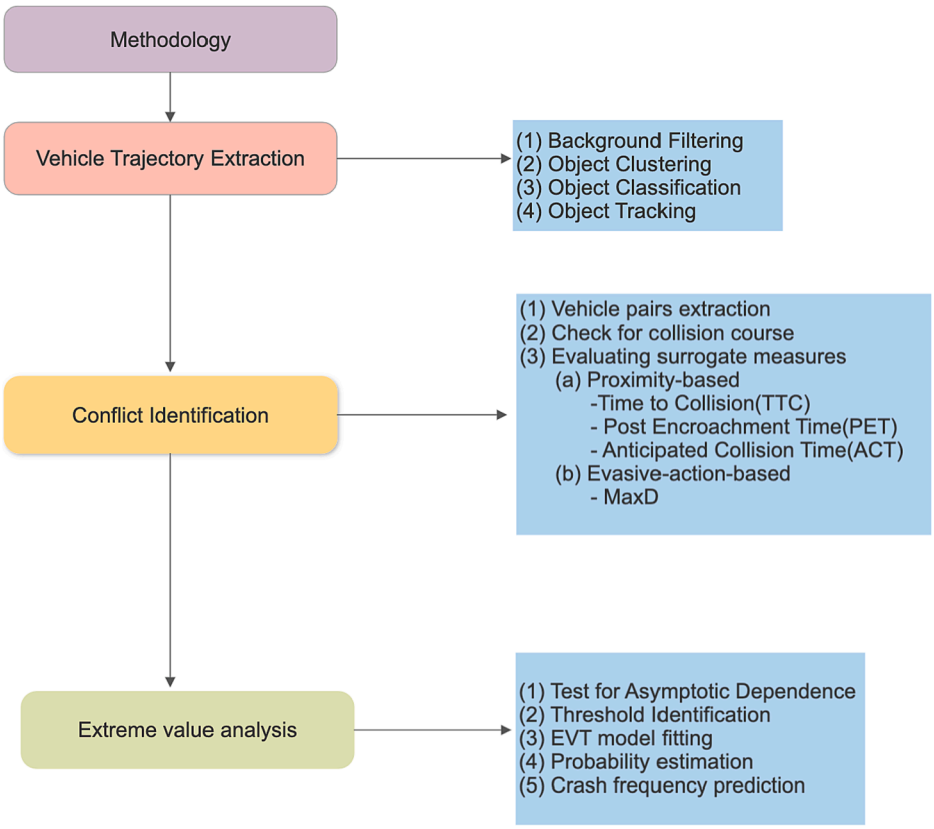

Fig. 1. Framework of the methodology.

The raw 3-D point clouds obtained from the Lidar sensor, as shown in Fig. 2(b), were then processed to obtain vehicle trajectories. This study used a generalized four-step algorithm developed by Zhao et al. (2019) for the processing of roadside-Lidar data. The four steps include: Background Filtering, Object Clustering, Object Classification, and Object Tracking. A point-density based filtering technique was used to filter both static and dynamic background points from all the frames (Wu et al., 2017). After background filtering, the remaining points were clustered using a modified DBSCAN algorithm with varying parameters depending on their proximity to the sensor. For classification, an artificial neural network model was developed using the parameters: number of points in the cluster, distance to the lidar sensor, and direction of cluster points distribution. The model classified the clusters to respective road user group: vehicles, pedestrians, bicycle. Lastly, a Global Nearest Neighbor (GNN) algorithm was applied for tracking, which uses the spatial and temporal relations between the reference points of each object within simultaneous data frames to match the closest pairs of points. A more detailed explanation of the vehicle trajectory extraction procedure can be found in the previous studies of the authors (Zhang et al., 2022; Zhang et al., 2023).

As such, every road user's trajectory data was automatically extracted using the above-described data processing pipeline. Cubic spline interpolation was used to populate the missing datapoints caused by occlusion and LOESS regression filter was used to smooth the vehicle trajectories. Additional description about the smoothing procedure can be found in a previous study by the authors (Bhattarai et al., 2023). Lastly, the vehicle trajectories obtained for each intersection were georeferenced using the respective reference markers (traffic signal poles; Fig. 2(c)) shows an example of the vehicle trajectories collected at one of the study sites. Table 1 presents a sample of the processed trajectory data showing all the variables used in this study.

2.2. *Conflict identification using surrogate safety indices*

Surrogate safety measures (SSMs) are widely adopted in traffic safety research for identifying traffic conflicts and safety performance evaluation. The availability of a high-resolution trajectory dataset, as prepared for this study, allows to continuously evaluate the conflict threat at a microscopic timescale with the help of surrogate safety indicators. The following sections describe the procedure carried out in this study for conflict identification.

2.2.1. *Identification of vehicle pairs*

A vehicle continuously interacts with multiple neighboring vehicles in the same, adjacent, or opposite lane at one time. Consequently, all vehicles present at the intersection at one time may encounter a conflict. Therefore, all the vehicle trajectory pairs with common FrameID were initially extracted from the dataset and those exceeding 10 datapoints were recorded as a vehicle pair for conflict identification. For instance, let's suppose there are 5 vehicles (Veh_1, Veh_2, Veh_3, Veh_4, Veh_5) at an intersection at one time sharing common Frame IDs. 10 different combinations of two vehicle pairs are extracted which are then used for conflict identification. It should be noted that the study does not take vehicle-pedestrian conflicts or single-vehicle conflicts into consideration for the analysis. Likewise, as all the pairs identified are two-vehicle pairs, there is a chance that a multi vehicle (more than 2) conflict may be identified as separate conflicts. To account for this, conflicts involving same group of vehicles (in multiple pairs) are merged later to define as a single conflict. Table2 summarizes the number of vehicle pairs extracted during the data-collection period for all 5 intersections.

2.2.2. *Check for vehicle's collision course*

Prior to the evaluation of SSM indicators for conflict identification, it is necessary to check for the collision course of the vehicle pairs involved in a possible conflict to make sure if the vehicles are in a path leading to

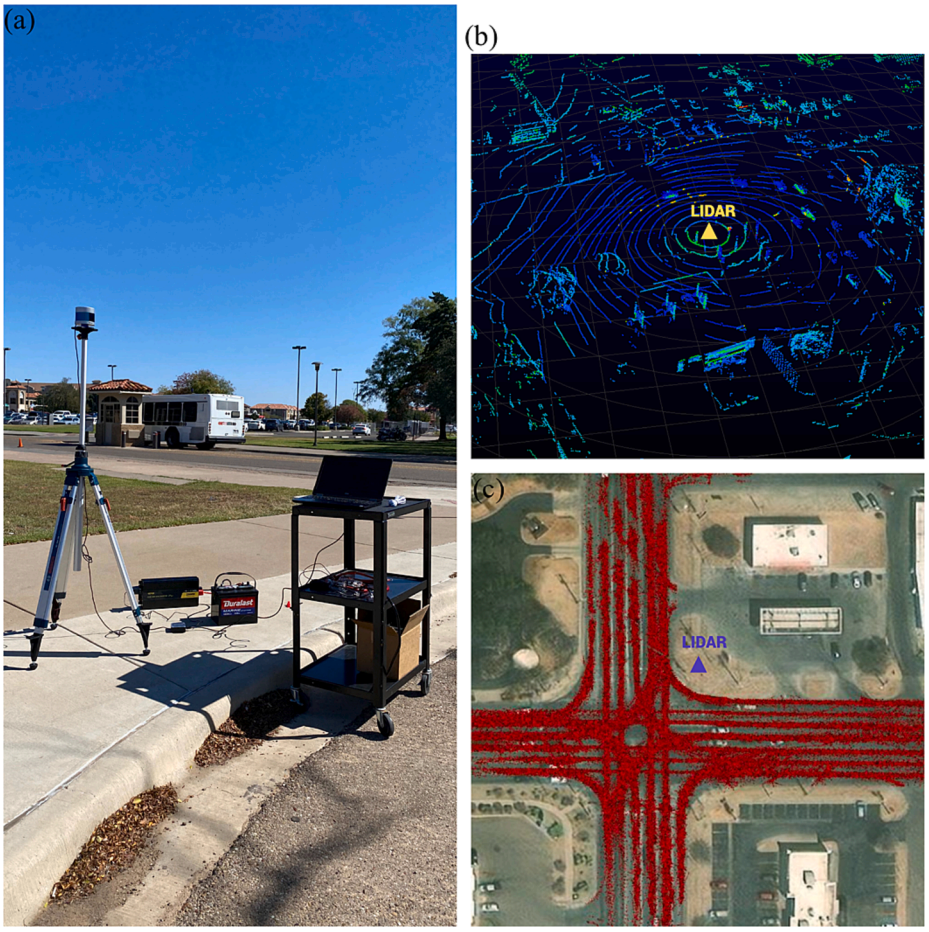

Fig. 2. (a) Roadside-Lidar sensor unit, (b) Raw point cloud data, (c) Processed vehicle trajectories.

Table 1
A sample of processed trajectory data.

Object ID	Frame ID	Object Length	Object Width	Coordinate_X (m)	Coordinate_Y (m)	Longitude	Latitude	Velocity (m/s)	Acceleration (m/s^2)	Average Acceleration (m/s^2)
1437	2294	6.589	1.617	21.9446	−21.83	−101.9396956	33.59273131	9.76	5.03	2.34
1437	2295	6.569	1.621	22.2027	−22.9514	−101.9396951	33.59274196	10.27	0.34	
1437	2296	6.609	1.263	22.5425	−24.1993	−101.9396951	33.59275397	10.3	2.13	
1437	2297	6.292	1.527	22.8613	−24.9837	−101.9396962	33.59276183	10.51	−0.86	
1437	2298	6.578	1.195	23.4056	−25.84	−101.9396995	33.59277098	10.43	6.46	
1437	2299	7.551	1.182	23.696	−26.7486	−101.9397003	33.59278002	11.07	5.53	
1437	2300	6.623	1.226	24.0463	−27.7323	−101.9397012	33.59278976	11.63	2.69	
1437	2301	7.066	1.171	24.415	−28.6885	−101.9397024	33.59279932	11.9	2.11	

Table 2
Summary of vehicle pairs extracted.

