THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Generation of Representative Pre-Crash Scenarios Across the Full Severity Range Using Real-World Crash Data:

Towards more accurate virtual assessments of active safety technologies

JIAN WU

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2024

Generation of Representative Pre-Crash Scenarios Across the Full Severity Range Using Real-World Crash Data:
Towards more accurate virtual assessments of active safety technologies
JIAN WU

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

Chalmers Reproservice
Göteborg, Sweden 2024

ABSTRACT

Virtual safety assessment is now the primary method for evaluating the safety performance of active safety technologies such as Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS), not the least because there are few alternatives. Generating representative crash scenarios is crucial for the assessment to produce valid results. However, the existing crash scenario generation methods face challenges such as limited and biased in-depth crash data and difficulties in validation. To meet these challenges, this thesis proposed a set of novel methods for generating representative synthetic crashes.

This thesis demonstrate the methods for a common crash type, the rear-end crash, in which the front of one vehicle collides with the rear of another. The process of generating synthetic rear-end crash scenarios consists of three main steps: 1) parameterizing the rear-end crashes by modeling the two involved vehicles using naturalistic driving and pre-crash kinematics data, 2) building multivariate distribution models for the parameterized crash data, and 3) generating representative synthetic crash scenarios.

Paper A utilized a piecewise linear model to parameterize the lead-vehicle speed profiles in rear-end crashes from two United States datasets. These parameterized speed profiles were then combined and weighted to create a comprehensive dataset representative of lead-vehicle kinematics in rear-end crashes across the full severity range, from physical contact to high severity. Synthetic speed profiles, generated using multivariate distribution models built on the dataset, were then compared with the raw profiles. The results show that the proposed lead-vehicle kinematics model accurately matches lead-vehicle kinematics in rear-end crashes across the full severity range, outperforming the conventional constant lead-vehicle acceleration/deceleration model in terms of both severity range and precision.

In Paper B, a following-vehicle behavior model was created by combining two existing driver behavior models. A representative dataset of the initial states (i.e., speeds of both vehicles and the following distance) of rear-end crash scenarios and the minimum accelerations of both vehicles was developed by weighting and combining crash data from various sources. The dataset was modeled to create a synthetic dataset with more samples. Crash scenarios were simulated based on this synthetic dataset, the following-vehicle behavior model, and the synthetic speed profiles from Paper A, creating a synthetic rear-end crash dataset. The dataset can be used for the safety assessments of ADAS and ADS and as a benchmark when evaluating the representativeness of scenarios generated through other methods.

Future work will aim to test ADAS and ADS with synthetic crash scenarios and validate existing crash scenario generation methods, especially those that are traffic-simulation-based.

Keywords: virtual safety assessment, crash scenario generation, pre-crash, data combination, data synthesis

Appended publications

This thesis consists of an extended summary and the following appended papers:

Paper A J. Wu, C. Flannagan, U. Sander and J. Bärgman (2024a). Modeling lead-vehicle kinematics for rear-end crash scenario generation. *IEEE Transactions on Intelligent Transportation Systems.* DOI: 10.1109/TITS.2024.3369097.

Paper B J. Wu, C. Flannagan, U. Sander and J. Bärgman (2024b). Model-based generation of representative rear-end crash scenarios across the full severity range using pre-crash data. *IEEE Transactions on Intelligent Transportation Systems (under review).*

Author's contribution

Paper A: wrote the first draft of the paper, designed the methods, and did most of the analysis.

Paper B: wrote the first draft of the paper, designed the methods, performed the simulations, and did most of the analysis.

Acknowledgements

I want to express my gratitude to my supervisors, Jonas Bärgman and Carol Flannagan, and my industrial supervisor, Ulrich Sander, for their inspiring guidance. I remember when we had the same thought and voiced it almost simultaneously. We also had moments of disagreement; fortunately, we never hesitated to express and challenge each other's ideas. You have all been outstanding mentors, research partners, and role models, and I am truly grateful for your support.

I want to thank András Bálint and Mikael Ljung Aust for their support and guidance at the beginning of my Ph.D. program. I am also grateful to Marco Dozza, my examiner, for his valuable insights into my research, and to my managers, Mattias Robertson and Ashok Chaitanya Koppisetty, for helping me balance my research and personal life more effectively.

Thanks to all my friends and colleagues at Chalmers and Volvo Cars for providing a fantastic working environment full of ideas.

Thanks to my family for their support and encouragement. Most importantly, I am grateful for my caring, wise, and farseeing wife, Yuan Liao, who is always there for me.

Malmö, June 2024
Jian Wu

Contents

CHAPTER 1

Introduction

1.1 Enhancing traffic safety through active safety technologies

Due to the rapid development of active safety technologies such as Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS), we are experiencing a transformation as vehicles from traditional manually operated machines to increasingly automated ones.

Systems ranging from adaptive cruise control (ACC) [1], lane keeping assist (LKA) [2], and automated emergency braking (AEB) [3] are all considered ADAS. These technologies help drivers control their vehicles by providing real-time alerts and interventions to prevent or mitigate potential collisions [4]. On the other hand, ADS take a significant leap forward by substantially reducing (and at the highest support level, eliminating) the need for human intervention, relying on advanced sensors and artificial intelligence (including machine learning) to navigate and operate vehicles autonomously [5–7].

By augmenting human driving abilities and introducing automation, these systems can address common causes of crashes, such as distracted driving [8], fatigue [9], and impaired driving [10]. Thus these technologies have the potential to substantially reduce the number of traffic accidents, injuries, and fatalities. In fact, the research has shown that ADAS systems such as LKA [11, 12] and AEB [13, 14] systems have already had a substantial positive impact on traffic safety.

1.2 Virtual safety assessments

It is essential to quantitatively assess the safety performance of ADAS and ADS, as the driving task is partially or completely transferred from the driver to the vehicle [15]. An assessment enables developers to thoroughly test the algorithms before deployment, ensuring their reliability and safety across various driving scenarios. Additionally, it can help policymakers and legislators prioritize the systems with the most substantial safety benefits and guide the widespread adoption of appropriate regulations and standards.

Virtual safety assessment is now the primary method for evaluating the safety performance of active safety technologies such as ADAS and ADS, not the least because there are few alternatives [16–20]. Typically, in a virtual assessment, a comparison of '*baseline*' and '*treatment*' scenarios is conducted. The baseline scenarios, without the technology being assessed, serve as a starting point for the simulations. (The same set of scenarios with the technology make up the treatment condition.) The baseline scenarios must match the assessment objective and include all relevant elements that may impact the performance of the technology under assessment [21].

Wimmer et al. [21] defined three main approaches to creating baseline scenarios (depending on the type of input data source, how the input data source is used, and how the data are processed):

- Approach A: Digitized real-world scenarios without modification.
- Approach B: Modified or varied real-world scenarios.
- Approach C: Synthetic scenarios.
  - Approach C1: A small number of scenarios covering a limited test space.
  - Approach C2: Many scenarios covering a large test space.

Approach A digitizes real-world scenarios without altering them, utilizing databases of recorded driving data or reconstructed crashes as sources. In Approach B, real-world scenarios are the basis as well, but the original data are modified (by altering existing properties or even adding new ones) to build the required baseline scenarios [21]. In Approach C, statistical information from real-world data is used to create synthetic cases instead of individual real-world scenarios. This approach has two variations, C1 and C2, which are distinguished mainly by the number of generated scenarios. Of the four approaches, C2 stands out: it can generate a large number of scenarios and cover a wide range of conditions, which is essential for making a statistically significant comparison between the baseline and treatment [21]. As a result, C2 is often preferred in virtual assessments [22, 23].

The traffic-simulation-based [24–26] and in-depth-crash-data-based (referred to as IDC-based) [27–32] methods are the two primary virtual assessment methods which assess the technology with a large number of scenarios; they both follow the C2 approach to create baseline scenarios. (Note that Approach B can also be IDC-based and generate many scenarios by manipulating individual original scenarios. But it is not included in this thesis.)

The traffic-simulation-based method simulates scenarios under assessment to create crash events in a (virtual) modeled driving environment [24, 26, 33, 34]. Typically, traffic simulation models for scenarios under assessment are built using naturalistic driving data (NDD) that contain a limited number of crashes, often of minor severity. To evaluate safety, simulations are often carried out over an extended period, measured in millions of simulation hours. Simulations are conducted with and without the specific ADAS or ADS, and the number of crashes experienced in each scenario is subsequently compared [34]. Sometimes, the outcome variables, such as Delta-v (i.e., the total change in vehicle velocity over the duration of the crash event) or injury risk, are also compared between the two conditions.

