CHALMERS
UNIVERSITY OF TECHNOLOGY

Modeling Lead-Vehicle Kinematics for Rear-End Crash Scenario Generation

Citation for the original published paper (version of record):
Wu, J., Flannagan, C., Sander, U. et al (2024). Modeling Lead-Vehicle Kinematics for Rear-End Crash Scenario Generation. IEEE Transactions on Intelligent Transportation Systems, In Press. http://dx.doi.org/10.1109/TITS.2024.3369097

Modeling Lead-Vehicle Kinematics for Rear-End Crash Scenario Generation

Jian Wu, Carol Flannagan, Ulrich Sander, and Jonas Bärgman

***Abstract*— The use of virtual safety assessment as the primary method for evaluating vehicle safety technologies has emphasized the importance of crash scenario generation. One of the most common crash types is the rear-end crash, which involves a lead vehicle and a following vehicle. Most studies have focused on the following vehicle, assuming that the lead vehicle maintains a constant acceleration/deceleration before the crash. However, there is no evidence for this premise in the literature. This study aims to address this knowledge gap by thoroughly analyzing and modeling the lead vehicle's behavior as a first step in generating rear-end crash scenarios. Accordingly, the study employed a piecewise linear model to parameterize the speed profiles of lead vehicles, utilizing two rear-end pre-crash/near-crash datasets. These datasets were merged and categorized into multiple sub-datasets; for each one, a multivariate distribution was constructed to represent the corresponding parameters. Subsequently, a synthetic dataset was generated using these distribution models and validated by comparison with the original combined dataset. The results highlight diverse lead-vehicle speed patterns, indicating that a more accurate model, such as the proposed piecewise linear model, is required instead of the conventional constant acceleration/deceleration model. Crashes generated with the proposed models accurately match crash data across the full severity range, surpassing existing lead-vehicle kinematics models in both severity range and accuracy. By providing more realistic speed profiles for the lead vehicle, the model developed in the study contributes to creating realistic rear-end crash scenarios and reconstructing real-life crashes.**

***Index Terms*— Rear-end crash, lead-vehicle kinematics, data combination, multivariate distribution modeling, data synthesis, virtual safety assessment.**

I. Introduction

VIRTUAL safety assessment has emerged as the primary approach for evaluating the safety of Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS) due to its cost-effectiveness and efficiency compared to traditional field testing [1], [2], [3], [4]. The two main approaches to such assessment are traffic-simulation-based [5], [6], [7] and in-depth-crash-data-based (referred to as IDC-based) [8], [9], [10].

The traffic-simulation-based approach simulates daily driving in order to create crash events in a virtual naturalistic driving environment [5], [6], [7]. Typically, traffic simulation models are built using naturalistic driving data (NDD), which includes few crashes, and those captured are typically of low severity. For safety evaluation, simulations are often conducted over an extended period (measured in millions of simulation hours) using the subject vehicle (i.e., the vehicle for which the system is assessed) both with and without the specific ADAS or ADS. The number of crashes experienced in each situation is subsequently compared.

This approach has three primary challenges. First, it is very inefficient; due to the high dimensionality of the environment and the rareness of safety-critical events, demonstrating the safety performance of autonomous vehicles requires hundreds of millions of miles [7]. To tackle this problem, Feng et al. [7] proposed a solution known as the naturalistic and adversarial driving environment (NADE), which introduces sparse but adversarial modifications in order to reduce the number of virtual test miles needed while maintaining unbiased evaluations. However, even with the NADE technique, a substantial number of test miles is still necessary. Second, utilizing NDD as the initial condition for generating crash scenarios may lead to stark differences in crash characteristics compared to real-world crashes, both at the individual level and in terms of their overall distribution. Olleja et al. [11] compared crash generation methods using normal driving data and near-crash incidents with crashes obtained from in-depth crash databases. The results showed substantial disparities: normal driving data failed to reflect the crash outcomes and criticality observed in crashes from in-depth crash databases. Third and finally, crashes generated by the traffic-simulation-based approach rely heavily on accurate models of road-user behaviors that can produce realistic crashes (representative of real-world scenarios). However, validation of the details of the generated crashes is infrequent.

In contrast to the traffic-simulation-based approach, the IDC-based approach uses in-depth crash data containing reconstructed (and sometimes, although much more rarely, recorded) information such as vehicle kinematics to generate virtual crashes, either directly (by constructing digital twins for individual crashes) or indirectly (by sampling from distributions of relevant crash characteristics). A simulation

Manuscript received 27 March 2023; revised 25 August 2023, 31 October 2023, and 8 January 2024; accepted 16 February 2024. This work was supported by the Fordonsstrategisk forskning och innovation (FFT) Program sponsored by Vinnova, the Swedish Governmental Agency for Innovation, as part of the Project Improved Quantitative Driver Behavior Models and Safety Assessment Methods for Advanced Driver Assistance Systems (ADASs) and Automated Driving (AD) (QUADRIS) under Grant 2020-05156. The Associate Editor for this article was C. Lv. *(Corresponding author: Jian Wu.)*

Jian Wu is with the Volvo Cars Safety Center, 41878 Gothenburg, Sweden, and also with the Department of Mechanics and Maritime Sciences, Chalmers University of Technology, 41756 Gothenburg, Sweden (e-mail: jian.wu.2@volvocars.com).

Carol Flannagan is with the University of Michigan Transportation Research Institute (UMTRI), Ann Arbor, MI 48109 USA, and also with the Department of Mechanics and Maritime Sciences, Chalmers University, 41756 Gothenburg, Sweden (e-mail: cacf@umich.edu).

Ulrich Sander is with the Volvo Cars Safety Center, 41878 Gothenburg, Sweden (e-mail: ulrich.sander@volvocars.com).

Jonas Bärgman is with the Department of Mechanics and Maritime Sciences, Chalmers University of Technology, 41756 Gothenburg, Sweden (e-mail: jonas.bargman@chalmers.se).

Digital Object Identifier 10.1109/TITS.2024.3369097

with the ADAS or ADS [4], [8], [9], [12], [13] is then run for each generated crash to answer the question, "What would happen if vehicle X were equipped with technology Y?"

The IDC-based approach, however, also has its challenges. First, the safety assessments of ADAS and ADS typically require more real-world crash instances than what is currently accessible. Furthermore, the limited availability of real-world crashes with in-depth information hampers the representation of the diverse range of crashes within specific scenarios. Consequently, "synthetic crashes", which can be viewed as variations of the original crashes, must be generated to fill in the gaps between real crashes [10]. Second, the selection criteria used in traditional in-depth crash databases inherently introduce a bias toward severe crashes. Relying solely on these databases to create synthetic crashes [10], [14], [15] skews the crash generation models, potentially distorting the overall analysis. A third issue relates to how crashes are generated and their representativeness in the real world. Using reconstructions to generate crashes can be problematic as it involves making assumptions about individual road users, such as their braking profiles, without relying on detailed pre-crash recordings. While the crash outcome, such as the change in speed during the crash, may be reasonably accurate, the pre-crash kinematics is influenced by the decisions made during reconstruction and by the reconstruction software itself, leading to models based on assumptions and software rather than on explicit descriptions of the pre-crash kinematics of real crashes. In one such model, Gambi et al. [14] proposed a model to efficiently generate crash scenarios by extracting crash information from police reports using Natural Language Processing (NLP) techniques. However, this model mainly relies on information from police reports and considers basic kinematics, overlooking driver behavior, leaving uncertainty about how accurately crashes generated from this "simplistic kinematics" method reflect the safety benefits of the assessed systems. A fourth issue is encountered when generating crashes at the tails of distributions. For example, Wang et al. [15] demonstrated that using independent component analysis (ICA) followed by kernel density estimation (KDE) to generate synthetic crashes can introduce biases, particularly near boundaries and in distributions with long tails [16]. Overall, methodological choices significantly impact the accuracy and generalizability of the crash scenarios produced for both scenario generation approaches.

To address the limitations of both approaches, a novel method combining data from naturalistic driving and recorded pre-crash kinematics from in-depth crash databases is proposed. This combined dataset covers the full severity range (from low to high severity levels) and aids in developing crash-generation models applicable to both traffic-simulation-based and IDC-based approaches. This study focuses specifically on rear-end crash scenario generation as an initial step in demonstrating the proposed method.

A rear-end crash, in which the front of one vehicle collides with the rear of another, is a common crash type. In the United States, for example, rear-end crashes accounted for 27.8 percent of all car crashes in 2020 [17]. Hence, studying rear-end crash scenario generation is essential. Moreover, the rear-end crash is relatively simple since it refers mainly to longitudinal maneuvers, and only two vehicles (the lead and the following vehicles) are involved. Consequently, synthetic rear-end crashes can be created based on models of the two involved vehicles, in which the lead vehicle is independent of the following vehicle, and the following vehicle responds to the presence and actions of the lead vehicle.

To address these issues, we propose a novel approach to rear-end scenario generation, which combines the lead-vehicle kinematics model and the following vehicle behavior model to obtain representative rear-end crash scenarios across the full severity range. The work presented in this paper addresses only the first step of this crash scenario generation approach. Many other studies have analyzed the following vehicle's behavior during rear-end emergencies (crashes and near-crashes) by means of a driver response model [18], [19], [20], [21], [22], [23], [24], [25]. For example, Markkula et al. [23] used a piecewise linear model and used driver glance behaviors to model the following vehicles' speed profiles in naturalistic rear-end emergencies. They found that braking typically started less than a second after the kinematic urgency reached certain threshold levels; faster reactions occurred at higher urgencies.

However, there has been a notable lack of research on the lead vehicle's behavior in such situations despite its significant influence on the following vehicle. In crash reconstruction and rear-end emergency studies, it is commonly assumed that the lead vehicle maintains a constant acceleration or deceleration before the crash [19], [21], [24], even though there is inadequate evidence to support this assumption. To address this knowledge gap, the objective of this study is to develop a model of the lead-vehicle kinematics in rear-end crashes across the full severity range as the first step in generating rear-end crash scenarios. Future work will use models of the following vehicle's behavior together with the lead-vehicle kinematics model from this work to generate rear-end crash scenarios.

In this study, a piecewise linear model was employed to parameterize the speed profiles of the lead vehicles in two established datasets of rear-end pre-crash/near-crash incidents. These datasets consist of recorded vehicle kinematics prior to the actual crash or near-crash events. The two datasets were combined to create a comprehensive dataset that spans the full range of severity, which was then categorized into multiple sub-datasets based primarily on lead-vehicle speed change patterns. Next, a multivariate distribution model of the parameters was built for each sub-dataset. Finally, a synthetic dataset was generated by sampling the synthetic lead vehicles' speed profiles created by the distribution models in proportion to the sample size of each sub-dataset. The synthetic dataset was then validated by comparing the parameter distributions and kinematics of the generated crashes with the original combined dataset.

II. Datasets

This study focuses on passenger vehicle rear-end crashes/near-crashes, and the data come from two sources: the Crash Investigation Sampling System (CISS) and the Second

Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS).

CISS is a nationally representative complex probability sample of general passenger vehicle crashes in the United States in which at least one light vehicle was towed away [26], [27]. The data collection started in 2017 and is still ongoing. The data are from in-depth crash investigations that include inspection of damaged vehicles and crash sites, as well as estimation of crash kinematics. In addition, Event Data Recorder (EDR) data were extracted and included in the sample whenever possible.

In the SHRP2 NDS, over 3,300 passenger vehicles were instrumented with a data acquisition system (DAS) that collected four video views (driver's face, driver's hands, forward roadway, and rear roadway) and information from vehicle networks and sensors [28]. Naturalistic driving data from six sites around the United States were collected through the participant vehicles between 2010 and 2013. Unlike the CISS dataset, which only contains crashes, the SHRP2 dataset contains both crashes and near-crashes. A dozen trigger algorithms were executed on collected trip files followed by manual annotation to identify crashes and near-crashes (defined as any circumstances that require a rapid evasive maneuver by the subject vehicle or any other vehicle, pedestrian, cyclist, or animal to avoid a crash [28]). It is worth mentioning that there is no overlap between the CISS and SHRP2 datasets since the data were collected at different times.

A. Data Selection

Rear-end pre-crash/near-crash data for incidents in which the subject vehicle that collected the data was the lead vehicle (struck vehicle) were extracted from both datasets. The subject vehicle's speed v (unit: m/s) was the only signal we used; the signal was directly estimated from the wheel speed or in-vehicle inertial sensor.

In the CISS dataset, only rear-end crashes in which the struck vehicle was equipped with an event data recorder containing recorded data from the pre-crash phase were extracted. Among these cases, the ones with a data frequency of no less than 5 Hz were selected for this study. Of the selected 52 CISS crashes, three have a frequency of 5 Hz, while the rest have a frequency of 10 Hz.

In the SHRP2 dataset, incidents (crashes and near-crashes) labeled "Rear-end, struck" were selected. The frequency of all the SHRP2 data used is 10 Hz.

The code for extracting cases from CISS and SHRP2 datasets is published online [29].