Intersection Name	Intersection Number	Vehicle Count	Pairs Identified
82nd st./Milwaukee ave.	1	18,342	612,659
82nd st./Slide ave.	2	19,618	735,620
Quaker ave./50th st.	3	16,849	559,558
Ave. Q/50th st.	4	12,619	330,864
Frankford ave./4th st.	5	13,920	401,093

collision. A novel technique was adopted in this study to check the collision course at every instance, for every vehicle pair, using the trajectory data-points. First, the coordinates of 5-consecutive trajectory points were mapped to generate a linear regression line and define the vehicles' tentative path as shown in Fig. 3(a and b). The points were assumed to follow a linear path, considering such short timeframe. The regression lines were then projected to identify the point of intersection (PoI). After obtaining the coordinates for the point of intersection, it's position with respect to the vehicle travel direction was noted. If the PoI lies in the same direction as the direction of both the vehicles, they are identified to be in a collision course. The comparisons are made using the x-coordinates of the initial and terminal points of each vehicle. Fig. 3 (a) shows an example of two vehicles not in a collision course, whereas Fig. 3(b) shows an example of two vehicles in an angle collision course. Fig. 3(c) shows the flowchart demonstrating the algorithm as described above. One limitation of this approach arises when the vehicle paths are parallel to each other, in which case the lines do not intersect, and PoI cannot be calculated. To deal with this, condition of parallel lines needs to be checked initially, and if true, two cases arise—(i) if the distance

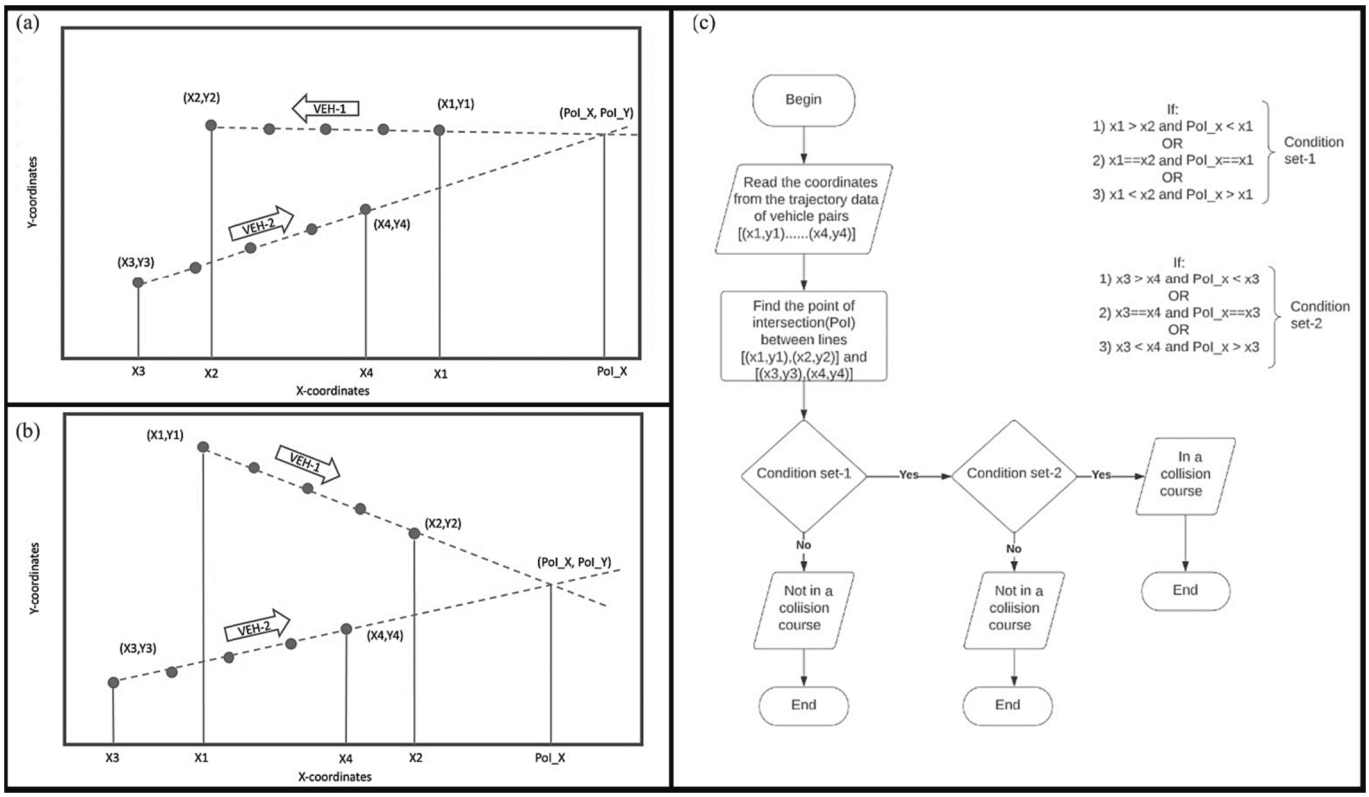

Fig. 3. (a) Trajectory of vehicles not in a collision course, (b) Trajectory of vehicles in collisions course, (c) Flowchart to check for vehicle collision course.

between the lines is greater than 6ft, they are not in a collision course, (ii) if the distance between the lines is smaller than 6ft, they are considered to be in a collision course.

2.2.3. *Proximity based conflict indicators*

Surrogate safety indicators are the metrics that are used to measure the riskiness of a driving maneuver or it's nearness to collision. Nearness to collision can be measured in terms of both time-based and deceleration-based proximity. Several SSMs have been developed over the past years to quantify the closeness to collision like TTC, PET, DRAC (Deceleration Rate to Avoid Crash), MADR (Maximum Available Deceleration Rate), SDI (Stopping Distance Index) etc. (Wang et al., 2021). The proximity-based indicators employed in this study have been described in the following sections:

Time to Collision (TTC)

Hayward (1972) defined TTC as "The time required for two vehicles to collide if they continue at their present speed and on the same path." Almost one-third of the studies related to surrogate safety measures have employed TTC measure either uniquely or in combination with other surrogate safety measures (Arun et al., 2021b). Studies have also reported that the conflicts identified using this measure correlate with the actual crashes at different levels (Johnsson et al., 2021; Wali et al., 2020). In this study, instantaneous TTC values were calculated as the minimum time required for a vehicle pair to meet at the projected point of intersection (PoI) at every timeframe (0.1 s).

Post Encroachment Time (PET)

PET is defined as the difference in time between the departure of one road user from a point of potential collision to the arrival of another road user to the same point (Allen et al. 1978). Previous applications have found PET to be an effective surrogate measure for crash prediction, especially at low thresholds (Peesapati et al., 2013). The instantaneous PET values were calculated based on the time difference between two vehicles to reach the encroachment area where the vehicle trajectories intersect.

Anticipated Collision Time (ACT)

Despite all the advantages of applying traditional surrogate indicators for crash frequency prediction, their dependency on the conflict type and traffic scenarios limits their applicability for comprehensive safety evaluations. The formulation of TTC, for instance, is based on car-following scenarios, which is suitable for detecting rear-end conflicts, but fails to capture other conflicts (Zheng et al., 2014b; Nadimi et al., 2020). Likewise, studies have determined PET to be more suitable for traversal trajectories (right angle or crossing conflicts) (Mahmud et al., 2017).This is primarily due to the one-dimensional nature of these surrogate indicators, whereas multi-vehicle interactions are mostly two-dimensional. To deal with this issue, Venthuruthiyil and Chunchu (2022) developed a new proximity-based surrogate indicator – Anticipated Collision Time (ACT). Similar to TTC, ACT measures the time remaining before two vehicles on a collision course collide with each other; however, it measures proximity in a two-dimensional plane. Evaluation of ACT is based on two variables: (i) shortest distance between the vehicles (δ) and (ii) the closing-in rate in the shortest distance direction ($\frac{\partial\delta}{\partial t}$). Mathematically,

$$ACT = \frac{\delta}{\frac{\partial\delta}{\partial t}} \tag{1}$$

$$\frac{\partial\delta}{\partial t} = \frac{\delta(i) - \delta(i-1)}{t(i) - t(i-1)} \tag{2}$$

where,

$\frac{\partial\delta}{\partial t}$ = closing-in rate.

$\delta(i)$ = shortest distance between the vehicles at current time frame.$t(i)$

$\delta(i-1)$ = shortest distance between the vehicles at previous time frame.$t(i-1)$

To determine the shortest distance first, a bounding box oriented along the horizontal axis was drawn around the reference point to locate the tentative vehicle corner points from the trajectory reference point

(x, y) as shown in Fig. 4(a). The length(l) and width(b) of the vehicles are obtained from the trajectory data. After that, the corner points were rotated by the orientation angle that was obtained from the slope of the regression lines denoting the vehicle path. Then, the rotated co-ordinates, as shown in Fig. 4(b) were obtained as shown in Eq. (3). Finally, the closest corners between a vehicle pair were identified and the distance between them was selected as the shortest distance.

$$\begin{matrix}(x^1,y^1)\\(x^2,y^2)\\(x^3,y^3)\\(x^4,y^4)\end{matrix} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} * \begin{Bmatrix} \left(x+\frac{l}{2}, y+\frac{b}{2}\right) \\ \left(x+\frac{l}{2}, y-\frac{b}{2}\right) \\ \left(x-\frac{l}{2}, y-\frac{b}{2}\right) \\ \left(x-\frac{l}{2}, y+\frac{b}{2}\right) \end{Bmatrix} \quad (3)$$

Closing-in rate is the rate at which vehicles approach one another. The previous study used the vehicle's state information (velocity, acceleration, heading angle and yaw rate) at consecutive timeframes to obtain the instantaneous closing-in rate value. In this study, an empirical approach was adopted leveraging the microscopic resolution of the Lidar based trajectory data. As shown in Fig. 5(a), shortest distance between the vehicle-pair is evaluated at n consecutive timeframes ($\delta_{i-1} \cdots .. \delta_{i-n}$) prior to the current timeframe (i). Then, the closing-in rate at timeframe (i) is calculated as the average rate of change of shortest distance between the vehicles observed during the preceding n number of time frames (5 frames used in this study).