In contrast to the traffic-simulation-based method, the IDC-based method uses detailed crash information, which includes reconstructed or recorded data such as vehicle kinematics. Statistical distributions of the relevant crash attributes are created and then sampled to generate virtual crashes. Following this, a simulation is conducted for each generated crash that involves the ADAS or ADS under assessment to assess whether the crash can be avoided [27–32].

The two methods are compared with respect to the following four qualities (summarized in Table 1.1).

Data accessibility and sufficiency: Because the traffic-simulation-based method mainly builds simulation models using NDD, there is a wealth of information realatively easy to access.

On the other hand, the IDC-based method faces a challenge regard-

	Traffic-simulation-based	IDC-based
Data accessibility and sufficiency	High	Low
Efficiency	Low	High
Spatiotemporal continuity	Yes	No
Bias	Non-severe crashes	Severe crashes

Table 1.1: Comparison between the two methods.

ing data accessibility and sufficiency, as the safety assessments for ADAS and ADS using the IDC-based method generally require more real-world crash instances than are currently available [35]. The limited accessibility of comprehensive real-world crashes means that the full range of crashes within a specific scenario is not always fully represented—primarily due to privacy concerns and the high costs of data collection, storage, and dissemination. In such cases, synthetic crashes are required to 'bridge the gaps' between real crashes [36, 37]. In the C2 approach, synthetic crashes can be seen as interpolations or extrapolations of the original crashes.

Efficiency: The traffic-simulation-based method is highly inefficient since demonstrating the safety performance of autonomous vehicles requires hundreds of millions of miles due to the high dimensionality of the environment and the rarity of safety-critical events. To tackle this issue, Feng et al. [26] proposed a solution known as the naturalistic and adversarial driving environment (NADE), which introduces sparse but adversarial modifications to reduce the number of virtual test miles needed while maintaining unbiased assessments. However, even with the NADE technique, many test miles are still necessary.

In contrast, the IDC-based method is more efficient since it directly generates crash scenarios.

Spatiotemporal continuity: In the traffic-simulation-based method, the traffic simulator can simulate not only individual scenarios but also entire road networks comprising numerous road users, allowing the continuous spatiotemporal assessment of active safety technologies [15].

However, the IDC-based method, like all scenario-based methods, is spatiotemporally discontinuous. Because the scenarios are extracted from actual traffic, this method has the disadvantage of not capturing all the context [15].

Bias: The traffic-simulation-based method is generally biased towards non-severe crashes. Thus, using 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 their overall distribution. Olleja et al. [38] compared crash generation using normal driving data and near-crash incidents with crashes from in-depth crash databases. The results showed substantial disparities: normal driving data failed to reflect the crash outcomes and criticality observed in crashes. In addition, crashes generated by the traffic-simulation-based method rely heavily on multiple models of road-user behaviors, which must be accurate in order to produce realistic crashes (representing the real world). Formal validation of the details of the generated crashes is usually overlooked [15]: instead, visualization techniques, such as histograms [26, 36, 39] and scatter plot [40], are used to compare the generated and real-world crashes. Also, the validations is almost always not considering the outcome severity (e.g., Delta-v or injury risk).

The IDC-based method, on the other hand, may over-represent severe crashes. In traditional in-depth crash databases, selection criteria inherently introduce a bias towards severe crashes. For instance, the Crash Investigation Sampling System (CISS), a crash database in the United States, focuses on incidents with at least one light vehicle towed from the scene [41]. Relying solely on these databases to create synthetic crashes [36, 40, 42] can skew crash generation models, biasing the overall analysis towards severe crashes.

Additionally, whether pre-crash data accurately reflect real-world scenarios depends on how they were produced. Generating crashes using reconstructed crash data can be problematic. While crash outcomes like Delta-v might be reasonably accurate, reconstructed pre-crash kinematics are heavily influenced by the reconstruction software and assumptions about the behaviors of the involved road users made during reconstruction, especially when detailed pre-crash recordings are unavailable. Thus, the generated crashes might depend more on assumptions and software than on the pre-crash kinematics in real-world crashes [42].

Lastly, generating crashes at the tails of distributions can be challenging. For instance, Wang et al. [40] used independent component

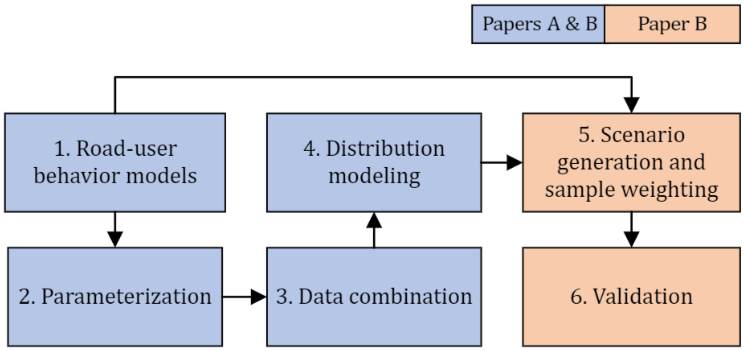


Figure 1.1: Illustration of the method in this work. Colors indicate which parts are covered by which paper.

analysis (ICA) followed by kernel density estimation (KDE) to generate synthetic crashes. However, the KDE distribution modeling method can introduce biases, particularly near boundaries and in distributions with long tails [43].

1.3 Aims and scope

This thesis aims to address the mentioned challenges for the two methods by creating synthetic crashes that are representative of real-world crashes across the full severity range, from physical contact to high severity. This work can thus improve the accuracy of virtual assessments of active safety technologies.

To achieve this, this thesis proposes a novel and comprehensive method combining naturalistic driving and pre-crash kinematics data. The NDD are used to obtain the distribution of crash severity levels (indicated by Delta-v). This distribution is then used to mitigate bias when combining naturalistic driving and pre-crash kinematics data. In addition, both reconstructed and recorded pre-crash kinematics data are used to reduce the influence of the reconstruction software and the assumptions made during reconstruction. the assumptions made during reconstruction. The method consists of six parts (see Figure 1.1):

1. **Road-user behavior models**: Create road-user behavior models for road users involved in a particular crash scenario.
2. **Parameterization**: Parameterize the crash scenario into a multi-

dimensional vector.

3. **Data combination**: Combine and weight naturalistic driving and pre-crash kinematics data to create datasets of the parameterized crashes representative of reality (referred to as the '*reference datasets*').

4. **Distribution modeling**: Build multivariate distribution models for the reference datasets.

5. **Scenario generation and sample weighting**: Generate synthetic crash scenarios using the built distribution models and the road-user behavior models, then weight the generated crashes to match the reference datasets.

6. **Validation**: Validate the synthetic crash scenarios by comparing them with the reference datasets.

This thesis demonstrates the utility of the proposed method for generating synthetic scenarios of a common crash type: the rear-end crash, in which the front of one vehicle collides with the rear of another. The rationale for choosing this scenario is twofold. Firstly, rear-end crashes account for 27.8 percent of all reported car accidents in the United States in 2020 [44]. This high frequency means it is imperative to understand how ADAS and ADS handle this scenario. Secondly, a rear-end crash is relatively simple since it mainly involves the longitudinal maneuvers of two vehicles (lead and following).

Specifically, this thesis addresses the issue of synthetic rear-end crash scenario generation through two scientific publications:

- In **Paper A**, we developed a lead-vehicle kinematics model, built a reference dataset of parameterized lead-vehicle speed profiles, and created a synthetic lead-vehicle speed profile dataset.
- In **Paper B**, we developed a following-vehicle behavior model, parameterized rear-end crashes, built reference datasets of the parameterized rear-end crashes, and eventually generated a synthetic rear-end crash dataset matching the reference datasets.

In summary, this thesis addresses the following **research questions**:

- How can rear-end crashes be parameterized?
- How can parameterized rear-end crash data be used to build reference datasets?

- How can representative synthetic rear-end crashes be generated?

CHAPTER 2

Methodology

This chapter introduces the data, road-user behavior models, and key methods used to generate representative rear-end crash scenarios in this thesis.