B. Data Groups

In the study, the crashes extracted from the SHRP2 dataset were originally labeled according to severity level: I (most severe), II (police-reportable), and III (minor). Note that level IV (low-risk tire strikes) is not included. A crash that involves an airbag deployment, injury to the driver, pedal cyclist or pedestrian, vehicle rollover, high Delta V, or requires vehicle towing is classified as severity level I. A level II crash is any police-reportable crash that does not meet the level I crash requirements. All other crashes that involve physical contact with minimal damage are considered level III crashes [31]. In this study, a crash is indexed as 'Severe' if it fulfills the SHRP2 severity level I definition and 'Non-severe' otherwise. In addition, the severity level of any near-crash is designated as 'None'.

TABLE I
ALL EXTRACTED EVENTS

Group	Notation	Source	Severity level[a]	Sample size[b]
1	CISS_sc	CISS	Severe	49/52
2	SHRP2_sc	SHRP2	Severe	20/24
3	SHRP2_nsc	SHRP2	Non-severe	63/106
4	SHRP2_nc	SHRP2	None	171/272

[a] The severity level here does not correspond with the Abbreviated Injury Scale (AIS) [30]. A crash is indexed as severe if it fulfills the 'SHRP2 severity level I' definition [31]; otherwise, non-severe. Furthermore, the severity level for a near-crash incident is designated as 'None'.
[b] Valid sample size/raw sample size.

The data were separated into four groups according to source and severity level, as Table I shows. Group 1, CISS_sc, comprises the extracted CISS crashes. They fulfill the SHRP2 severity level I definition and are, therefore, considered severe crashes. Extracted SHRP2 severity level I crashes belong to Group 2, SHRP2_sc. The SHRP2 crashes at severity levels II and III make up Group 3, SHRP2_nsc (non-severe crashes). Group 4, SHRP2_nc, consists of SHRP2 near-crashes. The raw sample sizes of Groups 1-4 are 52, 24, 106, and 272, respectively. However, not all samples are valid; the conditions for selecting valid samples are introduced in the following subsection.

C. Event Data Extraction

Time zero, for a crash, is the impact moment. For a near-crash, it is the moment when the following vehicle reaches the minimum distance to the subject vehicle. This moment was annotated manually according to the video.

CISS data typically contains five seconds before the impact, while SHRP2 data contains a longer duration. To make all events equivalent, -5 s was set as the start-point of all events. The closest data point to the impact moment was excluded to avoid a possible sharp acceleration pulse near impact. Because the lowest data frequency is 5 Hz, raw data were extracted only up to 0.3 s before impact ($t = -5$ s to $t = -0.3$ s) for each crash. For a near-crash, the extracted duration is from $t = -5$ s to $t = 0$ s.

The extracted events fulfilling the following conditions were considered valid and selected for further analysis:

- The total sample duration should be no less than three seconds (due to missing or invalid data).
- The fitted accelerations should range between -1 g and 1 g, where g is the gravitational acceleration. (This study simplifies the lead-vehicle speed profile as a sequence of straight lines. The fitted accelerations are the slopes of those lines. More details on fitted acceleration are in Section III.)

There were 49, 20, 63, and 171 valid samples for Groups 1-4, respectively (see Table I).

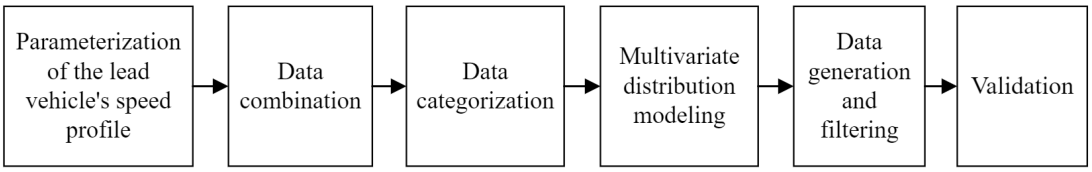


Fig. 1. Flowchart of six steps of data analysis performed in the study.

III. Methodology

The following six steps (see Fig. 1) were performed successively:

1) Parameterization of the lead-vehicle speed profiles
2) Data combination
3) Data categorization
4) Multivariate distribution modeling
5) Data generation and filtering
6) Validation

The methods used in each step will be presented in this section, with the exception of the data categorization step, which is based on the results from other steps and introduced in Section IV.

A. Parameterization of the Lead-Vehicle Speed Profiles

For each case, the lead-vehicle speed profile was fitted to a piecewise linear model, similar to that used by Markkula et al. [23]. However, the applications are different. Their work modeled the following vehicle's braking behavior (acceleration, not speed) as only one braking phase with constant deceleration. The lead vehicle's braking behavior can have multiple phases, however, making the piecewise linear model here more complicated.

1) Piecewise Linear Model: This model simplifies the lead-vehicle speed profile as several consecutive straight lines. The connection points of these lines are named breakpoints. The model contained the following steps:

Step 0: Start.

Step 1: Set sample weight to emphasize the different importance levels of different samples. The closer to time zero, the more relevant the sample is to the crash/near-crash, and thus the more important and the greater the weight. In this case, we can prioritize capturing the speed changes closer to time zero and avoid overfitting the early samples. The weight of sample i is defined as

$$w_i = (0.1 - t_i)^{-0.5}, \tag{1}$$

where w_i and t_i are the weight and time of sample i respectively. (More details about setting the sample weights are in Appendix A-A.)

Step 2: Fit the lead-vehicle speed profile using weighted piecewise linear regressions with the number of breakpoints n_b from zero to the pre-configured maximum number of breakpoints $n_{b,max}$. The weighted piecewise linear regression is based on the "piecewise-regression" package in python [32], which fits a curve as several consecutive straight lines. The option to consider sample weights has been added to the package locally.

Step 3: Select the best regression. Compute the loss L according to (2) for each regression, and choose the regression with the minimum loss.

$$L = (\epsilon + \lambda \cdot \frac{\max(v)}{\Delta v + \epsilon}) \cdot n_b - R^2, \tag{2}$$

where ϵ is a small positive value to avoid a zero denominator or a zero penalty for the number of breakpoints when $\max(v)$ is zero, λ (> 0) is a pre-configured penalty coefficient, $\Delta v (= \max(v) - \min(v))$ is the maximum speed change, and R^2 is the R-squared of the regression (indicating the fitting accuracy). The loss function penalizes excessive breakpoints to avoid overfitting, especially when $\max(v)$ is large and Δv is small. (The justification for the loss function is in Appendix A-B.)

Step 4: Modify the fitting results to avoid any negative estimated speed $\hat{v}$ (due to estimation error) in the modeling duration, -5 to 0 s (the sampled duration might be shorter than 5 seconds, yet the model can predict the speed for the missing part). There are two sub-steps:

1) Add one breakpoint in the start or end segment where $\hat{v}$ is 0 m/s if there is any negative $\hat{v}$ at the start-point or end-point of the modeling duration. Then set $\hat{v}$ to 0 m/s from the newly added breakpoint to the start-point or end-point.
2) Change $\hat{v}$ at that breakpoint to 0 m/s if there is any negative estimated speed value at any breakpoint. Then connect the modified breakpoint with other points.

Given that $\hat{v}$ is non-negative at the start-point, end-point, and all breakpoints, $\hat{v}$ should be non-negative during the whole modeling duration.

Step 5: End.

This piecewise linear model aims to fit the lead-vehicle speed profile with the simplest regression model possible. Consequently, we set $n_{b,max} = 3$, $\lambda = 0.006$, and $\epsilon = 1 \times 10^{-6}$ m/s. $n_{b,max}$ was set as the maximum number of breakpoints annotated among a small sub-dataset that was randomly sampled, while λ was set as the elbow of the curve, representing the total number of breakpoints for all events plotted against λ. (More details of the selection of pre-configured parameters are in Appendix A-C.)

2) Parameters: The piecewise linear model consists of a maximum of four consecutive lines with three breakpoints. Given the priorities of simplification and sample weights, it is unnecessary to include the segments relatively far from time zero when there are more than three segments. If the lead vehicle reaches a steady speed at the last segment (S), at most, three segments closest to time zero are selected; otherwise,

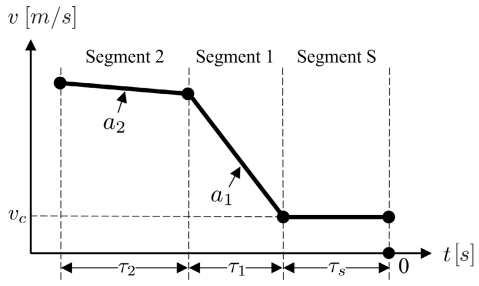


Fig. 2. Three selected segments.

at most, two segments closest to time zero are selected (i.e., segment S is removed). Fig. 2 shows an example of three selected segments. The following are explanations of each segment (backward in time from time zero), including two descriptive parameters, in the context of lead-vehicle pre-crash kinematics:

- Segment S: The lead vehicle maintains a steady speed in this segment. τ_s is the segment duration, and v_c is the lead vehicle's estimated speed at time zero.
- Segment 1: The lead vehicle keeps a non-zero constant acceleration in this segment. τ_1 is the segment duration, and a_1 is the constant acceleration.
- Segment 2: The lead vehicle keeps a constant acceleration in this segment. τ_2 is the segment duration, and a_2 is the constant acceleration.

The six-parameter vector $[v_c, a_1, a_2, \tau_s, \tau_1, \tau_2]$ is used to represent an event. It is important to note that not every event contains all three segments. Segment S and Segment 1 can exist independently, while Segment 2 can only exist when Segment 1 exists. There were five possible combinations, including their proportions: Segment S (8.9%); Segment S & 1 (8.3%); Segment S & 1 & 2 (22.8%); Segment 1 (6.9%); and Segment 1 & 2 (53.1%). The parameters of any non-existent segments are defined according to the following rules:

1) If Segment S is non-existent, $\tau_s = 0$ s.
2) If Segment 1 is non-existent, $\tau_1 = 0$ s and $a_1 = 0$ m/s^2.
3) If Segment 2 is non-existent, $\tau_2 = 0$ s and $a_2 = a_1$.

B. Data Combination

The CISS and SHRP2 datasets can be interpreted as two perspectives of crashes in the United States. By weighting the CISS and SHRP2 crashes appropriately, we can combine these two perspectives to achieve a combined dataset that describes the full range of passenger vehicle crashes (from non-severe to severe), leveraging each dataset's individual strength. For the sake of simplicity, the terms 'crash,' 'near-crash,' and 'incident' (without any further specification) refer to the lead vehicle's behavior (speed profile) for each respective type.

Assuming that severe crashes in both the CISS and SHRP2 datasets (CISS_sc and SHRP2_sc) come from the same distribution is crucial before combining the two datasets. The rationale for this assumption is the similarity between the definitions of CISS crashes (police-reported towed vehicle crashes) and severe SHRP2 crashes (SHRP2 severity level I crashes). Moreover, non-parametric tests were conducted to determine if CISS_sc and SHRP2_sc are significantly different. The results do not show any significance. (More details of the comparison are in Appendix B.)

The two datasets were combined in the following three steps: 1) pre-processing crash data, 2) reweighting crash data, and 3) adding selected near-crashes as variations of crashes. Steps 1-2 combined crashes in the CISS and SHRP2 datasets into one dataset, the combined crash dataset. Step 3 added selected SHRP2 near-crashes to the combined crash dataset, and the new dataset, including the selected near-crashes, is called the combined incident dataset. It contains the parameterized (not raw) incident and their sample weights and is published online and available to the public [29].

1) Pre-Processing Crash Data: Generally, sample weights are used so that the weighted data represent the frequency of occurrence. Before combining the crashes in the two datasets, we pre-processed the crash data so that the sample weights of the two datasets were compatible. Preprocessing tunes the existing sample weights (for the CISS dataset) or sets sample weights (for the SHRP2 dataset) so that the sum of sample weights equals the valid sample size for each respective dataset.

According to [26], CISS crashes were sampled using a probability sampling method with sampling features such as stratification, clustering, and unequal selection probabilities. Thus, the CISS sample is not a simple random sample; each CISS crash was assigned a sample weight to produce an unbiased estimation. These original sample weights range expansively from 21.8 to 3833.6. However, extreme variation in sampling weights can result in excessively large sampling variances when the data and the selection probabilities are not positively correlated [33]. A weight-trimming approach proposed in [34] was applied to reduce sampling variance. (More details of this approach are in Appendix C.) The next is to scale the trimmed weights so that the sum of the scaled weights equals the valid sample size.

$$w_{1,i} = n_{1,vld} \cdot \frac{w_{1,i,t}}{\sum_i w_{1,i,t}}, \tag{3}$$

where $w_{1,i}$ is the weight of crash i in CISS_sc after scaling, $w_{1,i,t}$ is the weight of crash i in CISS_sc after trimming, and $n_{1,vld}$ is the valid sample size of CISS_sc.