Mathematically, it can be presented as:

$$\left(\frac{\partial \delta}{\partial t}\right)_i = \frac{\sum_{s=i-n}^{s=i} \frac{\delta_s - \delta_{s+1}}{t_{s+1} - t_s}}{(t_i - t_{i-n}).f} \quad (4)$$

where,

f = number of data-frames per second (10 in this study).

t_a = time instance recorded at a^{th} frame.

It is to be noted that the vehicle's closing-in rate is assumed to remain constant up to the collision time. The closing in rate is negative when the vehicles are moving away from each other and are not in a collision course, as shown in Fig. 5(b). However, a positive close in rate does not mean the vehicles are in a collision course. For example, Fig. 5(c) shows a scenario where vehicles are closing in first and then moving away. Since, the vehicles are not in a collision course, the evaluated closing-in rate is not significant. In this manner, the instantaneous ACT values were calculated at every time step for all vehicle trajectory pairs.

2.2.4. Evasive action-based conflict indicator

Past studies have shown that proximity-based indicators fail to capture the evasive actions (Tageldin and Sayed, 2016). An evasive action generally involves a significant change either in the speed or in the direction of the road user to avoid the collision. Evasive-actions capture nearness to collision from a different perspective, therefore making it essential to include them in the crash-prediction models based on traffic conflicts. In this study, we use the surrogate measure MaxD (Maximum Deceleration) to capture the evasive actions by measuring the intensity of vehicle braking during different driving maneuvers. The instantaneous acceleration/deceleration obtained from the trajectory data are used to obtain the MaxD values for a defined time interval.

2.2.5. Conflict-Extraction

After separating the vehicle pairs and checking for collision courses at every frame (0.1 s), all proximity based SSMs (TTC,PET and ACT) are evaluated separately for each vehicle pair. Such short-term interactions, however, cannot be defined as conflicts; the exposure period of such interactions needs to be taken into account. Therefore, the entire time-frame for a vehicle pair is segmented into temporal windows of 10 data points (1 s) and only those windows with consistent risk exposure are filtered out as conflicts and the minimum value of the corresponding proximal-SSM over the window is noted. To check for risk exposure, first, the collision course values at all instances are inspected and only those windows with consistent collision course are filtered out. Then, windows with either of TTC, PET or ACT exceeding the thresholds at all instances are selected as conflicts. Subsequently, the maximum deceleration (minimum acceleration) value out of all 10 instances in each temporal window is selected as the MaxD value for the corresponding window. A MaxD value of 0 is noted if none of the instances have negative-acceleration (deceleration) values.

2.3. Extreme value analysis.

Extreme Value Analysis can be used to predict the probability of rare events based on the distribution of frequently occurring events. In terms of its application in this study, conflicts that occur quite frequently are used to predict traffic crashes that are very rare. There are two main approaches within the EVT context:

- Block-Maxima (BM) approach: The distribution is divided into several blocks and maximum values in each block are used as the extreme values that are modeled with a generalized extreme value (GEV) distribution. A detail on this approach can be found on (McNeil, 1999).
- Peak-over threshold (POT) approach: Extreme values are selected based on a suitable threshold value. Values above a certain threshold in the distribution are considered as extremes, that are modeled with a generalized pareto distribution (GP) distribution.

Previous applications of EVT for conflict-based crash estimation have established that the peak-over threshold method performs better than the block-maxima method, especially for dataset from small observation-periods (Wang et al., 2019b). Therefore, a POT based EVT approach was adopted in this study to predict crashes at signalized intersections from identified conflicts. The following sections provide a brief summary of the peak over threshold approach, a more detailed explanation can be found in (Coles et al., 2001).

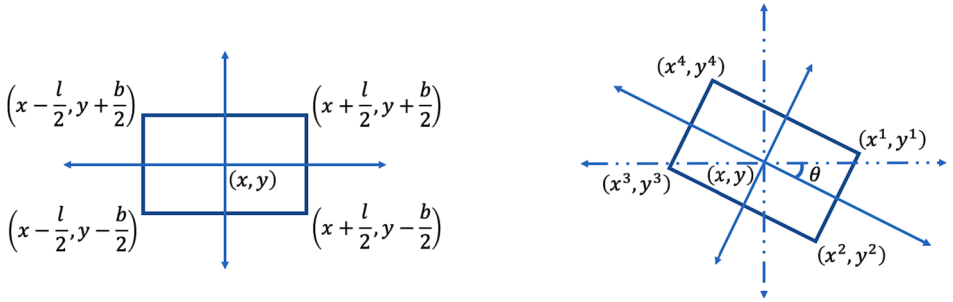

Fig. 4. (a) Bounding box to define tentative vehicle corner points, (b) Rotation along vehicle orientation angle.

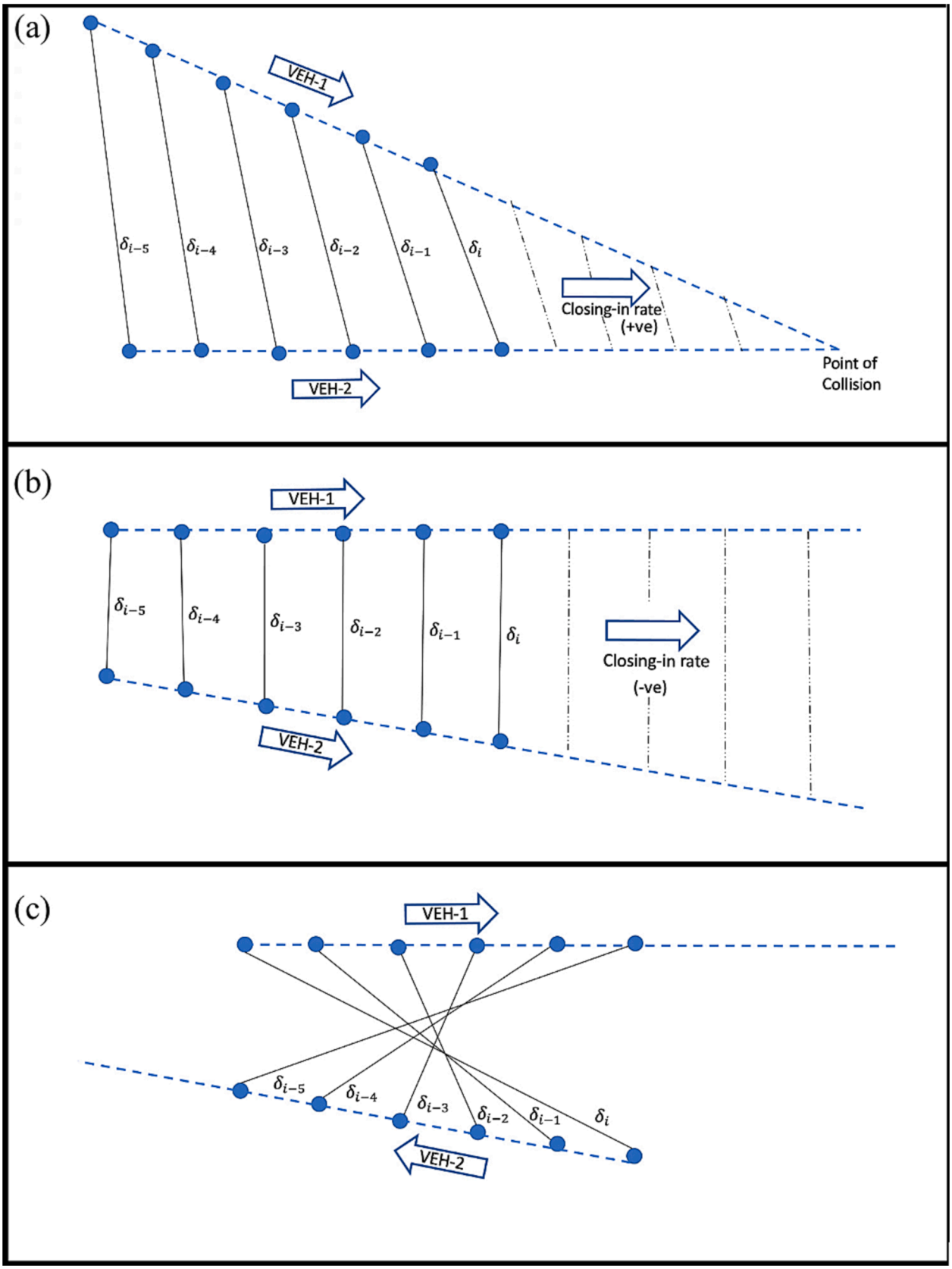

Fig. 5. Closing in rate of vehicle pairs.

Univariate POT approach

For the univariate approach, let's suppose that a sequence of independent and identical random variables $X_1, X_2 \cdots\cdots X_n$ are defined by a distribution function $F(x)$. All the values exceeding some high threshold u are considered as extreme values, defining the threshold exceedance as: $Y_i = X_i - u$. For large enough values of u, the distribution function of $Y_1, Y_2 \cdots\cdots Y_n$ is approximately a Generalized Pareto distribution (GP):

$$F(y) = 1 - \left(1 + \xi\frac{y}{\sigma}\right)^{\frac{-1}{\xi}} \tag{5}$$

where,

ξ = shape parameter ($-\infty < \xi < \infty$).

σ = scale parameter ($\sigma > 0$).