2.1 Data

The data used are from three sources: the Crash Investigation Sampling System (CISS) [41], the Second Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS) [45], and the German In-Depth Accident Study (GIDAS) Pre-Crash Matrix (PCM) [46].

CISS is a dataset representing a comprehensive study of car crashes across the United States. It focuses on incidents in which at least one light vehicle was towed from the scene [41]. The dataset is derived from thorough crash investigations that provide a detailed understanding of damaged vehicles, crash sites, and the complex dynamics of each incident. Its aim is to uncover the numerous elements that contribute to car crashes across the United States; several studies have used CISS data for various applications [11, 47, 48].

One essential aspect of CISS's methodology is extracting and integrating data from Event Data Recorders (EDRs) [49], electronic devices embedded in vehicles that record data about a vehicle's movement and the actions of its driver. Integration of EDR data collected before a crash provides a nuanced and data-rich perspective on the events leading up to it. Including EDR data increases the level of detail in CISS and provides information on pre-crash vehicle dynamics.

However, EDR data are only available for a subset of crashes, and there are even fewer crashes in which EDR data are available for both vehicles. For instance, of the 1,125 rear-end crashes in the CISS dataset, only 10.0% (113) contain EDR data for both vehicles. This issue restricts the ability of CISS EDR data to describe the kinematics of both vehicles. Another issue is that most EDRs collect pre-crash data at a relatively low frequency, ranging from 1 to 10 Hz. Most cases have a frequency lower than 5 Hz; only 0.4% (five) have a frequency of 5 Hz or higher. The limited temporal resolution makes it difficult to capture rapid changes in vehicle dynamics and driver behavior leading up to a crash. As a result, the ability to create detailed reconstructions of the pre-crash kinematics is compromised, potentially limiting the depth of analysis and the insights that can be derived from the data.

The second source, the SHRP2 NDS in the United States, has also delivered significant insights into pre-crash dynamics. As part of the broader SHRP2 program, the SHRP2 NDS involved equipping more than 3,300 passenger vehicles with a sophisticated Data Acquisition System (DAS) to capture a wide range of pre-crash information, including four different video perspectives: the driver's facial expressions and hand movements, the forward roadway view, and the rear roadway view [45].

From 2010 to 2013, the SHRP2 NDS collected data from participants' instrumented vehicles in six locations across the US. The dataset is extensive, providing a broad representation of naturalistic driving scenarios which covers incidents of varying severity—ranging from near-crashes to low-severity crashes and even (a few) high-severity crashes. The study used event identification algorithms to systematically identify incidents within the dataset and a meticulous manual annotation process to classify these instances into different severity levels [45].

The German In-Depth Accident Study (GIDAS), one of the most detailed and widely recognized in-depth crash datasets globally, gathers on-scene accident cases with personal injury in Hannover and Dresden. It provides a comprehensive view of accident sequences and causation through meticulous on-scene investigation and full accident reconstruction [50]. The GIDAS PCM dataset, initiated in 2011, is a subset of the GIDAS dataset, containing all relevant data in a database format to simulate the pre-crash phase until the first collision of the accident for a maximum of two participants [46]. The dataset includes the definition of the participants and their characteristics, the dynamic behavior of the participants as a time-dependent

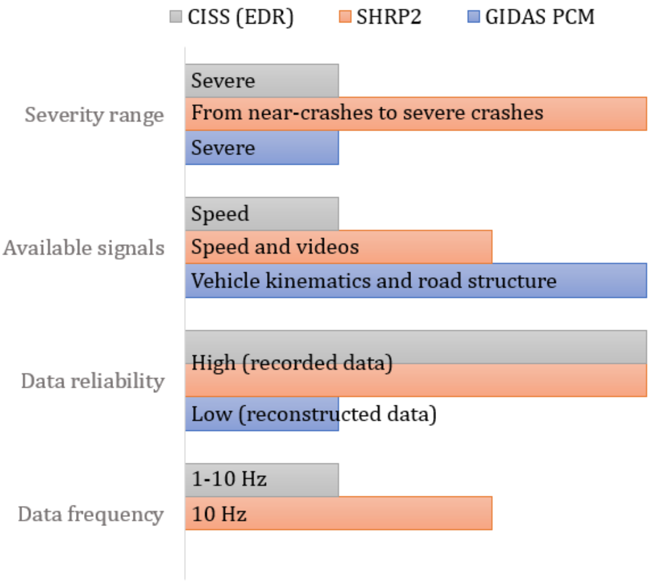


Figure 2.1: Comparison between CISS (EDR), SHRP2, and GIDAS PCM.

course for five seconds before the crash, as well as the geometry of the traffic infrastructure [46]. The GIDAS PCM dataset is widely used to analyze the causes and consequences of road traffic accidents [51, 52], assess the effectiveness of active safety technologies [53–55], and develop strategies for preventing accidents [56].

It is important to note that the data in the GIDAS PCM dataset are reconstructed based on evidence from the accident site and eyewitness accounts. As mentioned in Chapter 1.1, the reconstructed data may not be as dependable as recorded data. Therefore, instead of the entire speed profiles from the GIDAS PCM dataset, the minimum accelerations of both vehicles were used.

Figure 2.1 compares the three data sources: CISS (EDR), SHRP2, and GIDAS PCM.

- **Severity range**: In this thesis, the severity level does not correspond with the Abbreviated Injury Scale [57]. A crash is indexed as '*Severe*' if it fulfills the SHRP2 severity level I definition (i.e., a crash that involves an airbag deployment, injury to the driver, pedal cyclist, or pedestrian, vehicle rollover, high Delta V, or requires vehicle towing) and '*Non-severe*' otherwise. In addition, the severity level of any near-crash is designated as '*None*'. CISS specifically targets crashes with at least one light vehicle towed away from the scene, whereas GIDAS PCM focuses on crashes resulting in personal injury. Therefore, CISS and GIDAS PCM datasets contain only severe crashes. Unlike CISS and GIDAS PCM, SHRP2 does not filter or censor crash data, including all crashes in the instrumented fleet of vehicles, from the lowest possible to the most severe. Consequently, the SHRP2 dataset covers near-crashes to severe crashes but has few severe crashes.

- **Available signals**: CISS EDR data contains only the speed signal of the subject vehicle (i.e., the vehicle that recorded the data), SHRP2 data contains the speed and four different video perspectives of the subject vehicle, while the GIDAS PCM dataset contains not only the kinematics of all vehicles involved in the crash but also road structure information.

- **Data reliability**: CISS EDR and SHRP2 data are considered reliable as they were directly recorded. Meanwhile, the reliability of the GIDAS PCM's pre-crash kinematics data is less certain, as they were reconstructed based on scene investigation and certain assumptions.

- **Data frequency**: CISS EDR data ranges from 1 to 10 Hz, and most cases have a frequency of less than 5 Hz. In contrast, SHRP2 and GIDAS PCM data have a frequency of 10 and 100 Hz, respectively. However, because the GIDAS PCM is reconstructed, the frequency is "artificial": any sampling frequency could have been chosen.

Rear-end pre-crash data (five seconds before the crashes) from the three data sources were extracted, including the time-series data of the (longitudinal) distance between the lead and following vehicles and/or the speeds for the two vehicles (as available). The advantages and disadvantages of these data sources are described in the data combination step (see Chapter 2.4).

2.2 Road-user behavior models

Road-user behavior models play a critical role in virtual safety assessments. These models aim to replicate the behavior or kinematics of various road users, including drivers, pedestrians, and cyclists, in vir-

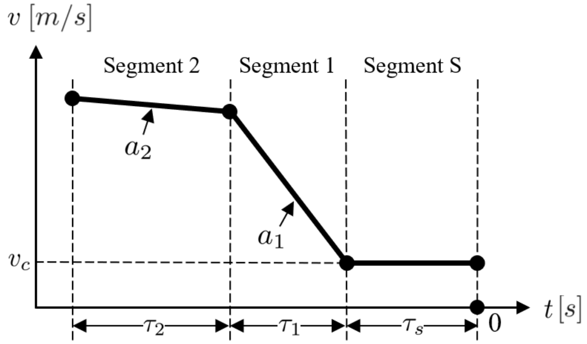


Figure 2.2: Three selected segments.

tual environments [58]. This thesis focuses on vehicle behavior models for longitudinal control of cars because a rear-end crash mainly involves the longitudinal maneuvers of the following and lead vehicles.