SHRP2 crashes (SHRP2_sc and SHRP2_nsc), since they occurred during the data collection period in the SHRP2 project, were not sampled but directly collected. Consequently, there are no sample weights in the SHRP2 dataset, yet all raw crashes can be seen with the same weight. The crashes in Group i ($i = 2, 3$) are assigned with the same weight w_i, which is computed as

$$w_i = (n_{2,vld} + n_{3,vld}) \cdot \frac{n_i}{n_2 + n_3} \cdot \frac{1}{n_{i,vld}}. \tag{4}$$

(As shown in Table I, among the $n_2 = 24$ and $n_3 = 106$ samples in SHRP2_sc and SHRP2_nsc, respectively, there were $n_{2,vld} = 20$ and $n_{3,vld} = 63$ valid samples, respectively.) The rationales for this sample weighting design are 1) the valid samples are selected to represent the raw samples, 2) the weighted SHRP2 data should have the same proportions

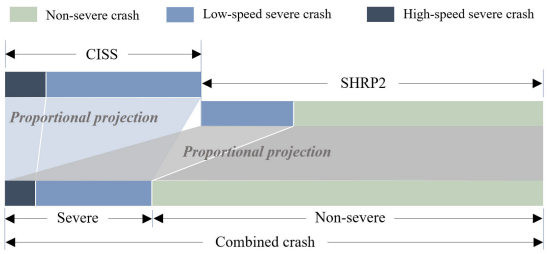


Fig. 3. Combination of crashes in the CISS and SHRP2 datasets. The dataset's icon length corresponds with its valid sample size. The combined crash dataset keeps 1) the same proportions of low-speed and high-speed severe crashes as the CISS dataset and 2) the same proportions of severe and non-severe crashes as the SHRP2 dataset.

of severe and non-severe crashes as the raw SHRP2 data, and 3) the sum of the weights equals the valid sample size $(n_{2,vld} + n_{3,vld})$.

2) Reweighting Crash Data: This step combined the CISS and SHRP2 crashes into a combined crash dataset. The combined dataset includes more information than either the CISS or SHRP2 dataset alone, but it also should retain the raw distributions. The lead vehicle's estimated speed at time zero, v_c, is the most important and representative of the six parameters. The objective of retaining the raw distributions of the original datasets can be described as two sub-objectives. For the combined crash dataset:

1) Keep the same distribution of v_c for severe crashes as the CISS dataset.
2) Keep the same proportions of severe and non-severe crashes as the SHRP2 dataset.

Compared with SHRP2_sc, CISS_sc has a wider range of v_c: SHRP2_sc [0, 7.9] m/s, CISS_sc [0, 30.4] m/s. Further, the crashes in CISS_sc and SHRP2_sc can be divided into two types:

- Low-speed: Crashes where v_c is lower than the maximum v_c in SHRP2_sc.
- High-speed: Crashes where v_c is higher than the maximum v_c in SHRP2_sc.

Per definition, SHRP2_sc contains only low-speed severe crashes, while CISS_sc contains both low-speed and high-speed severe crashes. Thus, the first sub-objective is equivalent to keeping the same proportions of low-speed and high-speed severe crashes given crashes in the CISS_sc and SHRP2_sc are from the same distribution. Fig. 3 shows the two sub-objectives of combining crashes in the CISS and SHRP2 datasets.

Consequently, this step reweighted the CISS and SHRP2 crashes and combined them as Fig. 3 shows. The proportion of non-severe crashes in SHRP2 crashes, η_{ns}, and the proportion of high-speed severe crashes in CISS, η_{hss}, are computed according to (5) and (6) respectively.

$$\eta_{ns} = \frac{n_3}{n_2 + n_3}, \tag{5}$$

$$\eta_{hss} = \frac{W_{hss}}{n_{1,vld}}, \tag{6}$$

where W_{hss} is the total weight of high-speed severe crashes before combination; it is computed as

$$W_{hss} = \sum_{i|v_{c,1,i} > max(v_{c,2})} w_{1,i}, \tag{7}$$

where $v_{c,1,i}$ is the estimated lead vehicle's speed at time zero for crash i in Group 1, CISS_sc, and $v_{c,2}$ is the estimated lead vehicle's speed values at time zero for crashes in Group 2, SHRP2_sc.

As (8) shows, the sample size of the combined crash dataset, n_{cmb}, is the sum of the valid samples of CISS_sc, SHRP2_sc and SHRP2_nsc.

$$n_{cmb} = \sum_{i=1}^{3} n_{i,vld}. \tag{8}$$

Because the combined crash dataset should retain the same η_{ns} and η_{hss}, the equivalent sample sizes can be computed as

$$n'_{ns} = n_{cmb} \cdot \eta_{ns}, \tag{9}$$

$$n'_{hss} = n_{cmb} \cdot (1 - \eta_{ns}) \cdot \eta_{hss}, \tag{10}$$

$$n'_{lss} = n_{cmb} \cdot (1 - \eta_{ns}) \cdot (1 - \eta_{hss}), \tag{11}$$

where n'_{ns} is the sample size of non-severe crashes (SHRP2_nsc) after combination, and n'_{hss} and n'_{lss} are the sample sizes of high-speed and low-speed severe crashes after combination. The weights can then be deduced according to (12-14).

$$w'_3 = \frac{n'_{ns}}{n_{3,vld}}, \tag{12}$$

$$w'_2 = n'_{lss} \cdot \frac{w_2}{W_{lss}}, \tag{13}$$

$$w'_{1,i} = \begin{cases} n'_{lss} \cdot \dfrac{w_{1,i}}{W_{lss}}, & \text{if } v_{c,1,i} \le max(v_{c,2}); \\ n'_{hss} \cdot \dfrac{w_{1,i}}{W_{hss}}, & \text{if } v_{c,1,i} > max(v_{c,2}). \end{cases}, \tag{14}$$

where W_{lss}, computed according to (15), is the total sample weight of low-speed severe crashes before combination.

$$W_{lss} = w_2 \cdot n_{2,vld} + \sum_{i|v_{c,1,i} \le max(v_{c,2})} w_{1,i}. \tag{15}$$

3) Adding Selected Near-Crashes as Variations of Crashes: After the last step, the combined crash dataset was acquired. However, the sample size of the combined crash dataset, namely the sum of the valid samples from CISS_sc, SHRP2_sc, and SHRP2_nsc, is only 132. As this is a low number for modeling distributions of the six parameters, $[v_c, a_1, a_2, \tau_s, \tau_1, \tau_2]$, we increased the number of samples by adding near-crashes similar in terms of the lead vehicle's behavior to any crash (more details in Section III-B.3). Sample weights must be adjusted so that the valid sample size (sum of sample weights) remains the same and the added near-crashes do not change the distributions in the combined crash dataset. More details of the rationale and consequences of this action will be discussed in Section V.

In the SHRP2 dataset, not all near-crashes were captured. The SHRP2 near-crashes were automatically captured by designed trigger specifications and then validated by manual annotation. In other words, those trigger specifications

contribute to a certain sampling bias of near-crashes in the SHRP2 dataset. For the rear-end near-crashes, the primary trigger specification is the "Longitudinal Deceleration," which requires the level of longitudinal acceleration to be less than or equal to -0.65 g, and the threshold is exceeded for at least one timestamp [28]. Consequently, near-crashes in which the lead vehicle does not brake harshly enough to reach that threshold are under-represented because they are not captured. Therefore, to avoid introducing bias, we cannot simply add all SHRP2 near-crashes directly to the combined crash dataset to form an even more comprehensive dataset covering all incident severities from near-crashes to severe crashes. However, lead-vehicle behaviors near-crashes in some near-crashes are similar to those observed in crashes; these near-crashes can be added to the combined crash dataset as variations of crashes. The implementation of the merging of crashes with a subset of the near-crashes was done with the following steps.

1) For near-crash i, defined by the parameter space $[v_c^i, a_1^i, a_2^i, \tau_s^i, \tau_1^i, \tau_2^i]$, find the most similar crash across the combined crash dataset using the Euclidean distance computed based on the standardized six parameters (z-score). The crash most similar to near-crash i (of all the crashes in the combined crash dataset) is defined as the one with the minimum Euclidean distance, $d_{i,min}$, and is called the 'most similar crash' of near-crash i.
2) Near-crashes with a minimum Euclidean distance less than a set threshold d_{thd} ($d_{i,min} \leq d_{thd}$) are considered similar enough to crashes to be selected for addition to the combined crash dataset.
3) For crash j, if there are in total n_{nc}^j ($n_{nc}^j \geq 0$) near-crashes whose most similar crash is crash j, the weights of crash j and those near-crashes in the combined incident dataset are set as $\frac{w^j}{1+n_{nc}^j}$, where w^j is the weight of crash j in the combined crash dataset.

The threshold d_{thd} was set as 0.78, and it was according to the analysis of the similarity between crashes in the combined crash dataset. As in the first step, we computed the minimum Euclidean distances for the lead-vehicle behaviors in crashes from the combined crash dataset. Then d_{thd} was set based on the cumulative distribution function (CDF) of the mentioned distances considering sample weights in the combined crash dataset. (More details of the choice of d_{thd} are in Appendix D.)

In the first two steps, to select near-crashes similar to crashes in the combined crash dataset, the Euclidean distance was used to measure the similarity between incidents. The last step was to adjust the sample weights to retain the raw distributions in the combined crash dataset.

C. Multivariate Distribution Modeling

The combined incident dataset was categorized into several sub-datasets (more details on this process in Section IV-C). For each sub-dataset, a multivariate distribution model was built to generate synthetic lead-vehicle speed profiles. A synthetic incident dataset was then built by sampling the generated speed profiles in proportion to the sample size of each sub-dataset.

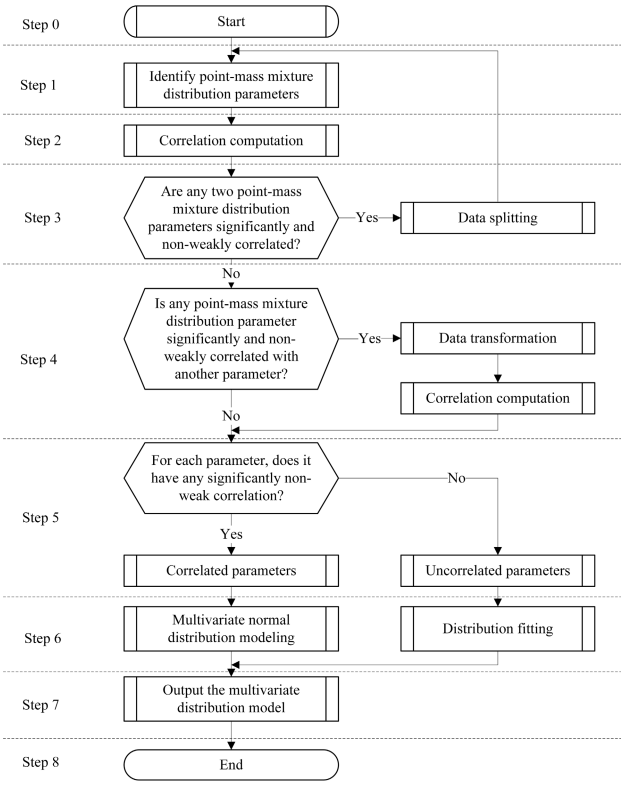


Fig. 4. The procedure of the multivariate distribution modeling.

The modeling principle here is to make things as simple as possible because a large amount of data is required to create a complicated model, while the actual amount of data available is very limited. Several simplifications were made in the modeling process and are discussed in Section V.

Two terms used in the modeling procedure are defined. A *point-mass mixture distribution parameter* contains a point-mass (a particular value with more observations than a continuous distribution can describe), which requires a mixture distribution model to describe its distribution. It is generally difficult to model the relationship between this parameter type and some other type.

A correlation coefficient is a numerical measure ranging from -1 to 1 that measures the strength and direction of a linear relationship between two quantitative variables [35]. A large significant absolute coefficient indicates a strong linear relationship between the measured variables. The sign indicates whether the relationship is positive or negative. A *significant and non-weak correlation* is defined as a correlation whose coefficient has a p-value less than 0.05 (significant) and an absolute value greater than or equal to 0.3 (non-weak).

1) Procedure: The steps in the procedure for multivariate distribution modeling (shown in Fig. 4) are listed below. More details of some steps are in the next subsections.

Step 0: Start.
Step 1: Identify point-mass mixture distribution parameters in the input data.
Step 2: Correlation computation. The correlation coefficients between every two parameters are computed. We use the "weights" package in R [36] to compute

the weighted (Pearson) correlation coefficients between all two parameter combinations, considering the sample weights.

Step 3: Check if any two point-mass mixture distribution parameters are significantly and non-weakly correlated. If so, split the data into two sub-datasets based on whether the parameter equals its point-mass value for either point-mass mixture distribution parameter; then, for each sub-dataset, go through the modeling from Step 1; otherwise, go to Step 4.

Step 4: Check if any point-mass mixture distribution parameter is significantly and non-weakly correlated with another parameter. If not, go to Step 5. Otherwise, perform data transformation followed by a correlation computation of the transformed data, then go to Step 5. The data transformation aims to decorrelate the point-mass mixture distribution parameter from any other parameter so that it can be modeled independently.

Step 5: Classify each parameter as correlated or uncorrelated. A parameter is categorized as a correlated parameter if it exhibits a significant and strong correlation with any other parameter and as uncorrelated if no such correlation is observed.