Bivariate POT approach

Building upon the univariate approach, the bivariate approach involves modeling the joint distribution of two extreme variables. Let's suppose that a sequence of independent and identical random variable pairs (X_1, Y_1), (X_2, Y_2)..... (X_n, Y_n) are defined by a joint distribution function $F(x,y)$. Similar to the univariate case, for suitable thresholds u_x and u_y, the marginal distributions of $F(x,y)$ each follow a univariate generalized pareto distribution with respective location, shape and scale parameters (μ_x, ξ_x, σ_x) and $\left(\mu_y, \xi_y, \sigma_y\right)$. The marginal distributions can be generalized with the parameters as:

$$\widetilde{X} = -\left(log\left\{1 - \mu_x\left[1 + \frac{\xi_x(X - u_x)}{\sigma_x}\right]^{\frac{-1}{\xi_x}}\right\}\right)^{-1} \tag{6}$$

$$\widetilde{Y} = -\left(log\left\{1 - \mu_y\left[1 + \frac{\xi_y(Y - u_y)}{\sigma_y}\right]^{\frac{-1}{\xi_y}}\right\}\right)^{-1} \tag{7}$$

This results in a distribution function:

$$G(x,y) = exp\{-V(\widetilde{x},\widetilde{y})\}, x > u_x, y > u_y \tag{8}$$

where,

$$V(x,y) = 2\int_0^1 \max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right)dH(\omega) \tag{9}$$

And H is a distribution function on the interval [0,1], satisfying the mean constraint:

$$\int_0^1 \omega dH(\omega) = \frac{1}{2} \tag{10}$$

Several parametric families can be used to define $G(x,y)$ satisfying the conditions in Eqs. (9) and (10). Logistic and Asymmetric logistic models are the most common models used.

The logistic GP model has the form of:

$$G(x,y) = \exp\left[-\left(x^{\frac{1}{r}} + y^{\frac{1}{r}}\right)^r\right] \tag{11}$$

where, $r \in (0,1)$ and the level of dependency between the variables increases from 0 (independent) to 1 (perfect dependency).

And the asymmetric logistic GP model has the form of:

$$G(x,y) = exp\left\{-(1-a)x - (1-b)y - [(ax)^{\frac{1}{r}} + (bx)^{\frac{1}{r}}]^r\right\} \tag{12}$$

where, $r \in (0,1]$,$0. \leq a,b \leq 1$.

The variables are independent when either $r = 1$, a = 0 or b = 0 and the variables are completely dependent when a = b = 1 and $r = 0$.

Model estimation

For a univariate case, the marginal GPD model can easily be estimated using maximum likelihood estimator. However, for a bivariate case, it is possible that the pair may exceed a specified threshold along just one margin. It is therefore essential to apply a censored likelihood function Eq. (13)). Depending on whether the observations exceed threshold u_x, u_y, both or none, four different regions can be defined on the plane, as shown in Eq. (14), and the parameters can be estimated by maximizing the log likelihood function.

$$L(\theta) = \prod_{i=1}^{n}\psi(\theta;(X, Y)) \tag{13}$$

where,

$$\psi(\theta;(X,Y)) = \begin{cases} \frac{\partial^2 F}{\partial x \partial y}\|_{(x,y)} if X > u_x, Y > u_y \\ \frac{\partial F}{\partial x}\|_{(x,y)} if X > u_x, Y \leq u_y \\ \frac{\partial F}{\partial y}\|_{(x,y)} if X \leq u_x, Y > u_y \\ F\left(u_x, u_y\right) if X \leq u_x, Y \leq u_y \end{cases} \tag{14}$$

Surrogate variable pairs

In this study, the selection of bivariate pairs was made based on the combinations of proximity-based and evasive-action based surrogate indicators. Three pairs—(i) ACT Vs MaxD (ii) PET Vs MaxD (iii) TTC Vs MaxD were used together to develop different bivariate POT models for each intersection. A threshold value of 3 s was used for ACT,TTC and PET to filter out the conflicts and obtain the distribution. As the methodology focuses on the extreme events that have small values of these surrogate indicator values, by selecting 3 s thresholds, it can be ensured that all extreme events are captured while avoiding unnecessary data processing.

Similarly, a 0 s value of ACT, TTC and PET indicate crash occurrence, therefore the thresholds for the time-proximity indicators were set as 0 s

Table 3
Descriptive statistics of surrogate measures.

Intersection	Indicator	Observations	Statistics		
			min	max	mean
1	PET	160	0.6	2.9	1.87
	PET_MaxD	160	0.03	2.72	1.18
	TTC	267	0.45	2.98	2.03
	TTC_MaxD	267	0	2.89	0.93
	ACT	210	0.55	2.99	2
	ACT_MaxD	210	0.02	2.69	0.98
2	PET	172	0.6	2.9	1.92
	PET_MaxD	172	0.02	2.37	1.10
	TTC	254	0.48	2.99	2.12
	TTC_MaxD	254	0	2.62	0.82
	ACT	235	0.67	2.99	1.96
	ACT_MaxD	235	0	2.8	0.82
3	PET	135	0.5	2.9	1.83
	PET_MaxD	135	0.01	2.19	1.30
	TTC	220	0.73	2.96	2.05
	TTC_MaxD	220	0	2.73	1.04
	ACT	212	0.62	2.98	1.99
	ACT_MaxD	212	0.04	2.82	1.04
4	PET	95	0.6	2.9	1.77
	PET_MaxD	95	0	2.50	1.13
	TTC	185	0.51	2.99	2.11
	TTC_MaxD	185	0.02	2.68	0.94
	ACT	153	0.64	2.97	2.03
	ACT_MaxD	153	0	2.27	1.02
5	PET	102	0.7	2.9	2.01
	PET_MaxD	102	0.3	2.58	1.05
	TTC	143	0.38	2.99	1.97
	TTC_MaxD	143	0	2.74	1.29
	ACT	167	0.44	2.98	1.86
	ACT_MaxD	167	0	2.49	1.04

to evaluate the probability of crashes. Likewise, as shown in Table 3, the maximum MaxD values corresponding to each proximity indicator in are below $3m/s^2$. Therefore, MaxD value of $3m/s^2$ was selected as the extreme value to differentiate between crash and conflicts. The joint probability of threshold exceedance $(X > u_x \vee Y > u_y)$ for each pair was evaluated to predict the crash frequency at each intersection.

3. Results and analysis

3.1. Descriptive statistics of conflicts and comparison of surrogate safety indices

The methodology was applied to the vehicle trajectory dataset prepared as described in section 2.1. Data processing was carried out in batches, each batch containing trajectories for an intersection over a 30-minute period. The conflicts identified for every batch were aggregated together for each intersection separately. Table 3 presents the descriptive statistics of the surrogate measures evaluated for all five study intersections. PET, TTC, and ACT denote the proximal indicators whereas PET_MaxD, TTC_MaxD, and ACT_MaxD denote the maximum deceleration values respectively.

3.2. Test-for asymptotic dependency between the variables

Bivariate extreme value distributions are based on the assumption that the margins are asymptotically dependent or perfectly independent at the extreme levels. In case of asymptotic independence, it may not be appropriate to use a class of bivariate extreme value distribution to model the tail distribution of the data, resulting in underestimation of the probability of two extreme events occurring simultaneously (Dutfoy et al., 2014). A χ -statistics test was conducted to examine the asymptotic dependency between the pairs of surrogate measures, prior to modeling their dependence at the extreme levels. χ provides a measure of tail dependence of distributions that are asymptotically dependent, where

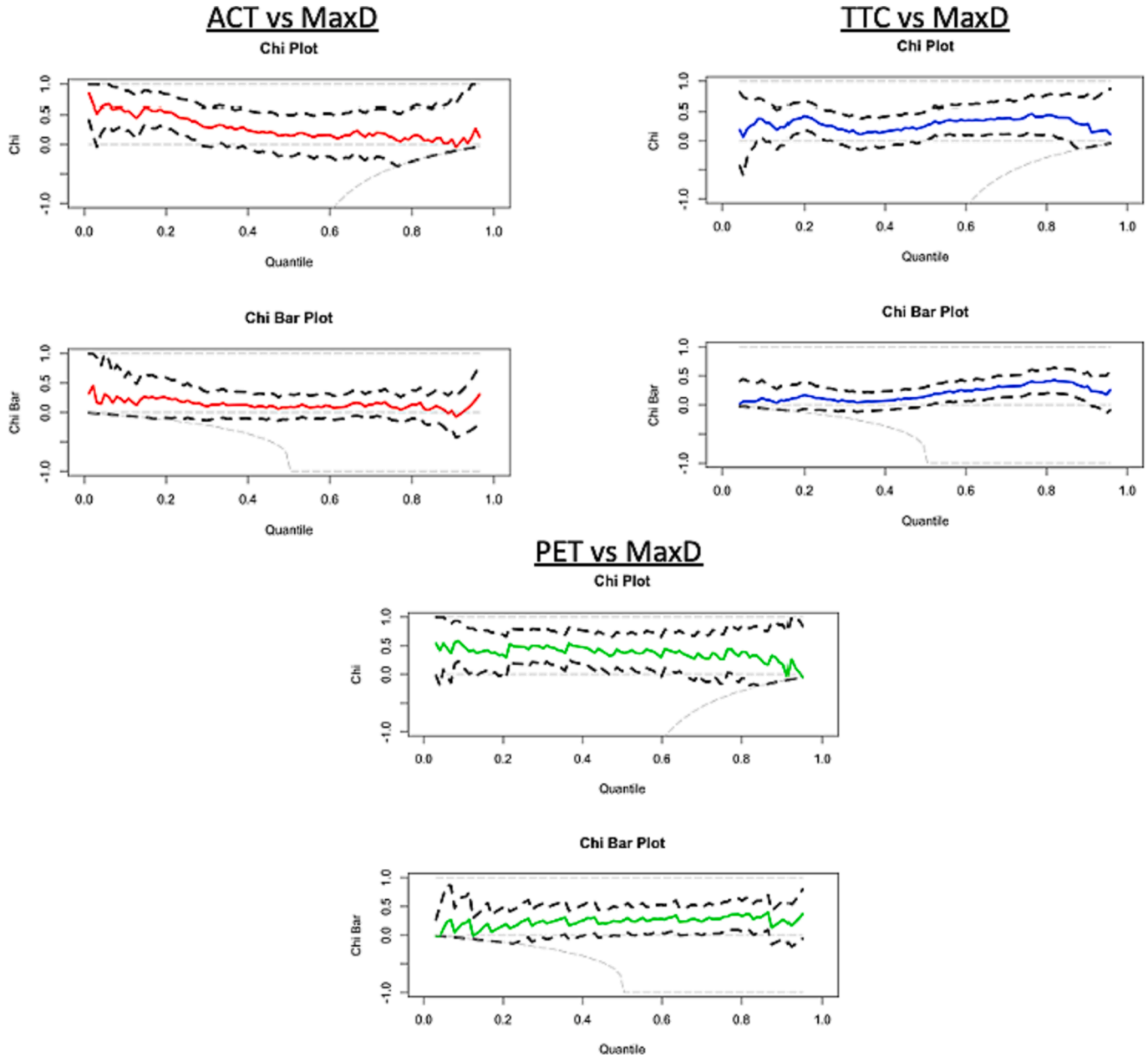

Fig. 6. Test for asymptotic dependency of conflicts at Intersection no. 1.

the value of χ increases with increasing strength of dependence at extreme levels. However, χ fails to provide the measure of tail dependence for asymptotically independent distributions, therefore an alternative $\overline{\chi}$ -statistics test was conducted to overcome this deficiency. If $\overline{\chi} = 1$ and $0 < \chi \leq 1$, then the variables are asymptotically dependent and χ gives the measures of the strength of dependence. On the other hand, if $-1 < \overline{\chi} \leq 1$ and $\chi = 0$, then the variables are asymptotically independent and $\overline{\chi}$ gives a measure of the strength of associations.