In virtual safety assessments, the vehicle behavior models can be used to simulate behaviors of both the ego and surrounding vehicles [58]. Existing models (including car-following [59–63], lane-changing [64–66], and cognitive models [23, 67, 68]) are typically expected to simulate realistic driving scenarios and interactions with various road users. Furthermore, the models sometimes may also need to consider different driver characteristics and cognitive capabilities to simulate different driving styles and possible human errors while driving [58].

Two vehicle behavior models were developed: a **lead-vehicle kinematics model** (presented in Paper A) and a **following-vehicle behavior model** (presented in Paper B).

The lead-vehicle kinematics model

In the rear-end crash scenario, the lead vehicle is mostly independent of the following vehicle. Thus, the lead-vehicle speed profile was modeled in Paper A without considering the vehicle's interaction with the following vehicle [35].

A **piecewise linear model** was used to simplify the lead-vehicle kinematics (i.e., speed profile) as several consecutive straight lines. Figure 2.2 shows an example with three segments (backward from time zero, defined as the impact moment). Unlike the conventional constant acceleration/deceleration model [69–71], in which the lead vehicle keeps a constant speed for a while and then transits to a constant acceleration/deceleration until the crash occurs or it comes to a complete stop, this model accurately matches the lead-vehicle speed profiles in rear-end crashes [35].

The following-vehicle behavior model

The following-vehicle behavior model was created by merging two existing driver behavior models:

1) **the Modified Intelligent Driver Model** [63]: This time-continuous car-following model is often used in traffic flow modeling. It computes a vehicle's acceleration based on its current velocity, maximum acceleration, desired velocity in free traffic, and the following distance from the lead vehicle.

2) **the Driver Pre-Crash Brake Response Model** [72]: This driver model quantitatively predicts when and how the driver will initiate and modulate a pre-crash brake response, considering the driver's off-road glance behavior. It uses the accumulation of the prediction error of looming (i.e., the relative expansion rate of the lead vehicle's image on the retina of the following vehicle's driver [73]) as the basis for the driver's braking response.

In addition to merging these two models, we adapted the resulting model to account for a particular type of rear-end crash in the GIDAS PCM dataset, which occurs in 9.2% of all crashes. In these, both vehicles were initially stationary; after a while, the following vehicle accelerated until it hit the lead vehicle. The driver of the following vehicle seemed to ignore the lead vehicle completely, possibly due to distraction. The new model includes the possibility of generating this 'abnormal' driver acceleration behavior.

2.3 Parameterization

Parameterization is a fundamental aspect of modeling and analysis across various disciplines [74]. It involves defining parameters within a system, model, or function to describe its behavior, characteristics, or properties. Parameterization enables models to be flexible, interpretable, and adaptable to different scenarios [75]. It allows model calibration, sensitivity analysis, and simulation, facilitating accurate predictions about and insights into complex phenomena. Overall, parameterization is essential for developing effective models that capture the complexity and variability of real-world systems [76].

Parameters for the two new models were established (six for the lead vehicle model and four for the following vehicle model). The only remaining elements in a rear-end crash scenario are two initial states: the following vehicle's speed and the initial following distance. Consequently, a rear-end crash was parameterized as a twelve-dimensional vector. (See Section III-B in Paper B for further information.)

2.4 Data combination

This step aims to create a reference dataset, representative of real-world crashes, for the parameterized rear-end crashes. Creating it directly from a single crash data source would be ideal. However, none of the available datasets alone could serve as a reference dataset since they all have limitations, such as bias and censoring, incomplete parameters, and low data quality. Consequently, it was not feasible to create a reference dataset of all twelve parameters directly from the available datasets. Therefore, several reference datasets of subsets of parameters were created by combining data from multiple sources as an intermediate step before building the final reference dataset of all parameters. (See Section III-C in Paper B for further information.)

Data combination [77] is the process of merging or integrating multiple datasets into one. This process, commonly used in fields such as statistics, data analysis, and machine learning, leverages information from different sources to optimize the resulting dataset's comprehensiveness and utility. It requires careful consideration of data quality, compatibility, and consistency to ensure that the combined dataset accurately reflects the underlying phenomena and supports valid analysis and interpretation.

Raw or combined raw datasets can often be biased in one or more parameters (especially the crash severity level in our case) [35]. The bias can be mitigated if the reference marginal or joint distribution of all or a subset of those parameters is available. This can be achieved by weighting the samples in the raw dataset so that the weighted data matches the known reference distribution [78]. The weighted dataset can be considered as the reference dataset of its contained parameters.

As noted, the SHRP2 dataset contains collected crash data from physical contact to a few with high severity, whereas the other data

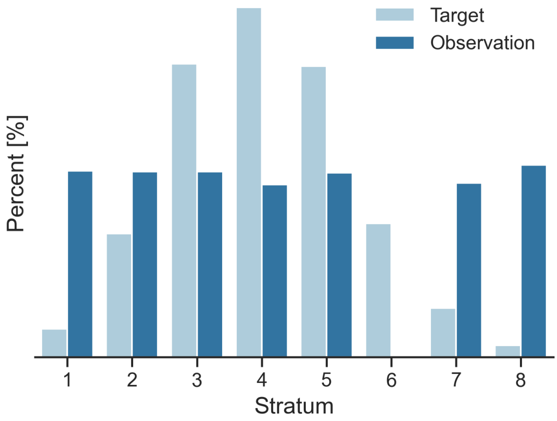


Figure 2.3: A binning example for unidimensional data. Strata 6 is the omitted stratum.

sources are biased towards severe crashes [41, 46]. Therefore, the distribution of the lead vehicle's Delta-v for SHRP2 rear-end crashes was used as the reference distribution for crash severity (i.e., the distribution representing severity levels of real-world rear-end crashes, from physical contact to high severity). The pre-crash data from the CISS and GIDAS PCM datasets were merged with the SHRP2 crash dataset to obtain more detailed information on severe crashes.

Data combination, including a sample-weighting process, was conducted to gather as much information as possible while leveraging the strength (and minimizing the bias) of each dataset, in order to create reference rear-end crash datasets across the full range of severity.

Post-stratification weighting

Since raw distributions from datasets can often be biased with respect to one or more parameters, the sample weighting process aims to assign weights to samples in raw distributions so that the weighted data match the known reference distribution of a subset of parameters.

Post-stratification weighting [78] is one such process. It is a statistical technique commonly used in survey research to reduce bias and improve the accuracy of population estimates. It involves dividing the target population into strata based on certain characteristics or variables, collecting data within each stratum, and then assigning weights (based on the target population distribution within each stratum) to the observations.

Typically, binning is used to create strata when the variables are continuous [79]. The weight assigned to observations in each stratum is calculated by dividing the target population total by the number of observations in the stratum. The raw samples (i.e., observations) are grouped into discrete bins based on the known reference distribution (i.e., target population). This method assumes that no strata are omitted, meaning that observations must cover all bins spanned by the target population. Otherwise, the process will attempt to divide by zero. However, in our case, omitted strata did exist in the combined data. (See Figure 2.3 for an example: Stratum 6 contains no observations.) Therefore, a novel method, the **k-nearest neighbors (KNN) sample weighting method**, was proposed to handle this issue. This method can be seen as a post-stratification weighting method with a dynamic binning strategy. Each sample extracted from the known reference distribution carries a weight of one. For each extracted sample, the k-nearest raw samples are grouped into one bin to share the weight based on their distance from the extracted sample. It is also worth mentioning that samples that have never been selected as the nearest neighbors of any extracted sample will have a weight value of zero.

This thesis also investigated another option: treating the parameters not included in the known reference datasets as 'missing' and imputing them using a data imputer based on the observations (i.e., the combined data). A classic data imputation method, the KNN data imputation method [80], was chosen to handle data with omitted strata. An empirical simulation study was carried out to compare the performances of the two options. The results demonstrate the advantages of the KNN sample weighting method. (See the Appendix for further information.)

2.5 Distribution modeling

Through the process of data combination, two reference datasets were created, each encompassing multiple parameters. One contains the lead-vehicle speed profiles, and the other comprises the initial states and the minimum accelerations for both vehicles. However, the two datasets do not contain enough samples to conduct virtual safety assessments of ADAS and ADS. Therefore, synthetic crash scenarios, essentially interpolated or extrapolated from actual crashes, are required to bridge the gaps. To generate synthetic data, the reference data must be modeled. The resulting model(s) can then be used to generate any number of synthetic crashes.