Step 6: Model the distribution. For correlated parameters, a multivariate normal distribution model is used (Multivariate normal distribution modeling in Fig. 4). The parameters are fitted to their own distributions separately (Distribution fitting in Fig. 4).

Step 7: Output the multivariate distribution model for the input data.

Step 8: End.

2) Data Splitting: Step 3 performed data splitting if two point-mass mixture distribution parameters have a significant and non-weak correlation. In this way, we built two sub-models for the two sub-datasets instead of building a complicated model for two correlated point-mass mixture distribution parameters. This simplification was performed because it is difficult to model the correlation between two point-mass mixture distribution parameters. In this case, the decorrelation method utilized in Step 4 is ineffective when dealing with the correlation between a regular parameter and a point-mass mixture distribution parameter. Because the point-mass of the point-mass mixture distribution parameter used as the independent variable cannot be produced in the generated data. However, future work should aim to incorporate this modeling aspect.

3) Data Transformation: The purpose was to ensure that no parameter is correlated with any point-mass mixture distribution parameter after transformation. For a parameter, x_i, that is significantly and non-weakly correlated with any point-mass mixture distribution parameter, the transformed parameter x_i' was computed according to

$$x_i' = x_i - f(X_{pm}), \tag{16}$$

where $f(X_{pm})$ is the estimated linear regression model, in which x_i and the vector of all point-mass mixture distribution parameters X_{pm} are the explanatory and dependent variables respectively. Consequently, x_i' is not correlated with any point-mass mixture distribution parameter.

In contrast, for a parameter that is neither significantly nor non-weakly correlated with any point-mass mixture distribution parameter, it is unchanged after transformation.

4) Distribution Fitting: The data for a parameter, which does not include a point mass, was fitted into a set of distributions, including normal, skew-normal, exponential-normal, and gamma distributions, using the maximum likelihood estimation (MLE). Akaike information criterion (AIC) was used to select the best-fitting distribution with the lowest AIC value. AIC is an estimator of prediction error and, thereby, the relative quality of statistical models for a given set of data [37].

For a point-mass mixture distribution parameter x_j, a mixture distribution model combining a binomial distribution and a continuous distribution was used. The mixture distribution model, in this case, is a hurdle model [38], in that all the point-mass values come only from the binomial distribution, while all other values come from the continuous distribution. Therefore, the two sub-distributions were fitted separately using MLE:

1) The binomial distribution's estimated success possibility is the point-mass value proportion.
2) For the continuous distribution model, select successively from the gamma, generalized gamma, and exponential distribution models using AIC.

5) Multivariate Normal Distribution Modeling: This process modeled all n correlated parameters as a multivariate normal distribution according to the following steps.

1) Fit the distribution for all correlated parameters with the same method used in the sub-step, distribution fitting.
2) Use the quantile transformation (also known as quantile mapping) [39] to transform the data to the standard normal distribution $\mathcal{N}(0, 1)$ for each parameter.
3) Compute the covariance matrix $\mathbf{\Sigma} \in \mathbf{R}^{n\times n}$ of the normalized parameters and set $\mathbf{\Sigma}_{ii} = 1$ $(i = 1, \ldots, n)$.
4) Build the multivariate normal distribution $\mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$, where $\mathbf{M} = (0, \ldots, 0)^T$.

The $\mathbf{\Sigma}_{ii}$ and $\mathbf{M}$ were set directly because the normalized parameters all belong to $\mathcal{N}(0, 1)$.

D. Data Generation and Filtering

Synthetic lead-vehicle speed profiles represented by the six parameters were generated in two steps based on each developed multivariate distribution model, which might contain a multivariate normal distribution (for correlated parameters) and several fitted distribution models (for uncorrelated parameters).

1) Data generation from the sub-model(s): For correlated parameters, the normalized data is generated from the multivariate normal distribution model. Then, we perform a quantile transformation (an inverse version of the quantile transformation done in the sub-step, multivariate normal distribution modeling) of the normalized data. For every uncorrelated parameter, the data is generated solely from its fitted distribution model.

2) Inverse transformation: For each parameter, perform the inverse transformation of any transformation conducted in the data transformation sub-step during the modeling.

Three types of constraints were set to remove any invalid generated speed profiles.

- Range constraints: Each parameter has its own range limit, such as $\tau_1 \geq 0$ s and $v_c \geq 0$ m/s.
- Physical constraints: First, the lead vehicle should not reverse; its speed should be no less than 0 m/s for the whole duration. Second, the physical constraint for vehicle acceleration applied to extracted events is also applied here: $a_i \in [-g, g]$ $(i = 1, 2)$.
- Categorization constraints: The whole dataset was categorized into multiple sub-datasets according to certain conditions that lead to the modeling sub-dataset, so the generated data should also fulfill these conditions. Some examples are shown in Section IV-C.

E. Validation

A synthetic dataset containing 10,000 synthetic lead-vehicle speed profiles was built by proportionally sampling the speed profiles generated by the distribution model of each sub-dataset. Besides descriptive statistics analysis, non-parametric tests, particularly the weighted two-sample Kolmogorov—Smirnov (KS) tests (using the "Ecume" package in R [40]), were conducted to test whether the synthetic and raw lead-vehicle speed profiles are from different distributions. While lack of significance in the KS test does not mean that the distributions are the same, it does mean that the sample distributions are similar enough that a conclusion of "different" cannot be made with high confidence. Since non-parametric tests generally have lower statistical power (the probability of a test correctly rejecting the null hypothesis) than parametric tests [41], and since similarity is of interest in this application, we adjusted the significance level (α) to 0.10 rather than 0.05. Doing so increases power and reduces the probability of a Type II error (a failure to reject a null hypothesis that is actually false) [42].

IV. Results

A. Parameterization of the Lead-Vehicle Speed Profile

Most events have a decent fitness level. 98.3% (298 out of 303) of the events have an adjusted R-squared $\bar{R}^2$ greater than 0.9. Fig. 5 shows several examples of the fit results. On the other hand, a few events have a lower adjusted R-squared value (an example is shown in Fig. 6). This is because of minor fluctuations (basically noise) during a very small (negligible) lead vehicle speed change during the event, which the piecewise linear model reasonably approximated with a straight line (constant acceleration). In addition, the events with a frequency of 5 Hz (3 out of 303) have an adjusted R-squared greater than 0.99, indicating that the lower-frequency cases do not have a 'worse' fit than the higher-frequency cases.

B. Data Combination

Eighty-two near-crashes, 62.1% of the total number of crashes, are selected as variations of crashes and added to the combined crash dataset, resulting in a sample size of 214 incidents in the combined incident dataset. Table II shows the composition. The weighted CDFs of the six parameters are checked for both the combined crash and combined incident datasets (for instance, the weighted CDFs of v_c are shown in Fig. 7). The difference between the two datasets regarding a single parameter's marginal distribution is negligible. However, there are noticeable variations in the joint distribution when considering multiple variables, as illustrated in Fig. 8. (More details of the comparison are in Appendix E.)

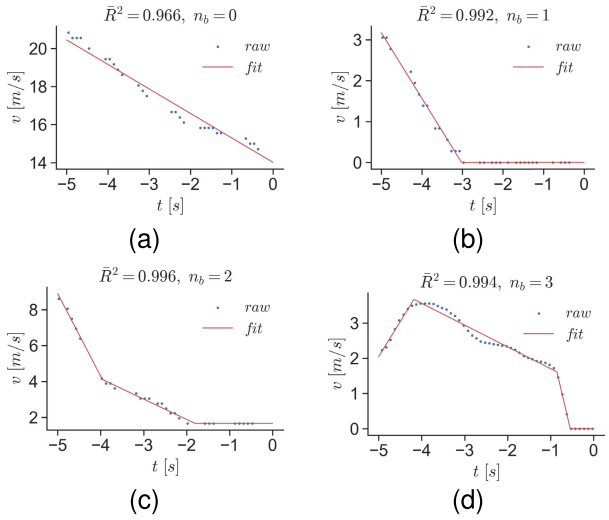


Fig. 5. Examples of fit results. (a)-(d) show the weighted piecewise linear regressions with zero to three breakpoints.

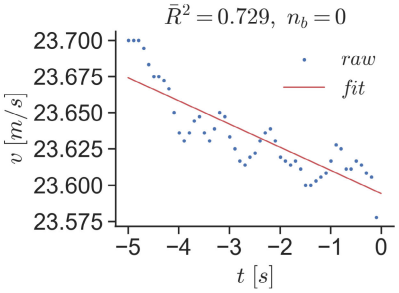


Fig. 6. An example of fit results with lower adjusted R-squared values.

TABLE II
Composition of the Combined Incident Dataset

Group	Notation	Sample size
1	CISS_sc	49
2	SHRP2_sc	20
3	SHRP2_nsc	63
4	SHRP2_nc	82

C. Data Categorization

There can be different patterns in the combined incident dataset, and it is difficult to build a comprehensive multivariate distribution model that covers all the data. However, creating sub-datasets allows a simpler model to be applied to each one.

The relationship of each pair of the six parameters in the combined incident dataset was checked. The relationship

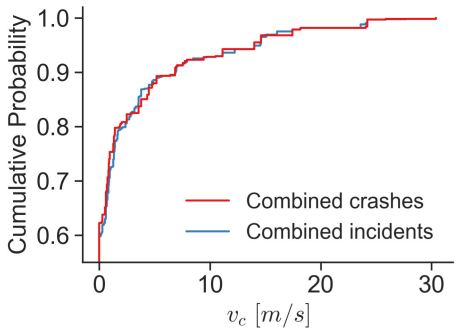


Fig. 7. Weighted CDFs of v_c of both the combined crash and combined incident datasets.

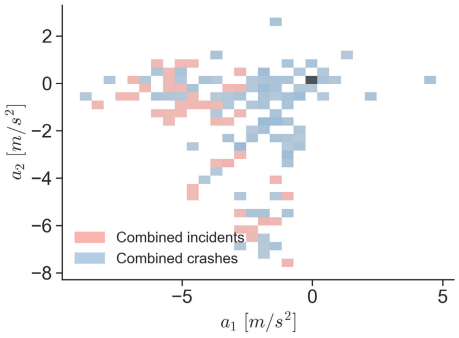


Fig. 8. Joint distributions of a_1 and a_2 of both the combined crash and combined incident datasets. The combined crash dataset is a subset of the combined incident dataset.

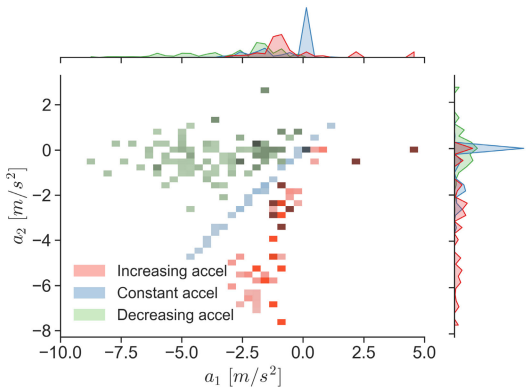


Fig. 9. Three patterns of the lead-vehicle speed change trend.

between a_2 and a_1, indicating the lead-vehicle speed change trend, shows the most distinct patterns, including the proportions in the combined incident dataset, as follows (see Fig. 9).

1) Constant Acceleration ($a_1 = a_2$, 46.2%): The lead vehicle keeps a constant acceleration until a steady speed is reached or an impact occurs. In this case, Segment 2 is non-existent ($\tau_2 = 0$ s), and a_2 is set to a_1.

2) Increasing Acceleration ($a_1 > a_2$, 20.3%): The lead vehicle increases its acceleration as time goes from Segment 2 to Segment 1. For example, the lead vehicle brakes harshly, followed by gentle braking or even acceleration.

3) Decreasing Acceleration ($a_1 < a_2$, 33.5%): The lead vehicle decreases its acceleration from Segment 2 to Segment 1. For example, the lead vehicle accelerates first and then starts to brake harshly.

TABLE III
COMPARISON BETWEEN THE RAW AND SYNTHETIC INCIDENTS

Parameter	Unit	Raw ($n = 132^a$)		Synthetic ($n = 10,000$)		statistic	p-value
		Mean	SD	Mean	SD		
v_c	m/s	2.01	4.69	1.79	3.91	0.05	0.98
a_1	m/s^2	-1.37	1.82	-1.57	1.71	0.10	0.25
a_2	m/s^2	-0.95	1.72	-1.04	1.69	0.07	0.75
τ_s	s	1.73	2.07	1.71	2.06	0.03	0.99
τ_1	s	1.98	1.64	1.95	1.62	0.04	0.99
τ_2	s	1.18	1.30	1.18	1.28	0.05	0.99

[a] Valid sample size = sum of sample weights.

However, the data of each pattern can be appropriately modeled if they are further categorized into smaller sub-datasets. In this case, there were seven sub-datasets (S1-S7), and we needed to model all of them except S1. All the sub-datasets, including the proportions in the combined incident dataset, are listed below.