Fig. 6 shows the empirical estimates of chi and chi-bar distribution for the variable pairs at Intersection 1. The solid lines represented by the colors red, blue, and green denote the χ and $\overline{\chi}$ values for ACT vs MaxD, PET vs MaxD and TTC vs MaxD respectively. On the other hand, the bold dotted lines indicate the 95% confidence levels of the χ and $\overline{\chi}$ values for each bivariate pair respectively. The $(\chi, \overline{\chi})$ values at higher quantile for PET Vs MaxD, ACT Vs MaxD and TTC Vs MaxD were (0.47,0.53), (0.38,0.64) and (0.36,0.73)respectively. All the plots seem consistent with $\chi > 0$ and the possibility that $\overline{\chi}(u) \to 1$ as $u \to 1$. These observations lend support to discard asymptotic independence while supporting the use of bivariate extreme value models to describe the dependence structure between the variable pairs.

3.3. Threshold selection for bivariate POT modeling

Threshold selection was conducted referring to the spectral measure plot (also known as bivariate threshold choice plot) with the approximation of the spectral measure. The empirical spectral value should be close to the theoretical spectral measure, which is equal to 2 in case of a bivariate extreme case. In other words, k value (k_o) corresponding to H ([0,1]) is noted, based on which the number of samples are selected as the extreme values.

Fig. 7 presents the optimal threshold selection based on the empirical spectral values for intersection 1. The k_o values obtained for ACT vs MaxD, TTC Vs MaxD and PET Vs MaxD were 52,58,62 respectively. Thus, observations with first respective largest radial coordinates were selected as the extreme cases, corresponding to the threshold values of (-1.42 s, 1.76 m/s^2), (-1.17 s, 1.97 m/s^2), (-1.28 s, 1.73 m/s^2), respectively for the three bivariate pairs. Likewise, a similar approach was adopted to evaluate the threshold values for each intersection.

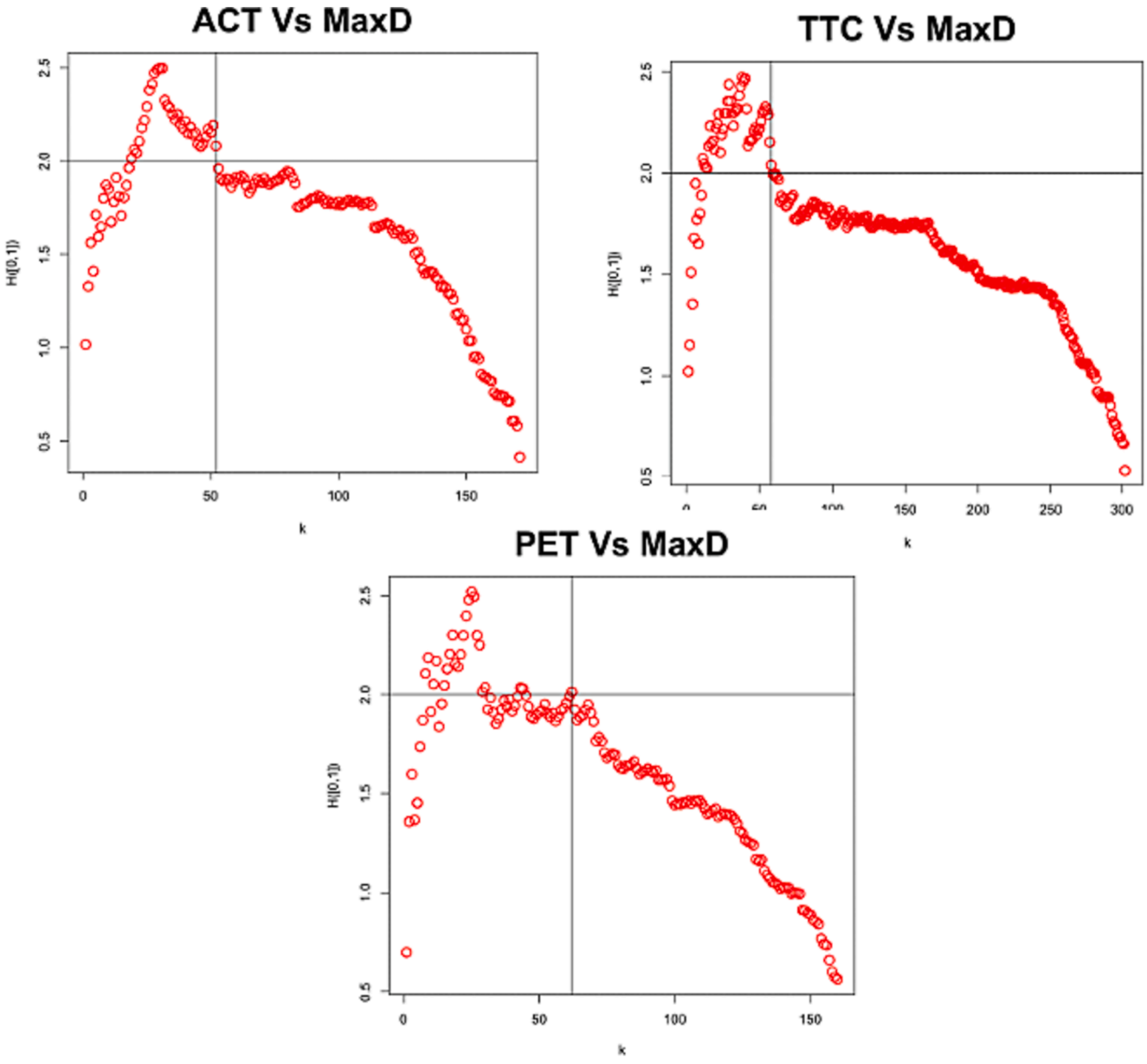

Fig. 7. Threshold selection for bivariate EVT models based on empirical spectrum value.

3.4. Fitting bivariate POT models

For the marginal distributions, two univariate GPD models were fitted using the thresholds determined for each pair. To model the dependence structure between the two margins there are several parametric families available such as Logistic, asymmetric logistic, Husler-Reiss, asymmetric negative logistic, bilogistic, negative bilogistic, and Coles-Tawn. All these parametric families were fitted on the observed extremes using censored maximum likelihood estimation method. The best model for each pair was selected based on the model AIC (Akaike Information Criteria) values. AIC provides a relative goodness of fit of the models with respect to each other thereby facilitating the selection of a suitable model; lower AIC values indicate a better fit (Wagenmakers and Farrell, 2004). Table 4 summarizes the AICs of estimated models involving different parametric distribution functions.

As shown on Table 4, Logistic model of the bivariate family performs the best to model the dependence structure between the variables providing best fit for the extreme values, indicated by lower AIC values. In some cases, other parametric models such as bilogistic, negative logistic and negative bilogistic values seem to provide better fit. However, the AIC values obtained for these models are comparable with the logistic model. Therefore, for uniformity across all study cases, the logistic model was chosen as the best model for all five intersections.

Table 5 presents the estimation results with the logistic model. The scale parameters(σ_x, σ_y), shape parameter(ξ_x, ξ_y) and dependence parameter (r) are given in the table. The dependence parameter ranges from 0.7 to 0.93 indicating an overall high dependence between the promity based and evasive action based indicators.

3.5. Evaluating crash probability

After fitting the extreme values with two GPDs at the margins and a logistic distribution to model the dependence structure of two variables, the probability densities at different levels were evaluated. Initially, the probability for each margin was calculated from the univariate GPD models fitted earlier, as:

$$CP_z = \Pr(Z \geq t) = 1 - F(t) = 1 - \left(1 + \xi\frac{(t-u)}{\sigma}\right)^{\frac{-1}{\xi}} \quad (15)$$

where, CP_z is the univariate crash probability, Z is either of the surrogate variable from the pair, u is the marginal threshold, ξ and σ are the shape and scale parameters of the fitted univariate model (margins), and t is the threshold value for crash occurrence (0 s ACT, TTC, PET and $12m/s^2$ for MaxD).

After evaluating the marginal probabilities, the joint probability was evaluated as:

$$CP_{z_1,z_2} = \Pr(Z_1 \geq t_1 \vee Z_2 \geq t_2) = 1 - G(t_1,t_2) = 1 - exp\left[-\left(\widetilde{X}^{\frac{-1}{r}} + \widetilde{Y}^{\frac{-1}{r}}\right)^r\right] \quad (16)$$

where, CP_{z_1,z_2} is the bivariate crash probability, Z_1 and Z_2 are the surrogate variable pairs, t_1 and t_2 are the thresholds for crash occurrence. The values of $\widetilde{X}$ and $\widetilde{Y}$ can be obtained using Eqs. (6) and (7) on the marginal distributions.

The crash probabilities contribute to the overall prediction of the average annual crash frequency for each intersection, since they reflect the expected number of crashes during a particular period of time. Fig. 8 presents the Bivariate excess probability density plots with probability densities and thresholds for the fitted logistic distribution at the extreme levels for each surrogate pair.