Parametric [81, 82], **non-parametric** [83, 84], and **copula-based** [85–88] methods are commonly used for modeling a multivariate distribution.

The parametric methods assume a specific parametric form for the joint distribution, such as the multivariate normal distribution [81], which can be advantageous when the underlying distribution is known or can be reasonably assumed. They often have fewer parameters than other methods, making them computationally efficient and less prone to overfitting when dealing with sparse data. However, they rely on strong assumptions about the underlying distribution, which might not hold in real-world scenarios. If the assumptions are violated, the parametric models can provide biased estimates. In addition, if the available data are sparse, they might not provide enough information to accurately estimate the chosen distribution's parameters.

Non-parametric methods, such as kernel density estimation [89] or nearest neighbor methods [90], make minimal assumptions about the underlying distribution and instead estimate it directly from the data. Nonparametric methods might require more data to achieve reliable estimates than parametric methods. They may not always provide accurate density estimates, especially for sparse or irregularly sampled data [91]. In addition, non-parametric methods can produce biased estimates, particularly near boundaries and in distributions with long tails [43]. Finally, interpreting the results can be more challenging (compared to parametric methods) since the distribution does not have an explicit functional form.

Copula-based methods allow the flexible modeling of dependence structure between random variables; they are often used when the marginal distributions are known or estimated separately [88]. However, the copula-based methods usually require a relatively large amount of data to accurately estimate the parameters of the copula function, especially for high-dimensional datasets [86]. These methods are also less easily interpreted than simpler parametric models.

In our situation, the amount of data available is limited, and some parameters are sparsely represented. Additionally, it is crucial to be able to interpret the distribution models and grasp the correlations among parameters, as they hold physical significance. Therefore, a **parametric multivariate distribution modeling method** was developed, which models the marginal distributions and the linear correlations among parameters.

The method first assesses the linear correlation between each pair of parameters using the **Pearson correlation coefficient** [92], which measures the strength and direction of a linear relationship between two variables. A high absolute coefficient shows a strong relationship. The sign indicates whether the relationship is positive or negative.

A parameter is categorized as **correlated** if it exhibits a **significant and non-weak correlation** with any other parameter and **uncorrelated** otherwise. Here, a significant and non-weak correlation is defined as one whose Pearson coefficient has a p-value less than 0.05 (significant) and an absolute value greater than or equal to 0.3 (non-weak) [35, 93, 94].

Finally, each uncorrelated parameter's data are fitted into a set of known distributions (such as normal and gamma distributions), and the best-fit distribution with the lowest Akaike information criterion value is selected. In contrast, the correlated parameters are modeled as a multivariate normal distribution. Additionally, the parametric method gives special consideration to the *point-mass mixture distribution parameters*. A point-mass mixture distribution parameter is a parameter that contains a point-mass (a particular value with more observations than a continuous distribution can describe), thus requiring a mixture distribution model to describe its distribution [35]. This method is explored further in the discussion chapter. (See Section III-C in Paper A for additional information regarding the distribution modeling method.)

Moreover, because of the large number of parameters and the complexity of the multivariate distributions (such as various patterns in the data), it is difficult to build a single multivariate distribution model that covers all the data. On the other hand, creating sub-datasets allows a simpler model to be applied to each. Therefore, each reference dataset was categorized into several sub-datasets modeled separately using the proposed parametric multivariate distribution method. (See Section IV-C in Paper A and Section IV-B in Paper B for further information regarding the categorization.) Then, for each reference dataset, the overall distribution model (which can be seen as a mixture distribution model) was derived by combining the distribution models for all sub-datasets according to their relative proportions in the reference dataset.

2.6 Crash scenario generation

Based on the mixture distribution models built for the two reference datasets in the previous step and the following-vehicle behavior model, a **matching algorithm** was used to:

1. Create various synthetic kinematic parameter configurations, essentially pairing the synthetic lead-vehicle speed profile from one mixture distribution model with the synthetic initial states and the minimum accelerations for both vehicles from the other.

2. Conduct simulations under those synthetic kinematic parameter configurations and search for **valid simulations** (described below) to create a set of synthetic rear-end crashes.

Not all of the synthetic crashes were valid. A valid simulation was characterized by 1) the occurrence of a crash and 2) the crash occurring approximately five seconds after the start of the simulation. (Recall that all of the pre-crash data in the original datasets were captured starting five seconds before the crash; see Section III-E in Paper B for further information.)

The matching algorithm generated 5,000 synthetic crashes. To ensure that these synthetic crashes accurately represent real-world occurrences, the synthetic dataset should be weighted so that it matches the reference datasets of all parameters.

The **Iterative Proportional Fitting (IPF)** [95–97] procedure can be applied in this situation. IPF is a statistical method used in many disciplines to adjust multi-dimensional or contingency tables to satisfy known marginal totals [95]. It is commonly employed when the joint distribution of variables is unknown, yet the marginal distributions are known or estimated [96]. It typically contains four steps:

1. **Initialization**: Start with an initial table of observed counts, which may not necessarily match the desired marginal totals.

2. **Iteration**: Iteratively adjust the cell values to minimize the discrepancy between the observed cell counts and the desired mar-

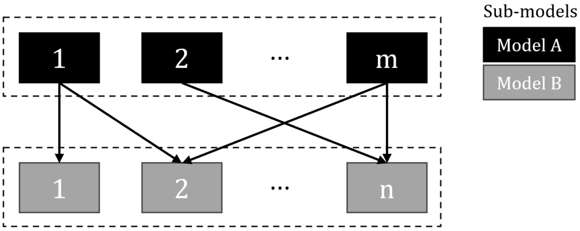


Figure 2.4: Illustration of pairing sub-models between the two mixture distribution models A and B. There are m and n sub-models in Models A and B, respectively.

ginal totals while maintaining the relative proportions within each dimension of the table.

3. **Convergence**: Stop iterating when a stopping criterion is met, such as a specified number of iterations or the cell values converge.

4. **Output**: Once the iterations have stopped, the adjusted table of counts represents a solution where the observed marginal totals match the desired totals as far as possible. This will typically not be an exact match [98].)

In this thesis, an **IPF-based sample weighting method** was used to weight the synthetic dataset to match the reference datasets. It is important to note that when modeling each reference dataset, the entire dataset was split into multiple sub-datasets so that simpler models could be employed for each sub-dataset. The overall distribution model was then derived by combining the distribution models for the sub-datasets according to their proportions in the corresponding reference dataset. Therefore, the weighting method must align each parameter's weighted marginal distribution with its reference marginal distribution for each sub-dataset.

Additionally, as illustrated in Figure 2.4, when pairing parameters drawn from the two mixture distribution models to form the kinematic parameter configurations in the synthetic crash dataset, a parameter combination from one sub-model of the first mixture distribution model (Model A) can be paired with parameter combinations from multiple sub-models of the second mixture distribution model (Model B). Ideally, the IPF procedure stops when the cell values converge; however, because the sub-datasets may not be isolated, reaching this point may be challenging. Therefore, the IPF procedure was designed to stop when the predefined maximum number of iterations was reached. A **loss function**, computed after every iteration, was used to select the optimal weighting result with the minimum loss value.

The loss function was designed based on the **Kolmogorov–Smirnov (KS)** statistics [99] (i.e., the largest absolute difference between two compared cumulative distribution function curves) for all parameters in each sub-dataset. It indicates the overall difference between the weighted synthetic crash dataset and the reference distributions. Here, the KS statistic was used to measure the difference between the weighted marginal and reference distributions for each parameter in each sub-dataset. (See Section III-E in Paper B for further information.)

2.7 Validation

We must ensure that the weighted synthetic rear-end crash dataset is similar to the reference datasets to validate its representativeness. Because the datasets are multidimensional, various aspects of their characteristics must be compared. Below are some of the most commonly used methods for this purpose:

- **Descriptive statistics**: Compare summary statistics such as means, medians, variances, and percentiles across the two datasets to assess their overall distributional similarity [100].
- **Visualization techniques**: Utilize histograms [26, 36, 39], scatter plots [40], and dimension reduction (such as **t-distributed stochastic neighbor embedding (t-SNE)** [100, 101] and **uniform manifold approximation and projection (UMAP)** [102]), to visualize the distributions and relationships within each dataset and identify any discrepancies or similarities.
- **Correlation analysis**: Calculate correlation coefficients between corresponding variables in the two datasets to measure the strength and direction of their linear relationships [40].
- **Statistical tests**: Perform statistical tests to formally compare the distributions of variables or features in the two datasets [103].