According to the number of actual parameters, the data of the constant acceleration pattern were divided into the following three sub-datasets:

- S1: Standstill; 25.5%.
- S2: Constant acceleration; 10.5%.
- S3: Constant non-zero acceleration then steady speed; 10.2%.

The increasing acceleration pattern data contains two sub-patterns depending on whether or not the lead vehicle is decelerating in Segment 1, as shown in Fig. 9. Therefore, the data were divided into two sub-datasets:

- S4: Increasing acceleration and $a_1 < 0$ m/s^2; 15.7%.
- S5: Increasing acceleration and $a_1 > 0$ m/s^2; 4.6%.

Finally, the decreasing acceleration pattern data contains two significantly and non-weakly correlated point-mass mixture distribution model parameters, v_c and τ_s. In the multivariate distribution modeling, the data of the decreasing acceleration pattern were divided into two sub-datasets based on whether τ_s equals its point-mass value, 0 s, or not:

- S6: Decreasing acceleration and $\tau_s = 0$ s; 13.3%.
- S7: Decreasing acceleration and $\tau_s > 0$ s; 20.2%.

As mentioned in III-D, along with two other constraints, categorization constraints were used to remove invalid speed profiles. For example, for sub-dataset S4, the categorization constraints were: $a_1 > a_2$ (increasing acceleration), and $a_1 < 0$ m/s^2.

D. Comparison Between the Synthetic and Raw Incidents

Table III compares the six parameters for the synthetic incident dataset and the combined (raw) incident dataset. There are minor differences between the raw and synthetic incidents for each parameter regarding the weighted mean and standard deviation (SD). Furthermore, the two datasets were subjected to weighted Kolmogorov-Smirnov tests to assess whether there are any significant differences in each of the six parameters. The results, presented as p-values in the table, do not indicate any significant difference. However, it is important to note that the lack of significance does not necessarily imply that

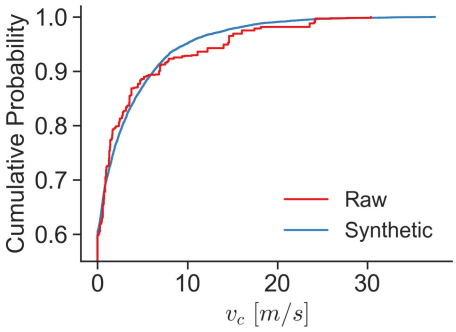


Fig. 10. Weighted CDFs of v_c of the raw and synthetic incidents. The raw incidents are properly weighted to be combined into one dataset. All synthetic incidents have a weight of 1.

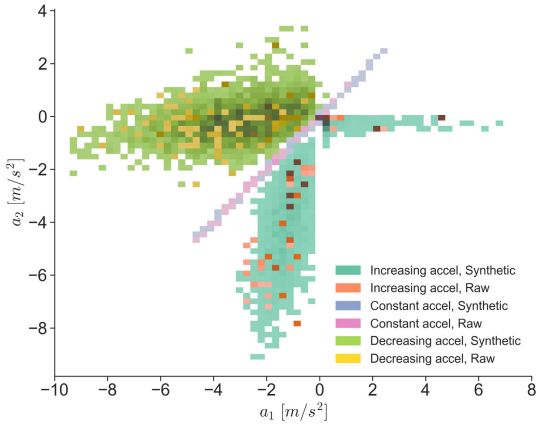


Fig. 11. Comparison between the raw and synthetic incidents: joint distribution of a_1 and a_2.

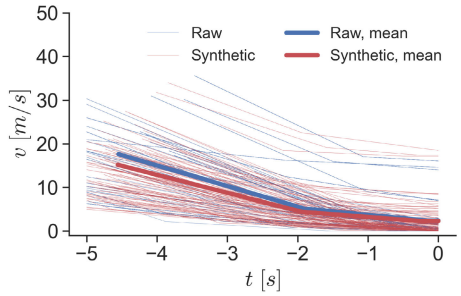


Fig. 12. Lead-vehicle speed profiles from the raw and synthetic incidents of sub-dataset S4 (Increasing acceleration and $a_1 < 0$ m/s^2). The bold lines are with the weighted mean values of parameters describing the speed profiles. The thin lines are 100 randomly sampled profiles for the raw and synthetic incidents respectively.

the datasets are from the same distribution. Despite this, the visual comparison of the well-aligned weighted cumulative distribution functions (CDFs) for each of the six parameters in the two datasets (an example of v_c is illustrated in Fig. 10) reveals substantial similarities. (More details of the weighted CDFs of each parameter are in Appendix E.)

Moreover, the joint distributions of every two parameters and the raw and synthetic speed profiles of every sub-dataset were compared. For instance, Fig. 11 shows the joint distribution of a_1 and a_2, and Fig. 12 shows the comparison results for S4.

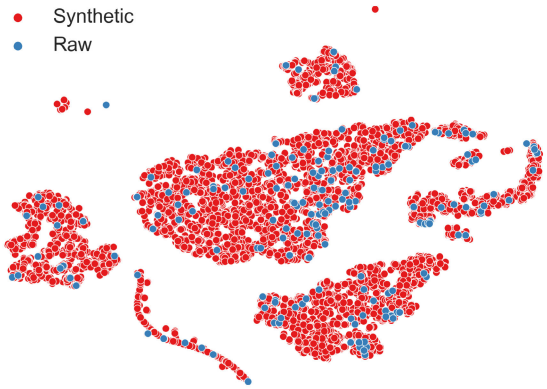


Fig. 13. t-SNE projection of the raw and synthetic incidents.

As a final step, t-distributed stochastic neighbor embedding (t-SNE) was used to visualize the raw and synthetic incidents in two dimensions; see Fig. 13. t-SNE is a statistical method for visualizing high-dimensional data by giving each data point a location in a two or three-dimensional map [43]. In Fig. 13, the blue dots (projection of raw incidents) are surrounded by the red dots (projection of synthetic incidents).

In summary, the synthetic and raw incidents are similar and well-aligned. (More details regarding the comparison are in Appendix E.)

V. Discussion and Conclusions

This study focuses on the lead vehicle's behavior in rear-end crashes, which is mostly independent of the following vehicle's behavior. Thus, this study models the lead-vehicle speed profile without considering its interaction with the following vehicle.

A piecewise linear model was used to represent the lead-vehicle speed profile in the pre-crash phase, providing a more accurate digital representation of the lead-vehicle kinematics than the conventional constant acceleration/deceleration model. Two datasets (CISS and SHRP2) were combined to produce a comprehensive rear-end critical incident (crash/near-crash) dataset that captures the full severity range. Multivariate distribution models were constructed to generate synthetic lead-vehicle speed profiles that were compared with the raw speed profiles.

The results show that the piecewise linear model has good fitting performance. The raw and synthetic incidents display a notable alignment. Moreover, a range of different lead-vehicle speed patterns were revealed, indicating the proposed piecewise linear model's greater accuracy compared to the conventional constant acceleration/deceleration model. For example, the lead vehicle could exhibit harsh braking followed by gentle braking (as shown in Fig. 12) or even acceleration. In addition, the lead vehicle does not necessarily brake harshly. In fact, in many cases, the lead vehicle keeps a constant speed or is at a standstill for a considerable time (up to five seconds) prior to the crash.

In summary, the proposed model accurately matches lead-vehicle kinematics from in-depth pre-crash/near-crash data across the full severity range, outperforming previously

TABLE IV
RESULTS OF BOOTSTRAPPING

Parameter	p-value	
	90% samples	80% samples
v_c	0.98	0.94
a_1	0.83	0.63
a_2	0.96	0.93
τ_s	0.80	0.75
τ_1	0.96	0.91
τ_2	0.73	0.67

existing lead vehicle models in terms of both severity range and precision. Furthermore, in addition to generating simulated rear-end crash scenarios, this model has the potential to aid substantially in the reconstruction of individual real-world crashes. That is, by offering more realistic speed profiles for reconstructed crashes (considering the speed at impact and other constraints as discussed in Section V-E), the model provides a means of generating a distribution of possible speed profiles during the reconstruction process instead of providing only a single speed profile.

A. Robustness of the Results

When conducting a study with a relatively small sample size, it is crucial to examine the robustness of the results. To do so, we conducted a bootstrapping study to test the multivariate distribution modeling method. This process is outlined below:

1) Create 200 new datasets by randomly sampling 100 times (without replacement) from all samples with a sample size reduction of 10% and 20%, respectively.
2) For every new dataset, go through the multivariate distribution modeling process and generate a new synthetic dataset with a sample size of 1,000.
3) Perform two-sample KS tests for each of the six parameters to compare each of the bootstrapped synthetic datasets with the original synthetic dataset generated using all samples.
4) Compute the p-value of bootstrapping, which represents the proportion of bootstrap samples for which the KS test is not significant ($p > 0.1$) [44].

The results do not show any significance (all p-values are larger than 0.1), as indicated in Table IV. Hence, the robustness of the proposed modeling method was demonstrated.

B. Sampling Bias of Driver Age in the SHRP2 Dataset

In addition to the sampling bias in near-crashes mentioned in Section III-B, there is also a driver-age sampling bias in the SHRP2 dataset. Both young and old drivers are over-represented [45]. We investigated the possible impact of this bias and concluded that it can be ignored in this study. (More details are in Appendix F.)

C. Adding Selected Near-Crashes as Variations of Crashes

In this study, near-crashes similar to a crash in the combined crash dataset were selected and added as variations of crashes with sample weighting adjustment. There are two rationales for doing so.

First, the lead-vehicle speed profiles in a near-crash and a crash can be similar. Given the same lead-vehicle behavior, a rear-end near-crash incident can easily turn into a crash if the following vehicle's driver reacts more slowly or brakes less harshly.

Second, the sample weighting adjustment practically mitigates the risk that the added near-crashes will change the raw distributions. With the sample weighting adjustment mentioned in Section III, the raw parameter distributions of the combined crash dataset are retained in the new, combined incident dataset. At the same time, with the added samples, there are more observed values. Thus, a more reliable distribution modeling can be achieved.

It is also worth mentioning the alternative to weigh the six parameters (based on prior knowledge) when computing the Euclidean distance between two events. For instance, v_c could have a larger weight than others because it directly relates to the impact result. Future work should address such a weighting method.

D. Limitations

In addition to the issue with correlated point-mass mixture distribution parameters (as discussed in Section III-C.2) and the reduced statistical power of the non-parametric tests used (as outlined in Section III-E), the following limitations are noteworthy:

1) Only the kinematics of the lead vehicle are considered in this work. Numerous variables can influence the occurrence of a crash, including road structure, traffic signals, and weather conditions. Future research should address these considerations when more comprehensive data are available. Nonetheless, a precise description of lead-vehicle kinematics by utilizing diverse data sources and considering crashes occurring in various situations is instrumental in capturing an important part of the overall variability observed in real-world rear-end crashes.
2) The modeled lead vehicle's acceleration is not consistently smooth. This could be attributed to the fact that the speed of the lead vehicle is modeled using a piecewise linear model, resulting in a sudden change in acceleration as it moves from one segment to another. Future work should aim to smooth the acceleration profile, potentially by introducing jerk during transitions.
3) To avoid any sharp acceleration pulse near impact and accommodate the lowest data frequency (5 Hz), we extracted raw data up to only 0.3 s before impact for each crash. To investigate the influence of the section of 0.3 s, we extracted raw data up to 0.2 s before impact for crashes with a data frequency of 10 Hz. Then we refitted those cases into the piecewise linear model and obtained a new set of six-dimensional vectors (parameterized speed profiles). At last, we checked the difference of each parameter between the raw and new fit results. Table V shows percentiles of the absolute value of the difference for each parameter, indicating negligible differences between the fit results. Therefore, the outcomes of this study are not sensitive to the selection of 0.3 s.

TABLE V
COMPARISON OF THE NEW AND RAW FIT RESULTS

Parameter	Unit	Percentile				
		90	92.5	95	97.5	99
$\lvert\Delta v_c\rvert$	m/s	0.10	0.10	0.13	0.27	0.47
$\lvert\Delta a_1\rvert$	m/s^2	0.10	0.12	0.16	0.29	0.55
$\lvert\Delta a_2\rvert$	m/s^2	0.02	0.03	0.03	0.07	0.15
$\lvert\Delta \tau_s\rvert$	s	0.04	0.05	0.06	0.07	0.14
$\lvert\Delta \tau_1\rvert$	s	0.07	0.07	0.08	0.14	0.34
$\lvert\Delta \tau_2\rvert$	s	0.05	0.07	0.07	0.11	0.13

4) The method of multivariate distribution modeling only considers the linear correlation between two parameters and disregards any potential nonlinear relationship between them, as well as weak or non-significant correlations. In addition, the correlated parameters are assumed to follow a multivariate normal distribution, which effectively models the parameters as linearly related. These simplifications are made to keep the model tractable and avoid over-interpreting the relationships between parameters, as it is not feasible to create a complex multivariate model with a small dataset without a substantial risk of overfitting. These simplifications may, however, reduce the accuracy of the model. Unfortunately, it is impossible to investigate the consequences with the available data, but future work should address this issue.
5) In terms of validation, the marginal distributions of each parameter between the raw and synthetic datasets were compared using the weighted two-sample KS test. Future studies should explore methods for statistically comparing multivariate joint distributions, rather than just marginal distributions.