3.6. Crash frequency prediction

As mentioned earlier, the observation period at each intersection was 1.5 h, therefore the joint probabilities also reflect the crash threat at the same time period. The following equation was used to predict annual crash frequency:

$$N = \frac{CP_{z_1,z_2} * 365 * 24}{O} \quad (17)$$

where, O denotes the observation period (1.5hrs in this study), N is the predicted annual crash frequency. For each intersection, annual crash frequencies were predicted based on univariate and bivariate model probabilities. The crash prediction results for all cases (3 univariate EVT models and 3 bivariate EVT models) are presented in Table 6.

3.7. Comparison with historical crash records

Crash records of the past five years (2016–2021) were collected from the Texas Department of Transportation (TxDOT)'s crash database – Crash Records Information System (CRIS). Annual crash frequencies of all five intersections at a radius of 100 m (328ft), matching the LiDAR's detection range, were extracted to compare with the predicted crash frequencies. Then, the average number of crashes at every intersection

Table 4
AIC values of estimated parametric models.

Intersection	Surrogate_Pairs	Logistic	Asymmetric logistic	Bilogistic	Negative logistic	Negative bilogistic	Asymmetric negative logistic	Husler-Reiss
1	TTC Vs MaxD	**288.14**	291.15	289.61	288.39	289.91	288.70	288.45
	PET Vs MaxD	**279.42**	281.42	281.33	281.71	283.45	280.92	286.77
	ACT Vs MaxD	**537.68**	543.19	540.93	538.54	539.77	542.58	538.49
2	TTC Vs MaxD	**412.70**	419.62	422.03	416.85	413.39	415.20	417.24
	PET Vs MaxD	352.16	355.57	352.87	357.95	365.25	**357.96**	352.85
	ACT Vs MaxD	**628.47**	668.45	678.06	654.86	645.55	635.85	652.88
3	TTC Vs MaxD	**400.88**	409.48	405.65	401.86	406.35	415.62	429.50
	PET Vs MaxD	258.47	**257.39**	260.88	258.72	260.52	258.34	258.38
	ACT Vs MaxD	**265.27**	266.85	265.30	266.05	268.26	266.85	266.35
4	TTC Vs MaxD	**438.41**	444.03	440.81	439.15	444.57	443.56	442.79
	PET Vs MaxD	281.83	282.81	**280.36**	281.07	282.93	281.45	280.71
	ACT Vs MaxD	**453.84**	457.54	454.81	456.51	454.43	454.89	453.94
5	TTC Vs MaxD	**261.99**	265.62	266.62	263.72	269.32	267.42	263.82
	PET Vs MaxD	**253.69**	255.90	257.73	254.13	265.89	274.76	262.74
	ACT Vs MaxD	421.39	436.30	**420.66**	423.11	430.42	459.17	449.53

Table 5
Bivariate model parameters.

Intersection No.	Surrogate_Pairs (A Vs B)	Thresholds		Estimates				
		A	B	σ_x	ξ_x	σ_y	ξ_y	r
1	TTC Vs MaxD	1.2	1.73	0.615	−0.174	1.246	−0.745	0.867
	PET Vs MaxD	1.17	1.97	0.337	−0.276	2.379	−0.784	0.932
	ACT Vs MaxD	1.42	1.76	0.055	−0.380	2.045	−0.643	0.763
2	TTC Vs MaxD	1.34	1.78	0.654	−0.265	2.067	−0.735	0.728
	PET Vs MaxD	1.3	1.83	0.489	−0.21	1.587	−0.465	0.775
	ACT Vs MaxD	1.67	1.67	0.307	−0.231	2.694	−0.314	0.765
3	TTC Vs MaxD	1.28	1.89	0.846	−0.167	0.483	−0.634	0.935
	PET Vs MaxD	1.12	1.66	0.796	−0.545	1.834	−0.562	0.709
	ACT Vs MaxD	1.32	1.94	0.356	−0.569	2.985	−0.424	0.901
4	TTC Vs MaxD	1.26	1.93	0.098	−0.884	1.972	−0.53	0.797
	PET Vs MaxD	1.13	1.76	0.346	−0.659	3.438	−0.365	0.843
	ACT Vs MaxD	1.23	1.88	0.897	−0.354	3.876	−0.563	0.912
5	TTC Vs MaxD	1.14	1.87	0.407	−0.51	2.643	−0.583	0.945
	PET Vs MaxD	1.22	1.95	0.637	−0.485	1.376	−0.987	0.875
	ACT Vs MaxD	1.34	1.9	0.415	−0.595	3.013	−0.882	0.834

was calculated. Lastly, Poisson confidence intervals over observed crash frequency were generated and used to compare the performance of the surrogate indicators.

The Poisson confidence intervals can be estimated according to the following equation (Songchitruksa and Tarko, 2006):

$$\lambda : \frac{1}{2N}\chi^2_{2y_i,1-\frac{\alpha}{2}} \leq \lambda \leq \frac{1}{2N}\chi^2_{2(y_i+1),\frac{\alpha}{2}} \tag{18}$$

where, y_i is the total observed crash frequency for i^{th} location for Nyears, λ is the mean annual observed crash frequency, α is the level of significance.

Table 6 summarizes the crash frequencies predicted using different set of variables, the observed crash frequencies (AACF), and the 95% Poisson confidence interval for the observed crashes at each intersection. Furthermore, the comparison between the different surrogate pairs was conducted by using the metrics Mean absolute error (MAE) and Root mean squared error (RMSE). MAE measures the average absolute difference between the predicted and observed crashes, whereas RMSE calculates the square root of the average of squared differences between predicted vs observed crashes with greater emphasis on larger errors. Both these metrics provide a way to quantitatively assess the accuracy of the model predictions and compare the performance.

As shown on Table 6, the bivariate EVTs models seem to perform better compared to the univariate EVT models in terms of crash frequency prediction. The inclusion of evasive action-based indicator MaxD has reduced the overestimation of crash frequencies. However, the estimations of the bivariate models still exceed the upper limit of the poisson interval. Out of the bivariate models, ACT Vs MaxD has the best performance with the lowest MAE of 27.39 and RMSE of 67.56,when compared with the observed crashes.

Fig. 9 compares the crash frequency prediction performance of all three bivariate models with the observed crash frequencies and vehicle count. It can be noted that the predictions from the TTC vs MaxD model captures the relative crash threat the intersections, whereas the predictions from the other two models follow the distribution of vehicle count. For instance, the TTC vs MaxD model predicts relatively higher crashes at intersection 4 compared to the other two models despite the fact that intersection 4 has the lowest vehicle count for the study period. As evident from the observed crash frequency, this reflects the actual crash threat at the intersection. Meanwhile, the predictions from the PET vs MaxD model seems to fail both in terms of overestimation of crash frequencies and capturing the crash trend across the intersections.

4. Discussion

The comparison results of the predicted Vs observed crashes lead to three major findings. First, the bivariate EVT models perform better crash predictions compared to univariate EVT models. The better prediction performance of the bivariate models can probably be attributed to the fact that they incorporate additional information from the relationship between proximity-based and evasive-action based indicator. This finding is also in line with the result of previous studies (Zheng and Sayed, 2019; Arun et al., 2022). The second finding is that ACT performed better for assessing the proximity of near crashes in terms of time, compared to TTC and PET. The reason could be that ACT adopts a two-dimensional time proximity-based threat detection while TTC and PET are based on one-dimensional equations of motion and are more suited to certain types of conflicts (TTC for rear-end conflicts and PET for crossing conflicts). The third finding is that the crash predictions from bivariate model with TTC and MaxD better reflects the relative safety levels of the intersections, irrespective of the traffic volume. An explanation for this can be attributed to the fact that rear-end crashes are the dominant crash types at the study intersections and the TTC vs MaxD model was able to capture the rear-end conflict threats, while overestimating other conflict risks. This highlights the issues with traditional surrogate measures being suitable for specific conflict types only.

The results of the study also show that there is an overestimation of crash frequencies. As shown in Fig. 9, all the bivariate models have predicted crashes beyond the Poisson confidence intervals of the observed crashes. Likewise, Table 6 shows a similar case among the univariate models as well. There are a few possible explanations for this. One possibility is that the data collection for the study was only conducted during PM peak hours, which may not be representative of daily traffic. The crash probabilities were extrapolated from the conflicts observed during this congested period and used to predict annual crashes. The high number of conflicts observed during peak hours may have led to an overestimation of crash probabilities and, subsequently, the predicted annual crashes. Another reason for the overestimation could be underreporting of crashes, as the data used in this study only includes police-reported crashes. Previous studies have established that a significant proportion of minor or no-injury crashes go unreported (Watson et al., 2015).