Statistical testing is unique among these methods as it formally

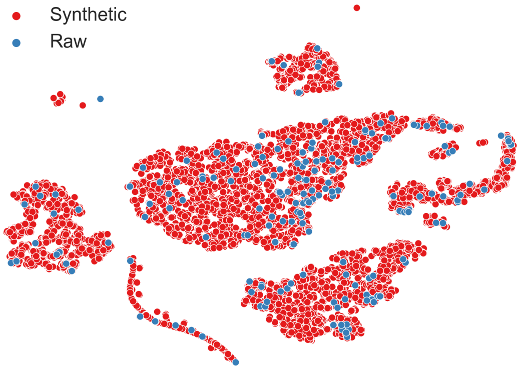


Figure 2.5: t-SNE projection of the raw and synthetic lead-vehicle speed profiles. The blue dots (projection of raw speed profiles) are surrounded by the red dots (projection of synthetic speed profiles).

compares two datasets. Two examples are the two-sample KS and **Anderson-Darling (AD)** tests; they are non-parametric tests commonly used to formally assess whether there is a significant difference between two distributions/datasets [103]. Both tests are based on empirical cumulative distribution functions. They are sensitive to differences in both the location and shape of the empirical cumulative distribution functions of the two samples [104–106]. In addition, the AD test emphasizes differences in the distribution's tails and is generally statistically more powerful than the KS test [103]. However, the former is also more sensitive to ties in the data (i.e., data with the same value), especially in the tails.

Both the synthetic and reference datasets are weighted data, which means that ties exist in each data point—largely increasing the possibility of the AD test rejecting the null hypothesis. Therefore, the weighted two-sample KS test was chosen.

However, the KS test faces limitations in handling high-dimensional data, such as computational complexity and increased sample size requirement [107, 108]. As the datasets being compared are high-dimensional, the KS testing was complemented with a dimensionality reduction technique (t-SNE), which was applied to transform the high-dimensional data into unidimensional data.

In this thesis, in addition to the commonly used methods, including general descriptive statistics (comparing means and variances of parameters) and a visualization technique (an example is shown in Figure 2.5), the synthetic and reference datasets were further compared through statistical tests from the following three perspectives:

1. **Marginal distributions**: For each parameter in each sub-dataset, the weighted two-sample KS test using the ”Ecume” package in R [109] was carried out to compare the weighted marginal and corresponding reference distributions (at the 0.05 significance level).

2. **Multivariate distributions**: For each sub-dataset containing multiple parameters, the t-SNE technique was applied to transform the multidimensional data into unidimensional data. Then, the weighted two-sample KS test on the transformed data was conducted to test whether the synthetic and reference data were significantly different (at the 0.05 significance level).

3. **Crash severity level distribution**: The weighted two-sample KS test was conducted to test whether the severity level (indicated by the lead vehicle’s Delta-v) distributions for the reference and synthetic datasets are significantly different (at the 0.05 significance level).

CHAPTER 3

Present Work

Paper A: Modeling Lead-Vehicle Kinematics For Rear-End Crash Scenario Generation

Introduction

The use of virtual safety assessment as the primary method for evaluating Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS) 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.

Methods

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 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, which were compared with the raw speed profiles.

Results

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 that the proposed piecewise linear model is more accurate than the conventional constant acceleration/deceleration model. For example, the lead vehicle could exhibit harsh braking followed by gentle braking or even acceleration; however, it 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.

Conclusions

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

Paper B: Model-Based Generation of Representative Rear-End Crash Scenarios Across the Full Severity Range Using Pre-Crash Data

Introduction

Generating representative rear-end crash scenarios is crucial for safety assessments of ADAS and ADS. However, existing methods face challenges such as limited and biased in-depth crash data and difficulties in validation. This study sought to overcome these challenges by combining naturalistic driving data and pre-crash kinematics data from rear-end crashes to create a representative distribution of rear-end crash scenarios in the United States. The resulting distribution can be used not only for safety assessments of ADAS and ADS, but also as a benchmark when evaluating the representativeness of scenarios generated through other methods.

Methods

The process of generating synthetic rear-end crash scenarios consists of three steps: 1) parameterizing the rear-end crashes through modeling the two involved vehicles, 2) building distribution models for the parameterized crash data, and 3) generating representative synthetic crash scenarios.

In the first step, a following-vehicle behavior model was developed by combining two existing driver models. Combining this model, the lead-vehicle kinematics model (created in a previous study [35]) and the initial states of rear-end crash scenarios created a twelve-dimensional vector representing a rear-end crash.

In the second step, parameterized crash data from multiple crash datasets were combined and weighted to create a reference dataset of the initial states (and minimum fitted accelerations of both vehicles) (REF_b). A synthetic dataset containing these data (REF_sb) was then created by sampling from the distribution model built for REF_b.

Lastly, simulations were conducted using the following-vehicle behavior model and the two synthetic datasets, REF_sb and REF_sl (a representative synthetic rear-end crash lead-vehicle speed profile dataset created in a previous study [35]). valid simulations were gathered and weighted using an IPF-based weighting algorithm to create a representative synthetic rear-end crash dataset.

Results

Sixty-one weighted two-sample KS tests were conducted to compare the synthetic crash dataset with reference datasets. The only two that showed a significant difference at the 0.05 significance level did so because the following vehicle's acceleration model could not imitate aggressive acceleration under certain situations. Comparing the weighted datasets for Δv_l showed no significant differences.

Conclusions

None of the available crash datasets in this study contain all necessary signals or are free of significant bias; these issues are commonly encountered in data-driven studies. To resolve them, we proposed a set of methods to combine and weight data from multiple crash datasets. These methods create a reference dataset of the initial states and minimum accelerations of the following and lead vehicles across the full severity range. Moreover, the data were weighted to match a reference dataset using the KNN sample weighting method to reduce bias. This method is particularly noteworthy because, unlike conventional post-stratification methods, it can be used to weight biased data to match a reference dataset even when omitted strata exist. The data combination methods we propose can also be applied to other situations with biased datasets and/or incomplete signals which require a multivariate joint distribution.

The created representative rear-end crash dataset can be used for the safety assessments of ADAS and ADS. Moreover, it can serve as a benchmark when evaluating the representativeness of scenarios generated through other methods.

CHAPTER 4

Discussion and Conclusions

Generating representative crash scenarios is crucial for evaluating active safety technologies using the virtual safety assessment method. However, existing crash scenario generation methods face challenges such as limited in-depth crash data, inefficiency, and biased outcomes [26, 35, 36, 39, 41, 46].

To address these challenges, this thesis proposed a set of novel methods that combine naturalistic driving and pre-crash kinematics data, build parametric multivariate distribution models for the combined data, and generate synthetic crashes using the synthetic data generated from the built distribution models. This thesis answers the three research questions posed at the start. For each research question, a concise explanation is provided regarding why it was asked, along with a brief summary of the steps taken to address it and a high-level overview of the results.

How can rear-end crashes be parameterized?

In this thesis, behavior models of the two involved vehicles (following and lead) were developed first. Then, combining the two models and the initial states of rear-end crash scenarios created a twelve-dimensional vector representing a rear-end crash.

The most important work in the parameterization was the two vehicle behavior models. Accurate vehicle models are essential for creating realistic crash scenarios which ensure reliable and valid simulation outcomes. In a rear-end crash scenario, the lead vehicle is mostly independent of the following vehicle, while the following vehicle responds to the presence and actions of the lead vehicle. Consequently, a lead-vehicle kinematics model and a following-vehicle behavior model were developed. Below is a further discussion of the two models.

Many studies used a driver response model [69–72, 110] to analyze the following vehicle's behavior during rear-end emergencies (crashes and near-crashes). The following-vehicle behavior model in this thesis was developed by combining two existing driver behavior models, the modified Intelligent Driver Model [63] and the driver pre-crash brake response model [72]. Moreover, as mentioned in Chapter 2.2, the behavior model was adapted to include the possibility of generating the 'abnormal' driver acceleration behavior, which was observed in 9.2% of all crashes.