E. Application

1) Data Combination Method: The proposed data combination method combines rear-end crashes from two datasets and includes selected rear-end near-crashes from the SHRP2 dataset as variations of crashes. This method is generic and can be adapted to other situations, such as combining multiple crash datasets of other crash scenarios. It is also important to mention that near-crashes are used as substitutes for crashes because of their strong connection and similarities. When applying this method, we need to ensure the data to be added can be used as substitutes.

2) Multivariate Distribution Modeling Method: The multivariate distribution modeling method proposed in this research can be easily adapted to other situations where building a distribution model from a relatively small dataset is needed and an understanding of the underlying distribution is available. For instance, this method can be used to analyze other crash scenarios.

3) Synthetic Data: The synthetic data generated in this study can be useful in both rear-end crash reconstructions and safety assessments of ADAS and ADS.

- For rear-end crash reconstructions, despite a relatively accurate estimation of the impact speed, it is rarely possible to reconstruct the speed profile of the vehicles during the pre-crash phase if no recorded data are available. Post-crash interviews and evidence from the on-scene investigation may be the only source of information. Usually, the lead vehicle would be assumed to be moving with a constant acceleration/deceleration before the crash when no information to the contrary is available. Synthetic data can provide alternative speed profiles given the speed at impact and other available constraints. The use of synthetic data will make the reconstruction easier and more reliable since they can provide prior knowledge of the lead-vehicle speed profile based on actual collected pre-crash data.
- For the safety assessments of ADAS and ADS, the synthetic data can be used to create virtual crashes for testing whether the crash can be avoided with a given ADAS or ADS.

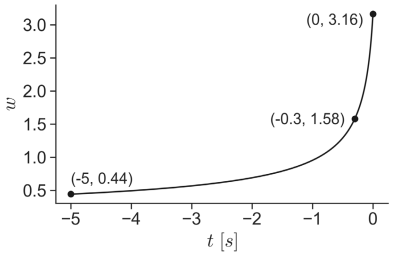


Fig. 14. Sample weight against time.

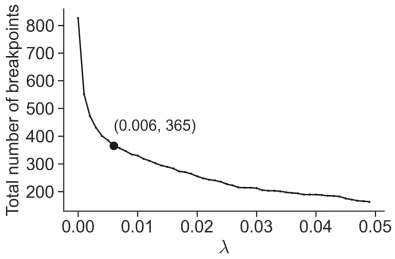


Fig. 15. Total number of breakpoints against λ. λ was set as the elbow of the curve, 0.006.

VI. FUTURE WORK

This study is the first step in generating rear-end crash scenarios for the safety assessments of ADAS and ADS. Future work will use models of the following vehicle's behavior together with the lead-vehicle kinematics model from this work to generate rear-end crash scenarios.

After completing these steps for the rear-end crash scenario, we will move on to other crash scenarios. Moreover, the parameterized data for the additional scenarios will be added to the same online combined incident dataset.

APPENDIX A
PARAMETERIZATION OF THE LEAD-VEHICLE SPEED PROFILES

A. Sample Weights

In each lead-vehicle speed profile, the weight of sample i is defined as a function of $t(i)$,

$$w_i = (0.1 - t_i)^{-0.5}. \tag{17}$$

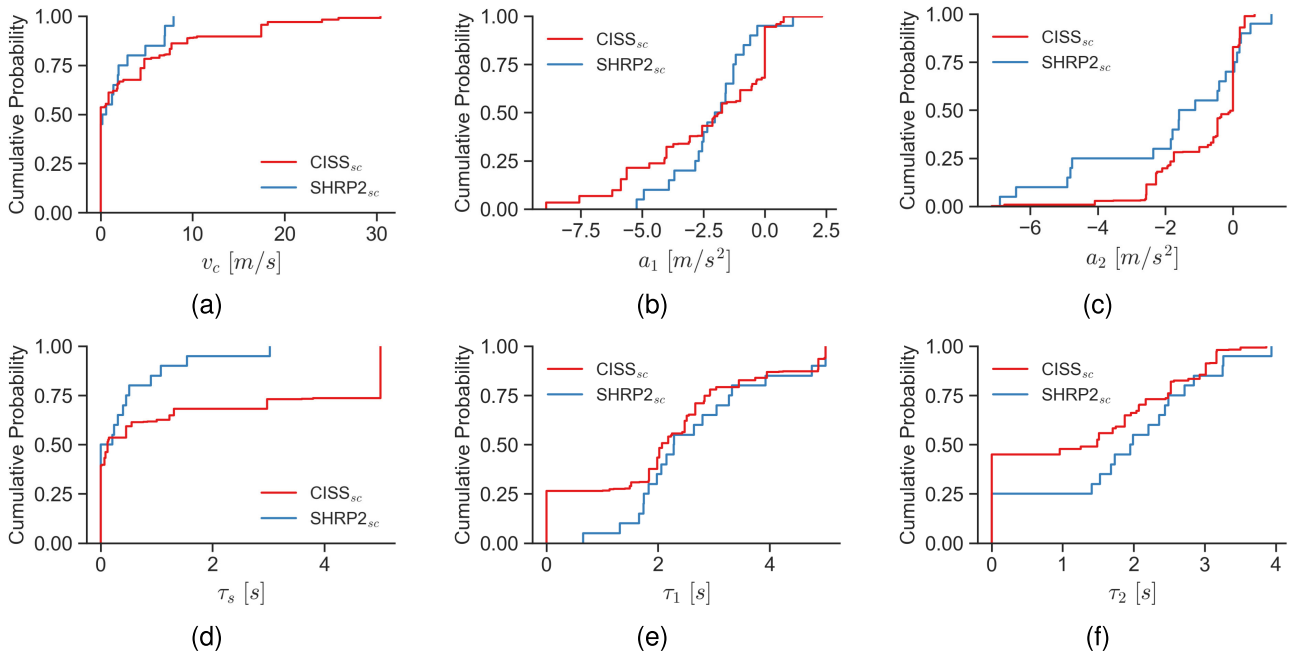


Fig. 16. Comparison between the weighted CDFs of the six parameters of CISS_sc and SHRP2_sc.

TABLE VI
WEIGHTED TWO-SAMPLE KS TESTS BETWEEN CISS_SC AND SHRP2_SC

Parameter	v_c	a_1	a_2	τ_s	τ_1	τ_2
statistic	0.15	0.28	0.27	0.27	0.26	0.26
p-value	0.97	0.33	0.40	0.37	0.41	0.44

*The sample sizes are as follows: 49 for CISS_sc and 20 for SHRP2_sc.

The function is plotted in Fig. 14; the weight ranges from 0.44 to 3.16 for near-crashes and from 0.44 to 1.58 for crashes. (The weight differences aren't huge.)

We wanted to fit the lead-vehicle speed profile into a piecewise linear model with as few breakpoints as possible, so we limited the maximum amount of breakpoints. Without sample weights, the piecewise linear model treats the speed changes close to time zero and those close to the start of the incident equally. In other words, the chance of the piecewise linear model adding a breakpoint is the same throughout the time period. However, with the sample weights considered, the model will prioritize capturing the speed changes closer to time zero by adding a new breakpoint.

B. Loss Function

In the piecewise linear model, a larger number of breakpoints covers more details (segments) of the speed changes with more parameters, leading to a more accurate fitting. However, as dimensionality (the number of parameters) increases, the complexity of the multivariate distribution modeling increases and the amount of data needed often grows exponentially [46]. Therefore, we introduced the loss function

$$L = (\epsilon + \lambda \cdot \frac{\max(v)}{\Delta v + \epsilon}) \cdot n_b - R^2. \tag{18}$$

to balance accuracy and complexity.

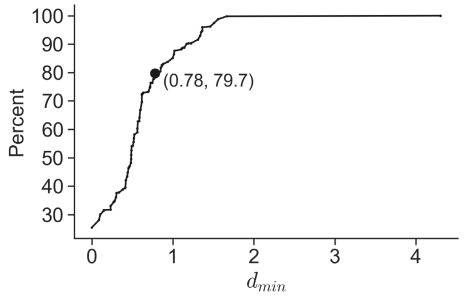


Fig. 17. CDF of the minimum Euclidean distance of the combined crashes. The threshold d_{thd} is set as the elbow of the curve, 0.780.

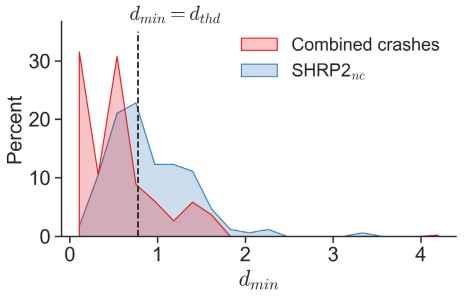


Fig. 18. Distributions of the minimum Euclidean distance of the combined crashes and SHRP2_nc.

The loss function selects the weighted piecewise linear regression with the smallest loss among several with various numbers of breakpoints. The function contains two parts: 1) a penalty for the number of breakpoints and 2) a reward for fitting accuracy.

The penalty part, $(\epsilon + \lambda \cdot \frac{\max(v)}{\Delta v + \epsilon}) \cdot n_b$, is based on the perceptibility of the lead vehicle's speed change to the following vehicle's driver. An easier perceptible speed change corresponds with a smaller penalty given the same number

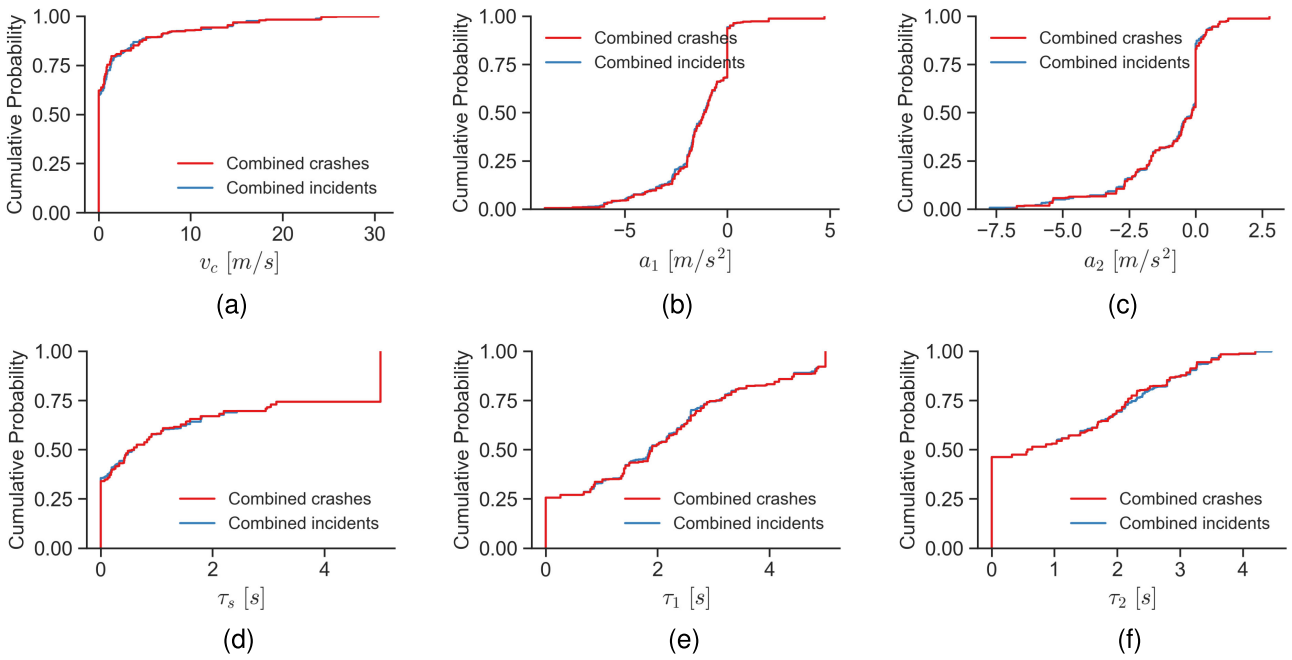


Fig. 19. Weighted CDFs of the six parameters of the combined crash and the combined incident datasets. The difference between the two datasets in terms of a single parameter's marginal distribution is negligible.

of breakpoints. According to Lee [47], the changing rate of the optical size (width) of the lead vehicle on the following vehicle's driver's retina is computed as

$$\theta' = \frac{d}{dt}(\tan^{-1}\frac{W}{d}) = \frac{W}{W^2+d^2}\cdot(v_f - v_l), \qquad (19)$$

where θ is the retinal image size, W is the width of the lead vehicle, d is the following distance, v_f is the following vehicle's speed, and v_l is the lead vehicle's speed. The lead vehicle's speed change will affect θ' and, therefore, be perceived by the following vehicle's driver. Considering that 1) a larger lead vehicle's speed change (Δv) leads to a larger change of θ', 2) a higher lead vehicle's speed ($\max(v)$) is usually associated with a larger following distance leading to a smaller change of θ', and 3) a lead speed change associated with a larger change of θ' is more perceptible, the penalty should increase with the decrease of Δv or the increase of $\max(v)$ given the same number of breakpoints n_b. Regarding the third point, it is worth mentioning that Tian et al. [48] discovered an interesting aspect regarding pedestrian crossing. As the approaching vehicle gets closer, the pedestrian's change rate of the vehicle's optical size initially increases, but after reaching approximately one meter, it decreases. This is because the pedestrian and the vehicle are not in the same lane. In contrast, our study focuses on two vehicles in the same lane, and we observed a monotonic increase in θ', thus confirming the third point. In addition, instead of $\max(v^2)$, $\max(v)$ is used in the penalty part because the following distance does not increase linearly with speed [49].