This study has a few limitations that can be addressed in the future

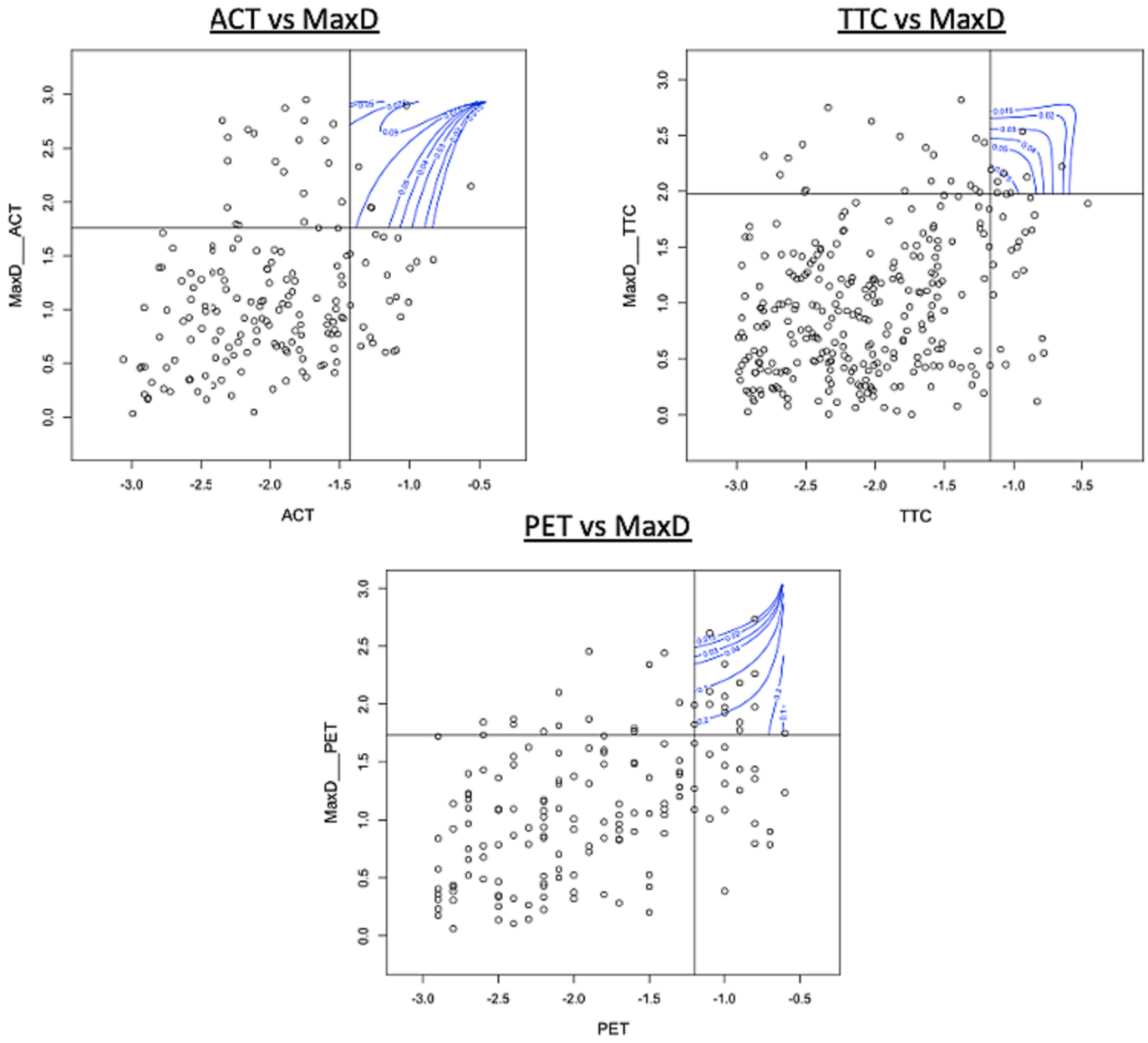

Fig. 8. Bivariate excess probability densities.

Table 6
EVT based crash predictions, Average annual crash frequency (AACF) and Poisson interval estimators.

Intersection No.	TTC	PET	ACT	TTC Vs MaxD	PET Vs MaxD	ACT Vs MaxD	AACF	Poisson Estimator	
								Lower	Upper
1	132.98	243.87	143.56	78.56	91.37	62.82	26.6	22.27	30.74
2	145.39	45.67	94.61	57.02	106.39	51.48	19.6	15.91	23.88
3	82.43	95.56	136.75	79.61	76.69	61.63	21	17.17	25.42
4	167.42	145.87	88.73	81.44	54.71	29.47	25	20.8	29.78
5	90.34	67.55	69.23	44.72	62.48	39.37	15.6	12.33	19.47
MAE	102.15	98.14	85.02	46.71	56.77	27.39			
RMSE	238.29	266.00	199.10	107.56	133.83	67.56			

research. First, the accuracy of extracted Lidar based trajectories at low resolutions is still not up to par. Especially, the instantaneous acceleration and deceleration values are noisy and unreliable, which led the authors to detect evasive actions using values aggregated at higher resolution. Further improvements in Lidar based detection and tracking algorithms are required to fully leverage the high-resolution microscopic trajectory information towards developing more reliable CPMs. The authors are currently working on comparing the performance of lidar-based trajectories across video-camera based trajectories in terms of data accuracy and computation speed. A future development might be integrating multiple data sources, such as on-board GPS, video cameras, and Lidar sensors, to collect vehicle trajectories. Likewise,

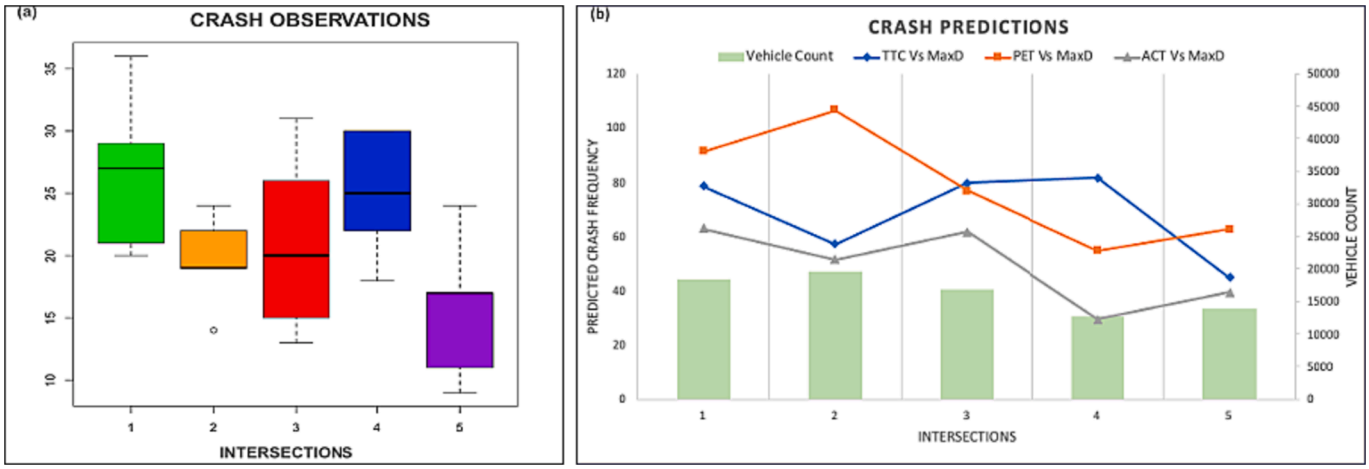

Fig. 9. Comparison of Observed and Predicted crashes.

integrating multiple Lidar sensors to collect trajectories from a single traffic facility can also be explored to improve data accuracy and solve occlusion issues. Another limitation of this study is that crash severity based surrogate measures were not tested in this study. Future studies could consider using multivariate EVT models to predict crashes frequencies at different severity levels. Likewise, all conflict types were analyzed together in this study. Future research could examine the suitability of different surrogate measures for different conflict types by using different models for each conflict type.

Overall, the methodology proposed in this study demonstrates the prospects of using roadside lidar trajectory data for near crash-based crash frequency prediction at signalized intersections. Using near crashes as crash surrogates can specially benefit the safety analysis of new roadway facilities that have not incurred any crashes till date. Although the methodology, in its current form, may not be suitable for practical implementation, it lays ground for the proactive safety analysis of any traffic facility leveraging the cost-effective emerging Lidar sensor technology. Some areas of potential applications of this study in the future could be near crash-based hotspot identification and before/after safety analysis of a facility after safety improvements.

5. Conclusion

This study developed a methodology to predict crash frequencies at signalized intersections using roadside Lidar based vehicle trajectory data. The extreme value theory was used to develop crash prediction models from near crashes identified using time-proximity and evasive action based surrogate measures such as TTC, PET, ACT and MaxD. The methodology was tested at five signalized intersections of Lubbock, Texas for crash frequency prediction. The bivariate EVT model using the surrogate pair of ACT and MaxD had the best results when compared to the actual number of crashes observed over the past five years. On the other hand, despite the considerable overestimation of crash frequencies, the model with TTC and MaxD was able to portray the relative level of safety among different intersections. The findings from this study demonstrate that the surrogate safety measures evaluated from roadside lidar based road-user trajectories at signalized intersections can help capture near-crashes, that can be used to predict crash frequencies at signalized intersections. The uniqueness of this research lies in the conjunctive application of proximity and evasive action based SSMs, extracted from high-resolution microscopic trajectory data, to predict crash probabilities using extreme value measures. By harnessing the roadside Lidar technology, we introduce a novel approach that surpasses traditional methods reliant on historical crash data alone. With further improvement to the current methodology in terms of including other surrogate measures and conflict-type based analyses, near-crashes can be used as a reliable alternative to actual crashes to foster a more comprehensive understanding of traffic safety dynamics at signalized intersections. Likewise, for future studies, larger observation periods must be considered, and the number of study intersections should be increased to gather enough samples for further validation. Regardless, this study lays the ground for using the emerging roadside Lidar technology as a cost-effective and easy-to-conduct approach for proactive safety analysis of signalized intersections while also highlighting the need for traffic safety researchers to explore the possibilities for incorporating near-crash-based crash prediction models as a part of the data-driven proactive safety analysis approaches.

CRediT authorship contribution statement

Nischal Bhattarai: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Yibin Zhang:** Validation. **Hongchao Liu:** Conceptualization, Supervision, Writing – review & editing. **Hao Xu:** Supervision, Software.

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.

References

Allen, Brian L, B Tom Shin, Peter J Cooper. 1978 Analysis of traffic conflicts and collisions. 0361-1981.

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

Arun, A., Haque, M.M., Washington, S., Sayed, T., Mannering, F., 2021b. 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.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: emerging technologies 138, 103653.

Bhattarai, Nischal, Yibin Zhang, Hongchao Liu, Yaser Pakzad, and Hao Xu. 2023 Proactive safety analysis using roadside lidar based vehicle trajectory data: A study on rear-end crashes. Transportation Research Record.

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 & Prevention 159, 106274.