On the other hand, there has been a notable lack of research on the lead vehicle's behavior, despite its significant influence on the following vehicle. In the crash reconstruction and rear-end emergency studies, it is conventionally assumed that the lead vehicle keeps a constant speed for a while and then transits to a constant acceleration or deceleration until the crash occurs or it comes to a complete stop [69–71]. In addition, a rapid, large deceleration is typically assumed when the lead vehicle decelerates. This conventional constant acceleration/deceleration model is usually a result of limited information about the pre-crash phase; there is no evidence in the literature for the validity of such a model.

Paper A addresses this knowledge gap by thoroughly investigating the lead-vehicle kinematics in rear-end crashes. A piecewise linear model was used to parameterize the lead-vehicle speed profiles five seconds before impact for two US rear-end pre-crash/near-crash datasets. The parameterized data were combined and weighted to create a reference dataset of lead-vehicle kinematics in rear-end crashes across the full severity range, from physical contact to high severity.

Figure 4.1 shows three lead-vehicle speed patterns from the reference dataset. Within the increasing acceleration pattern, the lead vehicle could brake harshly, followed by gentle braking or even acceleration. More importantly, unlike what has been conventionally assumed, the lead vehicle does not necessarily brake harshly, even in severe crashes. In fact, as shown in Section III-C of Paper B, there is only a weak (linear) correlation between the lead vehicle's Delta-v (which is an indicator of crash severity) and minimum acceleration (i.e., maximum deceleration).

Compared with the conventional constant acceleration/deceleration

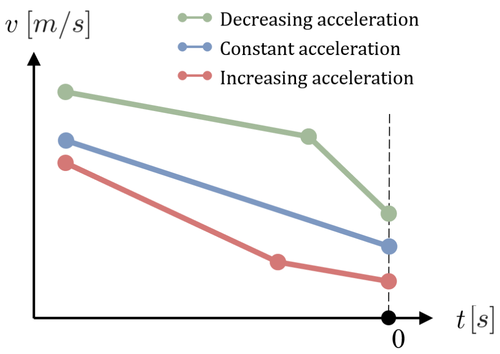


Figure 4.1: Three lead-vehicle speed patterns.

model, the proposed lead-vehicle kinematics model more accurately captures those patterns, better matching the lead-vehicle speed profiles in rear-end crashes across the full severity range. As a result, this new model lays the foundation for generating more realistic rear-end crash scenarios. Furthermore, it can facilitate the reconstruction of actual crashes by providing a set of more realistic speed profiles for the lead vehicle based on available information.

How can parameterized rear-end crash data be used to build reference datasets?

The obtained parameterized rear-end crash data were from three in-depth crash datasets. A comprehensive data combination method was proposed in this thesis to combine these crash datasets with a sample weighting process, creating reference datasets of the parameterized rear-end crashes that are representative of real-world rear-end crash scenarios. These reference datasets were used to create parameter combinations for simulations in order to generate synthetic crash scenarios.

The general idea was to gather as much information as possible from available data sources and minimize biases in the respective datasets and their combinations. As mentioned in Chapter 2.4, the three data sources used have their own limitations and strengths. Specifically, the SHRP2 dataset contains collected crash data from physical contact to a few with high severity, whereas the other data sources are biased towards severe crashes [41, 46]. The distribution of the lead vehicle's Delta-v for SHRP2 rear-end crashes was used as the reference distribution for crash severity.

Merging the pre-crash data from the CISS and GIDAS PCM datasets with the SHRP2 crash dataset provided the final dataset with more detailed information on severe crashes. The sample weighting process, using the KNN sample weighting method (presented in the Appendix), ensured that the weighted distribution of parameters in the final reference dataset matched their reference distributions.

How can representative synthetic rear-end crashes be generated?

A parametric multivariate distribution modeling method was used to build distribution models for the reference datasets, ensuring that the complexity and interdependencies inherent in the data are captured accurately.

These distribution models were used to create two synthetic datasets: one for the lead-vehicle kinematics and the other for the initial states of rear-end crash scenarios and the minimum accelerations of both vehicles. To obtain a set of synthetic rear-end crash scenarios, simulations based on these datasets and the following vehicle behavior model were conducted. These scenarios were further weighted using an Iterative Proportional Fitting (IPF)-based sample weighting method to match the known reference datasets, creating a representative synthetic rear-end crash dataset.

Finally, non-parametric statistic tests were employed to objectively compare the two datasets in three ways: 1) parameters' marginal distributions, 2) multivariate distributions, and 3) crash severity levels. The results indicate no significant difference between the reference and synthetic datasets. This finding validates the effectiveness of the proposed method for generating synthetic rear-end crash scenarios that are statistically indistinguishable from real-world data; thus, they provide a reliable basis for further research and analysis in traffic safety and crash prevention.

4.1 Contributions

The main contributions of this thesis are:

- **Data combination methods**: None of the available crash datasets contain all necessary signals without substantial bias; this shortcoming is common in data-driven studies [111, 112]. The data combination methods combine and weight data from multiple crash datasets to mitigate the biases and create a reference dataset of rear-end crash characteristics across the full severity range. Among these methods, the KNN sample weighting method is particularly noteworthy because it can be used to weight biased data to match a reference dataset when conventional post-stratification methods are not applicable due to omitted strata. The data combination methods can also be applied when a multivariate joint distribution is needed, but only datasets with biases or incomplete signals are available [35].

- **A parametric multivariate distribution modeling method**: As mentioned in Chapter 2.5, the data available are limited, with some parameters sparsely represented. Nonetheless, it is crucial to be able to interpret the distribution models and grasp the correlations among parameters, as they hold physical significance. The proposed method can handle all the issues above. Moreover, it can be readily modified and applied to the analysis of other crash scenarios and even to other analyses which require a distribution model when only a relatively limited dataset is available and an understanding of the underlying distribution is available. For example, this method can be applied to the analysis of other crash scenarios.

- **Two vehicle behavior models**: Vehicle behavior models are crucial for generating crash scenarios, and their accuracy greatly affects the realism of the generated crash scenarios at both individual and distribution levels [58]. The lead-vehicle kinematics and following-vehicle behavior models proposed in this thesis (developed or adapted for the reference pre-crash datasets) accurately cover the full severity range, in marked contrast to existing models Specifically, the lead-vehicle kinematics model captures various lead-vehicle speed patterns and accurately matches the lead-vehicle speed profiles. The following-vehicle behavior model combines two existing behavior models and incorporates the potential for generating ’abnormal’ driver acceleration behavior, a phenomenon observed in 9.2% of all crashes.

- **Formal validation methods**: Existing studies [26, 36, 39, 40, 100] have primarily employed informal methods (such as descriptive statistics and visualization techniques) to assess the similarity between synthetic and raw crash data. While these methods offer flexibility and accessibility, they suffer from drawbacks such as subjectivity, limited quantification, and the potential for misinterpretation [113, 114]. In contrast, the validation methods proposed in this thesis provide formal, objective comparisons between the synthetic and reference crash datasets. Furthermore, these methods contribute to ensuring the validity of research findings or simulations based on synthetic data in crash analysis (and related fields).

- **A synthetic rear-end crash dataset**: As mentioned in Chapter 1.2, most methods for generating synthetic crashes are inherently biased. Specifically, the crashes generated based on naturalistic driving data [24, 26, 33, 34] tend to over-represent low-severity crashes, thus inadequately reflecting the crash outcomes and criticality actually observed in crashes [38]. On the other hand, generating synthetic crashes with traditional in-depth crash databases often introduces a bias towards severe crashes [36, 40, 42]. Clearly, evaluating active safety technologies using a biased synthetic crash dataset yields biased outcomes. To address this issue, the naturalistic driving and pre-crash kinematics data were combined to create an unbiased (to the extent possible) synthetic crash dataset [115] representative of real-world crashes across the full severity range. Figure 4.2 illustrates this point, comparing the CDF curves of the lead vehicle's Delta-v in rear-end crashes from various datasets. The CISS and GIDAS datasets are biased towards severe crashes. Although the SHRP2 dataset is similar to the reference dataset, it lacks high-severity cases (the maximum Δv_l in the SHRP2 dataset is 7.4 m/s). In contrast, the synthetic crash dataset mirrors the reference dataset, encompassing crashes with Δv_l reaching 13.8 m/s. Therefore, we argue that the synthetic dataset is representative and can be used for the safety assessments of active safety technologies. Furthermore, the dataset can function as a benchmark when evaluating the representativeness of scenarios generated through other methods (such as those based on traffic simulation and machine learning), thereby aiding in understanding the limitations associated with those generation

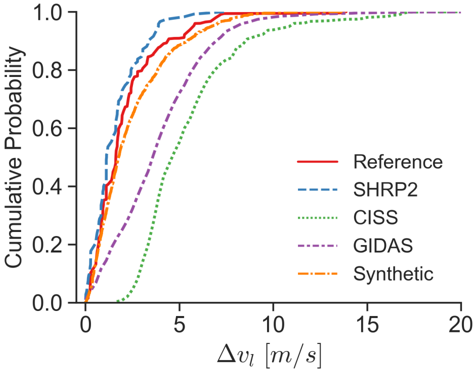


Figure 4.2: CDF curves for the lead vehicle's Delta-v (Δv_l) in rear-end crashes among various datasets.

methods.