The reward part, $-R^2$, rewards fitting accuracy. Given the same penalty, a larger R-squared indicates better fitting accuracy and, thus, a smaller loss.

TABLE VII
THREE AGE GROUPS IN THE SHRP2 CRASH DATASET

Group	Age	Sample size
Young	<25	24
Middle-aged	25-64	44
Senior	>64	15

TABLE VIII
WEIGHTED TWO-SAMPLE KS TESTS AMONG THREE AGE GROUPS IN THE SHRP2 CRASH DATASET

	p-value					
Comparison[a]	v_c	a_1	a_2	τ_s	τ_1	τ_2
Young V.S. Middle-aged	0.81	0.73	0.03[b]	0.15	0.57	0.02[b]
Young V.S. Senior	0.73	0.56	0.92	0.81	0.81	0.08[b]
Middle-aged V.S. Senior	0.96	0.96	0.44	0.49	0.49	0.96

[a] The sample sizes are as follows: 24 for the young group, 44 for the middle-aged group, and 15 for the senior group.
[b] Significant under the significance level of 0.10.

C. Selection of Pre-Configured Parameters

The two pre-configured parameters in the piecewise linear model are λ and $n_{b,max}$. To determine $n_{b,max}$, we randomly selected a small sub-dataset (30 events) across the CISS and SHRP2 datasets and manually annotated the preferred number of breakpoints for each event in the sub-dataset. It was found that most of the events can be covered with no more than two breakpoints, and very few require three breakpoints. Therefore, $n_{b,max}$ was set as 3. While λ was set as the elbow of the curve shown in Fig. 15 (the total number of breakpoints for all events against λ).

APPENDIX B
COMPARISON BETWEEN CISS_SC AND SHRP2_SC

The weighted two-sample Kolmogorov—Smirnov (KS) tests of each of the six parameters were conducted to determine

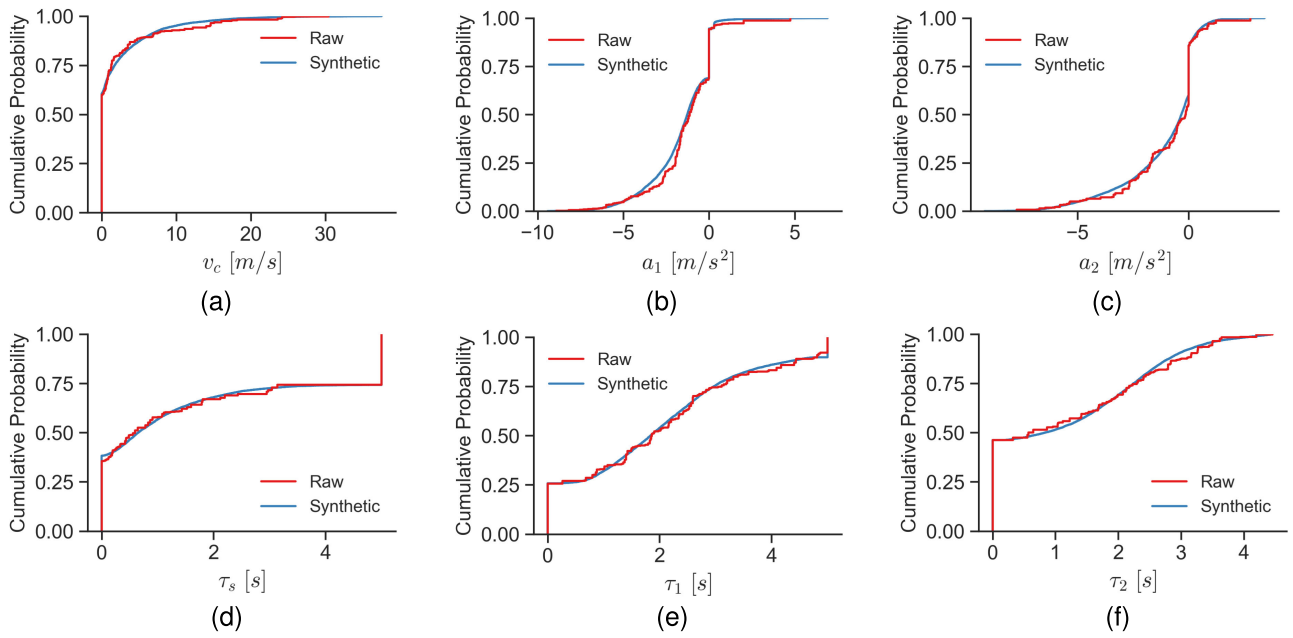


Fig. 20. Weighted CDFs of the six parameters of the raw and synthetic incidents.

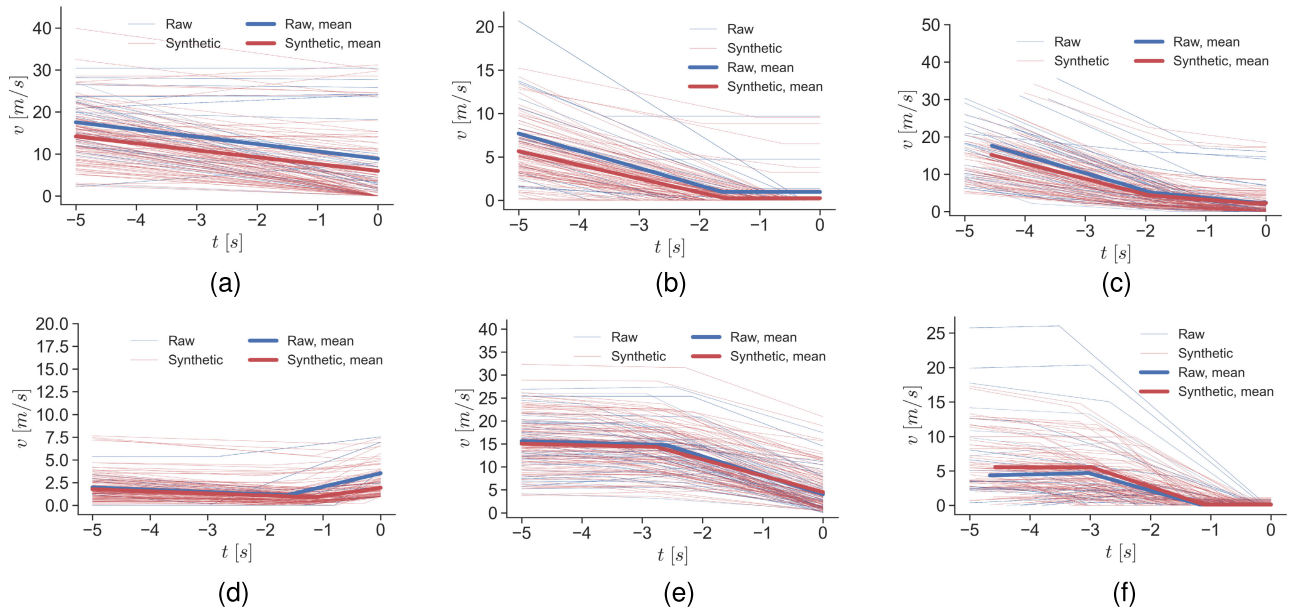


Fig. 21. Lead-vehicle speed profiles of the raw and synthetic incidents of all sub-datasets except S1. (a)-(b) show results from S2 to S7. The bold lines are with the weighted mean values of parameters describing the speed profile. The thin lines are 100 randomly sampled profiles for the raw and synthetic incidents respectively.

whether the severe crashes in the CISS and SHRP2 datasets are from the same distribution. The results in Table VI show no significant differences. The weighted cumulative distribution functions (CDFs) of the six parameters of CISS_sc and SHRP2_sc are shown in Fig. 16.

Appendix C
Weight Trimming of Raw Sample Weights in CISS_sc

The weight-trimming approach used in the study is proposed by Van de Kerckhove et al. In weight trimming, weights exceeding a specified cut-point are trimmed to that value, as expressed in (20).

$$w_{jt} = \begin{cases} w_0, & \text{if } w_j > w_0; \\ w_j, & \text{otherwise.} \end{cases}, \tag{20}$$

where w_j is the weight prior to trimming, and w_0 is the trimming cut-point which is defined as

$$w_0 = 3.5\sqrt{1 + CV^2(w_j)} \cdot \text{median}(w_j), \tag{21}$$

where $CV^2(w_j)$ is the coefficient of variation of w_j. In the case of CISS_sc, the trimming cut-point is 1797.3.

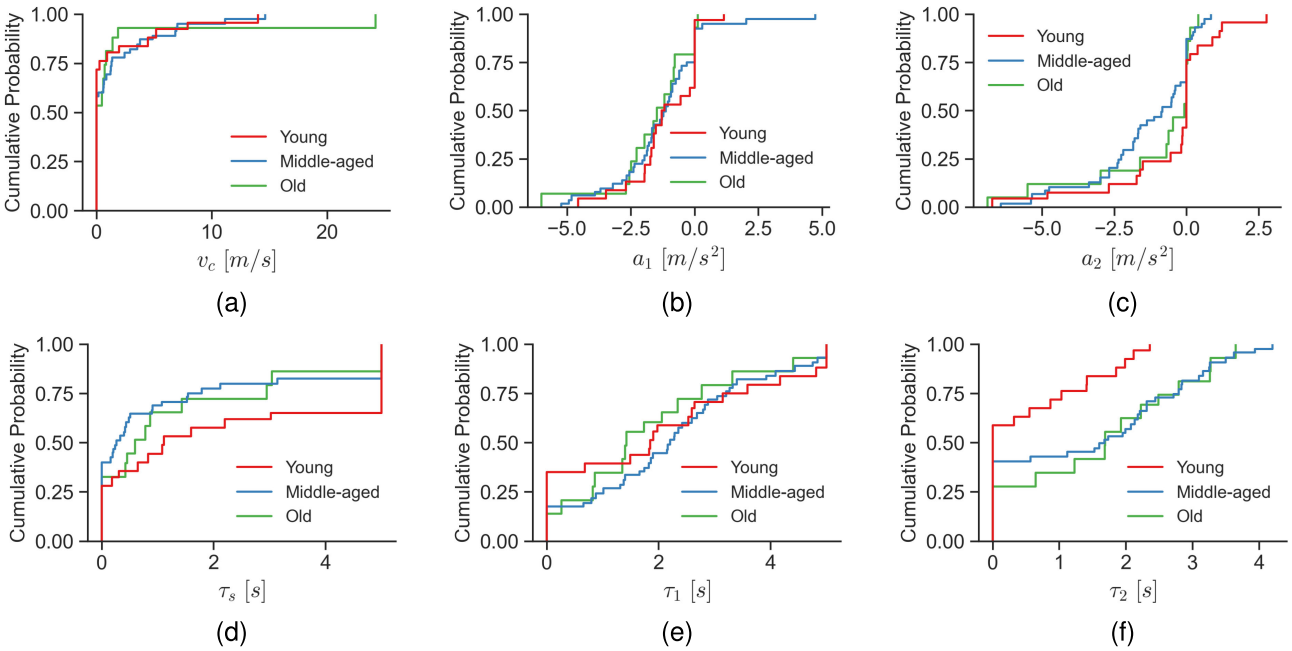


Fig. 22. Comparison among the weighted CDFs of the six parameters of different age groups in the combined crash dataset.

APPENDIX D
ADDING SELECTED NEAR-CRASHES AS VARIATIONS OF CRASHES

The CDF curve of the minimum Euclidean distance of the combined crashes is shown in Fig. 17. The pre-configured parameter d_{thd} was set as the elbow of the curve, where the curve slope slows down, and it can be seen as the start of the long tail. Moreover, the distributions of the minimum Euclidean distance of the combined crashes and SHRP2_nc are shown in Fig. 18.

APPENDIX E
MORE DETAILED RESULTS

The weighted CDFs of the six parameters of the combined crash and the combined incident datasets are shown in Fig. 19.

The weighted CDFs of the six parameters of the raw and synthetic incidents are shown in Fig. 20. Fig. 21 shows the lead-vehicle speed profiles for the raw and synthetic incidents of all sub-datasets except S1.