Campbell, S., O'Mahony, N., Krpalcova, L., Riordan, D., Walsh, J., Murphy, A., Ryan, C. (Eds.), 2018. 2018 Sensor Technology in Autonomous Vehicles: A Review. 2018 29th Irish SignAls And Systems Conference (ISSC). IEEE, pp. 1–4.
Cavadas, J., Azevedo, C.L., Farah, H., Ferreira, A., 2020. Road safety of passing maneuvers: a bivariate extreme value theory approach under non-stationary conditions. Accident Analysis & Prevention 134, 105315.
Coles, Stuart, Joanna Bawa, Lesley Trenner, Pat Dorazio. 2001. An introduction to statistical modeling of extreme values. Volume 208. Springer.
Dimitriou, L., Stylianou, K., Abdel-Aty, M.A., 2018. Assessing rear-end crash potential in urban locations based on vehicle-by-vehicle interactions, geometric characteristics and operational conditions. Accident Analysis & Prevention 118, 221–235.
Dutfoy, A., Parey, S., Roche, N., 2014. Multivariate extreme value theory-A tutorial with applications to hydrology and meteorology. Dependence Modeling 2 (1).
Embrechts, P., Resnick, S.I., Samorodnitsky, G., 1999. Extreme value theory as a risk management tool. North American Actuarial Journal 3 (2), 30–41.
Essa, M., Sayed, T., 2018. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. Transportation research part C: emerging technologies 89, 289–302.
Farah, H., Azevedo, C.L., 2017. Safety analysis of passing maneuvers using extreme value theory. IATSS research 41 (1), 12–21.
Gettman, Douglas, Lili Pu, Tarek Sayed, Steven G Shelby, and Siemens Energy. 2008 Surrogate safety assessment model and validation. Turner-Fairbank Highway Research Center.
Göhring, Daniel, Miao Wang, Michael Schnürmacher, and Tinosch Ganjineh. 2011 Radar/lidar sensor fusion for car-following on highways. The 5th International Conference on Automation, Robotics and Applications. 2011. 407-412. IEEE.
Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transportation Research Record 2147 (1), 66–74.
Hayward J. 1971 Near Misses as a results of Safety at Urban Intersections, Department of Civil Engineering, Pensylvania State University, University Park, PA.
Hayward, John C. 1972. Near miss determination through use of a scale of danger.
Johnsson, C., Laureshyn, A., Dágostino, C., 2021. A relative approach to the validation of surrogate measures of safety. Accident Analysis & Prevention 161, 106350.
Khan, M.A., Ectors, W., Bellemans, T., Ruichek, Y., Yasar, A.-u.-H., Janssens, D., Wets, G., 2018. Unmanned aerial vehicle-based traffic analysis: A case study to analyze traffic streams at urban roundabouts. Procedia computer science 130, 636–643.
Klauer, Charlie, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, and David J Ramsey. 2006 The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.
Krajewski, Robert, Julian Bock, Laurent Kloeker, and Lutz Eckstein. 2018 The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). 2018. 2118-2125. IEEE.
Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis & Prevention 102, 213–226.
Lv, B., Sun, R., Zhang, H., Hao, X.u., Yue, R., 2019. Automatic vehicle-pedestrian conflict identification with trajectories of road users extracted from roadside lidar sensors using a rule-based method. IEEE Access 7, 161594–161606.
Machiani, S.G., Abbas, M., 2016. Safety surrogate histograms (SSH): A novel real-time safety assessment of dilemma zone related conflicts at signalized intersections. Accident Analysis & Prevention 96, 361–370.
Mahmud, SM.S., Luis Ferreira, M.d., Hoque, S., Tavassoli, A., 2017. Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS research 41 (4), 153–163.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research 11, 1–16.
McNeil, A.J., 1999. Extreme value theory for risk managers. Departement Mathematik ETH Zentrum 12 (5), 217–237.
Nadimi, N., Ragland, D.R., Amiri, A.M., 2020. An evaluation of time-to-collision as a surrogate safety measure and a proposal of a new method for its application in safety analysis. Transportation letters 12 (7), 491–500.
Orsini, F., Gecchele, G., Gastaldi, M., Rossi, R., 2019. Collision prediction in roundabouts: a comparative study of extreme value theory approaches. Transportmetrica A: transport science 15 (2), 556–572.
Peesapati, L.N., Hunter, M.P., Rodgers, M.O., 2013. Evaluation of postencroachment time as surrogate for opposing left-turn crashes. Transportation research record 2386 (1), 42–51.
Pei, X., Wong, S.C., Sze, N.-N., 2011. A joint-probability approach to crash prediction models. Accident Analysis & Prevention 43 (3), 1160–1166.
Songchitruksa, P., Tarko, A.P., 2006. The extreme value theory approach to safety estimation. Accident Analysis & Prevention 38 (4), 811–822.
Sun, Y., Hao, X.u., Jianqing, W.u., Zheng, J., Dietrich, K.M., 2018. 3-D data processing to extract vehicle trajectories from roadside LiDAR data. Transportation research record 2672 (45), 14–22.
Sun, L., Zhang, D., Chen, C., Castro, P.S., Li, S., Wang, Z., 2013. Real time anomalous trajectory detection and analysis. Mobile Networks and Applications 18 (3), 341–356.
Tageldin, A., Sayed, T., 2016. Developing evasive action-based indicators for identifying pedestrian conflicts in less organized traffic environments. Journal of Advanced Transportation 50 (6), 1193–1208.
Tageldin, A., Sayed, T., Ismail, K., 2018. Evaluating the safety and operational impacts of left-turn bay extension at signalized intersections using automated video analysis. Accident Analysis & Prevention 120, 13–27.
van Haperen, W., Daniels, S., De Ceunynck, T., Saunier, N., Brijs, T., Wets, G., 2018. Yielding behavior and traffic conflicts at cyclist crossing facilities on channelized right-turn lanes. Transportation research part F: traffic psychology and behaviour 55, 272–281.
Vasconcelos, L., Neto, L., Seco, Á.M., Silva, A.B., 2014. Validation of the surrogate safety assessment model for assessment of intersection safety. Transportation Research Record 2432 (1), 1–9.
Venthuruthiyil, S.P., Chunchu, M., 2022. Anticipated Collision Time (ACT): A two-dimensional surrogate safety indicator for trajectory-based proactive safety assessment. Transportation research part C: emerging technologies 139, 103655.
Wagenmakers, E.-J., Farrell, S., 2004. AIC model selection using Akaike weights. Psychonomic bulletin & review 11 (1), 192–196.
Wali, B., Khattak, A.J., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Analytic methods in accident research 28, 100136.
Wang, C., Chengcheng, X.u., Xia, J., Qian, Z., Linjun, L.u., 2018. A combined use of microscopic traffic simulation and extreme value methods for traffic safety evaluation. Transportation Research Part C: Emerging Technologies 90, 281–291.
Wang, C., Liu, L., Chengcheng, X.u., Lv, W., 2019a. Predicting future driving risk of crash-involved drivers based on a systematic machine learning framework. International journal of environmental research and public health 16 (3), 334.
Wang, C., Chengcheng, X.u., Dai, Y., 2019b. A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. Accident Analysis & Prevention 123, 365–373.
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 & Prevention 157, 106157.
Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. Accident Analysis & Prevention 83, 18–25.
Wong, W., Shen, S., Zhao, Y., Liu, H.X., 2019. On the estimation of connected vehicle penetration rate based on single-source connected vehicle data. Transportation Research Part B: Methodological 126, 169–191.
Wu, Jianqing, Hao Xu, Jianying Zheng. 2017 Automatic background filtering and lane identification with roadside LiDAR data. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). 2017. 1-6. IEEE.
Wu, J., Hao, X.u., Zheng, Y., Tian, Z., 2018. A novel method of vehicle-pedestrian near-crash identification with roadside LiDAR data. Accident Analysis & Prevention 121, 238–249.
Wu, J., Hao, X.u., Zheng, Y., Zhang, Y., Lv, B., Tian, Z., 2019. Automatic Vehicle Classification using Roadside LiDAR Data. Transportation Research Record 2673 (6), 153–164.
Wu, J., Hao, X.u., Zhang, Y., Tian, Y., Song, X., 2020a. Real-time queue length detection with roadside LiDAR data. Sensors 20 (8), 2342.
Wu, J., Hao, X.u., Zheng, J., Zhao, J., 2020b. Automatic vehicle detection with roadside LiDAR data under rainy and snowy conditions. IEEE Intelligent Transportation Systems Magazine 13 (1), 197–209.
Yang, D.i., Xie, K., Ozbay, K., Yang, H., 2021. Fusing crash data and surrogate safety measures for safety assessment: Development of a structural equation model with conditional autoregressive spatial effect and random parameters. Accident Analysis & Prevention 152, 105971.
Zhang, Y., Bhattarai, N., Zhao, J., Liu, H., Xu, H., 2022. An Unsupervised Clustering Method for Processing Roadside LiDAR Data with Improved Computational Efficiency. IEEE Sensors Journal 22 (11), 10684–10691.
Zhang, Q.i., Bhattarai, N., Chen, H., Hao, X.u., Liu, H., 2023. Vehicle Trajectory Tracking Using Adaptive Kalman Filter from Roadside Lidar. Journal of Transportation Engineering, Part A: Systems 149 (6), 04023043.
Zhang, J., Xiao, W., Coifman, B., Mills, J.P., 2020. Vehicle tracking and speed estimation from roadside lidar. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 5597–5608.
Zhang, X., Yan, X., 2023. Predicting collision cases at unsignalized intersections using EEG metrics and driving simulator platform. Accident Analysis & Prevention 180, 106910.
Zhao, J., Hao, X.u., Liu, H., Jianqing, W.u., Zheng, Y., Dayong, W.u., 2019. Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors. Transportation research part C: emerging technologies 100, 68–87.
Zhao, J., Hao, X.u., Zhang, Y., Shankar, V., Liu, H., 2022. Automatic identification of vehicle partial occlusion in data collected by roadside LiDAR sensors. Transportation research record 2676 (5), 708–718.
Zheng, L., Sayed, T., 2019. From univariate to bivariate extreme value models: approaches to integrate traffic conflict indicators for crash estimation. Transportation research part C: emerging technologies 103, 211–225.
Zheng, L., Ismail, K., Meng, X., 2014a. Freeway safety estimation using extreme value theory approaches: A comparative study. Accident Analysis & Prevention 62, 32–41.
Zheng, L., Ismail, K., Meng, X., 2014b. Traffic conflict techniques for road safety analysis: open questions and some insights. Canadian journal of civil engineering 41 (7), 633–641.
Zheng, L., Sayed, T., Tageldin, A., 2018. Before-after safety analysis using extreme value theory: a case of left-turn bay extension. Accident Analysis & Prevention 121, 258–267.