APPENDIX F
DRIVER-AGE SAMPLING BIAS IN SHRP2

According to previous research [45], the SHRP2 dataset shows an over-representation of drivers under age 25 or over 64. To investigate the effects of the sampling bias on our study, the SHRP2 crash dataset is divided into three groups accordingly: young (under 25), middle-aged (over 24 and under 65), and senior (over 64), as shown in Table VII. The weighted two-sample KS tests were applied to determine whether there is any significant difference between any two of the three groups for each of the six parameters [v_c, a_1, a_2, τ_s, τ_1, τ_2]. Based on the results presented in Table VIII, it can be observed that only three out of eighteen comparisons depict notable dissimilarities at a significance level of 0.10. In addition, most of the weighted CDFs of the six parameters of different age groups shown in Fig. 22 are well-aligned, indicating decent similarities. Therefore, we argue that the sampling bias of driver age in the SHRP2 dataset can be ignored in this study.

REFERENCES

[1] A. Georgi, M. Zimmermann, T. Lich, L. Blank, N. Kickler, and R. Marchthaler, "New approach of accident benefit analysis for rear end collision avoidance and mitigation systems," in *Proc. 21st Int. Tech. Conf. Enhanced Saf. Vehicles*, 2009, p. 281.

[2] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 8, pp. 87456–87477, 2020.

[3] S. Riedmaier, D. Schneider, D. Watzenig, F. Diermeyer, and B. Schick, "Model validation and scenario selection for virtual-based homologation of automated vehicles," *Appl. Sci.*, vol. 11, no. 1, p. 35, 2020.

[4] P. Yves et al., "A comprehensive and harmonized method for assessing the effectiveness of advanced driver assistance systems by virtual simulation: The pears initiative," in *Proc. 24th Int. Tech. Conf. Enhanced Saf. Vehicles (ESV)*, 2015, pp. 1–12.

[5] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*. Berlin, Germany: Springer, 2018, pp. 621–635.

[6] W. Li et al., "AADS: Augmented autonomous driving simulation using data-driven algorithms," *Sci. Robot.*, vol. 4, no. 28, Mar. 2019, Art. no. eaaw0863.

[7] S. Feng, X. Yan, H. Sun, Y. Feng, and H. X. Liu, "Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment," *Nature Commun.*, vol. 12, no. 1, pp. 1–14, Feb. 2021.

[8] G. Savino, J. Mackenzie, T. Allen, M. Baldock, J. Brown, and M. Fitzharris, "A robust estimation of the effects of motorcycle autonomous emergency braking (MAEB) based on in-depth crashes in Australia," *Traffic Injury Prevention*, vol. 17, pp. 66–72, Sep. 2016.

[9] J. Bärgman, C.-N. Boda, and M. Dozza, "Counterfactual simulations applied to SHRP2 crashes: The effect of driver behavior models on safety benefit estimations of intelligent safety systems," *Accident Anal. Prevention*, vol. 102, pp. 165–180, May 2017.

[10] A. Leledakis, M. Lindman, J. Östh, L. Wågström, J. Davidsson, and L. Jakobsson, "A method for predicting crash configurations using counterfactual simulations and real-world data," *Accident Anal. Prevention*, vol. 150, Feb. 2021, Art. no. 105932.

[11] P. Olleja, J. Bärgman, and N. Lubbe, "Can non-crash naturalistic driving data be an alternative to crash data for use in virtual assessment of the safety performance of automated emergency braking systems?" *J. Saf. Res.*, vol. 83, pp. 139–151, Dec. 2022.

[12] L. Wang, H. Zhong, W. Ma, M. Abdel-Aty, and J. Park, "How many crashes can connected vehicle and automated vehicle technologies prevent: A meta-analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105299.

[13] T. Seacrist, R. Sahani, G. Chingas, E. C. Douglas, V. Graci, and H. Loeb, "Efficacy of automatic emergency braking among risky drivers using counterfactual simulations from the SHRP2 naturalistic driving study," *Saf. Sci.*, vol. 128, Aug. 2020, Art. no. 104746.

[14] A. Gambi, T. Huynh, and G. Fraser, "Generating effective test cases for self-driving cars from police reports," in *Proc. 27th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, Aug. 2019, pp. 257–267.

[15] X. Wang et al., "Autonomous driving testing scenario generation based on in-depth vehicle-to-powered two-wheeler crash data in China," *Accident Anal. Prevention*, vol. 176, Oct. 2022, Art. no. 106812.

[16] A. Z. Zambom and D. Ronaldo, "A review of kernel density estimation with applications to econometrics," *Int. Econ. Rev.*, vol. 5, no. 1, pp. 20–42, 2013.

[17] National Center for Statistics and Analysis. (2022). *Traffic Safety Facts 2020: A Compilation of Motor Vehicle Crash Data*. National Highway Traffic Safety Administration, Washington, DC, USA. DOT HS 813 375. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813375

[18] T. L. Brown, J. D. Lee, and D. V. McGehee, "Human performance models and rear-end collision avoidance algorithms," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 43, no. 3, pp. 462–482, Sep. 2001.

[19] J. D. Lee, D. V. McGehee, T. L. Brown, and M. L. Reyes, "Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 44, no. 2, pp. 314–334, Jun. 2002.

[20] G. Markkula, O. Benderius, K. Wolff, and M. Wahde, "A review of near-collision driver behavior models," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 54, no. 6, pp. 1117–1143, Dec. 2012.

[21] G. Li, W. Wang, S. E. Li, B. Cheng, and P. Green, "Effectiveness of flashing brake and hazard systems in avoiding rear-end crashes," *Adv. Mech. Eng.*, vol. 6, Jan. 2014, Art. no. 792670.

[22] J. Bärgman, V. Lisovskaja, T. Victor, C. Flannagan, and M. Dozza, "How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP2," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 35, pp. 152–169, Nov. 2015.

[23] G. Markkula, J. Engström, J. Lodin, J. Bärgman, and T. Victor, "A farewell to brake reaction times? Kinematics-dependent brake response in naturalistic rear-end emergencies," *Accident Anal. Prevention*, vol. 95, pp. 209–226, Oct. 2016.

[24] X. Wang, M. Zhu, M. Chen, and P. Tremont, "Drivers' rear end collision avoidance behaviors under different levels of situational urgency," *Transp. Res. C, Emerg. Technol.*, vol. 71, pp. 419–433, Oct. 2016.

[25] M. Svärd, G. Markkula, J. Bärgman, and T. Victor, "Computational modeling of driver pre-crash brake response, with and without off-road glances: Parameterization using real-world crashes and near-crashes," *Accident Anal. Prevention*, vol. 163, Dec. 2021, Art. no. 106433.

[26] F. Zhang, E. Y. Noh, R. Subramanian, and C.-L. Chen, "Crash investigation sampling system: Sample design and weighting," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, DC, USA, Tech. Rep. DOT HS 812 804, 2019.

[27] R. Subramanian and E. Acevedo-Díaz, "Crash investigation sampling system 2019 data manual," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. DOT HS 813 040, 2020.

[28] J. M. Hankey, M. A. Perez, and J. A. McClafferty, "Description of the SHRP2 naturalistic database and the crash, near-crash, and baseline data sets," Virginia Tech Transp. Inst., Blacksburg, VA, USA, Tech. Rep. S2-S31-RW-3, 2016.

[29] J. Wu. (2023). *QUADRIS Project Pre-crash/near-crash Dataset*. [Online]. Available: https://github.com/JianWu09/QUADRIS-project-Pre-crash-near-crash-database

[30] T. A. Gennarelli and E. Wodzin, "AIS 2005: A contemporary injury scale," *Injury*, vol. 37, no. 12, pp. 1083–1091, Dec. 2006.

[31] *Event Detail Table Documentation*. Accessed: Oct. 19, 2022. [Online]. Available: https://insight.shrp2nds.us/info/printable/38?type=dataset

[32] C. Pilgrim, "Piecewise-regression (aka segmented regression) in Python," *J. Open Source Softw.*, vol. 6, no. 68, p. 3859, Dec. 2021.

[33] F. Potter and Y. Zheng, "Methods and issues in trimming extreme weights in sample surveys," in *Proc. Amer. Stat. Assoc., Sect. Surv. Res. Methods*. Alexandria, VA, USA: American Statistical Association, 2015, pp. 2707–2719.

[34] W. Van de Kerckhove, L. Mohadjer, and T. Krenzke, "A weight trimming approach to achieve a comparable increase to bias across countries in the programme for the international assessment of adult competencies," in *JSM Proceedings, Survey Research Methods Section*. Alexandria, VA, USA: American Statistical Association, 2014, pp. 655–666.

[35] R. G. Easterling, "Passion-driven statistics," *Amer. Statistician*, vol. 64, no. 1, pp. 1–5, Feb. 2010.

[36] J. Pasek, A. Tahk, G. Culter, and M. Schwemmle. (2021). *Weights: Weighting and Weighted Statistics*. [Online]. Available: https://CRAN.R-project.org/package=weights

[37] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[38] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, vol. 53. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[39] H. A. Panofsky, G. W. Brier, and W. H. Best, "Some application of statistics to meteorology," Mineral Ind. Continuing, Philadelphia, PA, USA, Tech. Rep., 1958.

[40] H. Roux de Bezieux. (2021). *Ecume: Equality 2 (or K) Continuous Univariate Multivariate Distributions*. R Package Version 0.9.1. [Online]. Available: https://CRAN.R-project.org/package=Ecume

[41] L. M. Sullivan, *Essentials of Biostatistics for Public Health*. Sudbury, MA, USA: Jones & Bartlett, 2022.

[42] T. Dybå, V. B. Kampenes, and D. I. K. Sjøberg, "A systematic review of statistical power in software engineering experiments," *Inf. Softw. Technol.*, vol. 48, no. 8, pp. 745–755, Aug. 2006.

[43] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[44] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.

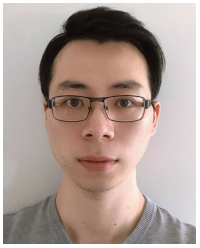
[45] C. Flannagan, J. Bärgman, and A. Bálint, "Replacement of distractions with other distractions: A propensity-based approach to estimating realistic crash odds ratios for driver engagement in secondary tasks," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 63, pp. 186–192, May 2019.

[46] M. Köppen, "The curse of dimensionality," in *Proc. 5th Online World Conf. Soft Comput. Ind. Appl. (WSC)*, vol. 1, 2000, pp. 4–8.

[47] D. N. Lee, "A theory of visual control of braking based on information about time-to-collision," *Perception*, vol. 5, no. 4, pp. 437–459, 1976.

[48] K. Tian et al., "Explaining unsafe pedestrian road crossing behaviours using a psychophysics-based gap acceptance model," *Saf. Sci.*, vol. 154, Oct. 2022, Art. no. 105837.

[49] A. Loulizi, Y. Bichiou, and H. Rakha, "Steady-state car-following time gaps: An empirical study using naturalistic driving data," *J. Adv. Transp.*, vol. 2019, pp. 1–9, May 2019.



Jian Wu received the B.S. and M.S. degrees in automotive engineering from Tsinghua University, Beijing, China, in 2013 and 2016, respectively. He is currently pursuing the industrial Ph.D. degree with the Volvo Cars Safety Center and the Department of Mechanics and Maritime Sciences, Chalmers University of Technology, Gothenburg, Sweden. He is the author or coauthor of four journal articles and two conference papers. His current research interests include driver behavior modeling, crash data synthesis, and the safety assessments of ADAS and ADS.

Carol Flannagan received the M.A. degree in statistics and the Ph.D. degree in mathematical psychology from the University of Michigan. She is currently a Research Professor with the University of Michigan Transportation Research Institute (UMTRI), Ann Arbor, MI, USA, and an affiliated Associate Professor with the Chalmers University of Technology, Gothenburg, Sweden. Her work in transportation research encompasses the analysis of a wide variety of transportation-related data and the development of innovative statistical methods for transportation research. She is also working on a number of projects related to safety assessment and benefits assessment for advanced technologies, including ADS.

Ulrich Sander received the B.S. degree in biomedical engineering from the University of Aachen, Germany, in 1997, the M.S. degree in accident research from Graz University of Technology, Austria, in 2008, and the Ph.D. degree in machine and vehicle systems from the Chalmers University of Technology, Gothenburg, Sweden. He has worked for over 20 years in different positions, such as a Data Analyst and a Senior Principal Researcher with Autoliv Research, Germany and Sweden. Since 2022, he has been a Technical Expert with the Safety Centre of Volvo Cars, leading the analysis of field data with a focus on crashes and their consequences.

Jonas Bärgman received the M.Sc. degree in mechanical engineering and the Ph.D. degree in machine and vehicle systems from the Chalmers University of Technology, Gothenburg, Sweden, in 1997 and 2016, respectively. After his degree, he was an industrial researcher in in-crash safety with Autoliv Research for three years and a software developer with AB Volvo for two years. At this point, he continued his career with Autoliv Research (again) in the domain of pre-crash safety, focusing on human factors and driver behavior. In 2009, he started working with the Chalmers University of Technology to build a research group on active safety, where he is currently a Professor. He is also the examiner for the course “vehicle and traffic safety” in the master’s degree program in Mobility Engineering. His main research interests include virtual safety assessment and its components, including driver behavior modeling, scenario generation, and statistical methods.