4.2 Limitations and future work

Numerous additional variables not considered here, including road structure, traffic signals, and weather conditions, can influence the occurrence of a crash. Future research should address these considerations when more comprehensive data are available.

The modeled lead vehicle's acceleration is not consistently smooth. This is due to the fact that the speed of the lead vehicle was 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 segment transitions.

The parametric multivariate distribution modeling method only considers the linear correlation between two parameters, disregarding any possible nonlinear relationships and weak or non-significant correlations between them. The set of correlated parameters was modeled with a multivariate normal distribution, which can effectively model the linearly related parameters. These simplifications, which keep the model tractable and avoid overinterpreting the relationships between parameters, may, however, reduce the accuracy of the model. However, creating a complex multivariate model with a small dataset without a substantial risk of overfitting is not feasible.

Unfortunately, the sparsity of the available data makes investigating the consequences of these simplifications impossible, but future work should address this issue.

Paper A relies on pre-crash data from the US to establish the reference dataset of lead-vehicle kinematics. However, there was a shortage of pre-crash data in the US that included both vehicles: only 37 samples were available. Therefore, the GIDAS PCM dataset was used in Paper B, even though it was reconstructed pre-crash kinematics data from Germany. The rear-end crashes in the US and Germany were assumed to have similar mechanisms, although their distributions may differ.

The modified Intelligent Driver Model was used to simulate the acceleration behavior of the following vehicle. However, it's important to note that this model cannot accurately mimic the highly aggressive accelerations observed in some real-world crash situations, since the model was designed and calibrated to replicate naturalistic car-following behaviors rather than crashes. Future research should aim to calibrate the acceleration model using near-crash or pre-crash data to address this limitation.

The Kolmogorov–Smirnov test was used to compare the synthetic and reference datasets. This type of test was originally designed to test whether two datasets are significantly different. It is important to note that a test outcome indicating 'no statistically significant difference' cannot be confidently interpreted as evidence supporting equivalence or the absence of differences between the datasets [116]. Future work should aim to explore other methods of equivalence testing. One candidate is Bayesian statistics using ROPE (Region of Practical Equivalence) [117, 118]; this approach to hypothesis testing and parameter estimation focuses on practical significance rather than purely statistical significance.

Additionally, future work should aim to assess active safety technologies with synthetic crash scenarios and validate existing crash scenario generation methods, especially those that are traffic-simulation-based. Beyond this licentiate thesis, the future PhD research will mainly focus on this last topic.

The traffic-simulation-based scenario generation method has attracted a great deal of attention from researchers; as mentioned in Chapter 1.2, the method can simulate not only individual scenarios but also entire road networks comprising numerous road users, generating crashes resulting from the interactions between the computational models that make up all road users. Theoretically, the continuous spatiotemporal assessment of active safety technologies can be assessed with these simulations [15]. However, validation of the characteristics of the generated crashes is usually overlooked.

Validation is crucial to ensure that synthetic crash scenarios accurately represent real-world conditions (vehicle dynamics, driver behavior, road conditions, and environmental factors). Equally importantly, the representativeness of these characteristics should be verified at both the individual and distribution levels, and not the least on the outcome severity (here the delta-v was used). Stakeholders can use the validation results to make crucial decisions regarding active safety technologies (during system development and policy-making processes, for example). Furthermore, validation can identify biases and/or inaccuracies in the simulation process; addressing them will improve the reliability of the simulations.

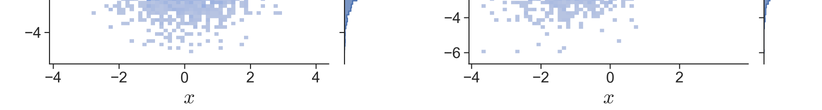
CHAPTER 5

Appendix

An empirical simulation study was conducted to compare the KNN imputation and KNN sample weighting methods for the specific problem described below.

5.1 Problem

Given 1) the target marginal distribution of x and 2) the observed dataset (x, y), how can the target joint distribution of (x, y) be obtained?



(a)

(b)

Figure 5.1: The two target datasets.

Type	Description	Sample size
S1	Unbiased (no omitted strata)	1,000
S2	Biased in x (with omitted strata)	1,000
S3	Biased in x and y (with omitted strata)	1,000

Table 5.1: The three types of observed datasets.

5.2 Dataset

5.2.1 The target datasets

- Joint distribution of $(x, y) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, where the means $\mu = [0,0]^T$, the covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} 1 & c_{xy} \\ c_{xy} & 2 \end{bmatrix}$, and the covariance c_{xy} is 0 or 0.7.
- Marginal distributions: $x \sim \mathcal{N}(0,1)$ and $y \sim \mathcal{N}(0,2)$. (Note that the marginal distribution of x is known.)

Figure 5.1 shows the two target datasets: 1) x and y are independent ($c_{xy} = 0$), and 2) x and y are correlated ($c_{xy} = 1$).

5.2.2 The observed datasets

Table 5.1 shows the three types of observed datasets, which were designed with different bias levels. Figure 5.2 shows the three observed datasets created for each target dataset.

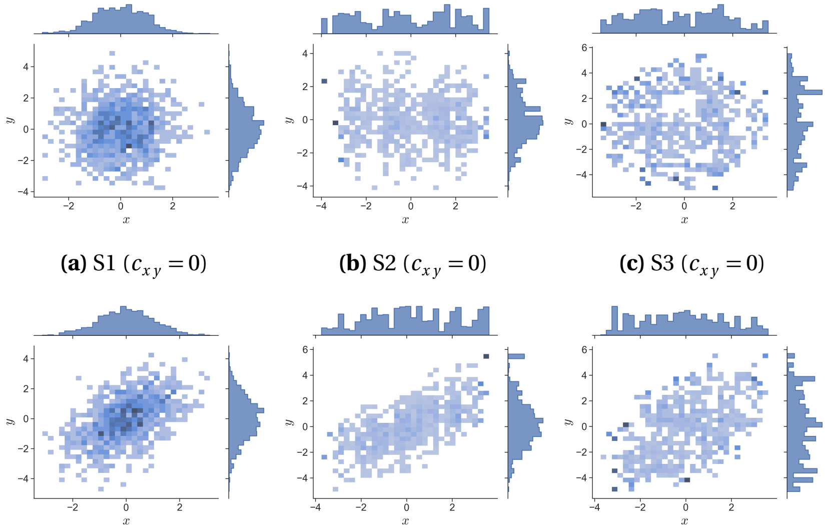
5.2.3 Simulations for the KNN imputation method

For each target dataset:

1. Build KNN imputers using the ”scikit-learn” package in Python [119] for the three observed datasets, respectively.
2. For $n = 100 : 100 : 1000$, repeat the following steps for 100 simulations:

   a. Generate n samples from the marginal distribution of x.

   b. Impute y values using the three data imputers, respectively.



(a) S1 $(c_{xy} = 0)$ **(b)** S2 $(c_{xy} = 0)$ **(c)** S3 $(c_{xy} = 0)$

(d) S1 $(c_{xy} = 0.7)$ **(e)** S2 $(c_{xy} = 0.7)$ **(f)** S3 $(c_{xy} = 0.7)$

Figure 5.2: The observed datasets for the two target datasets ($c_{xy} = 0$ and $c_{xy} = 0.7$).

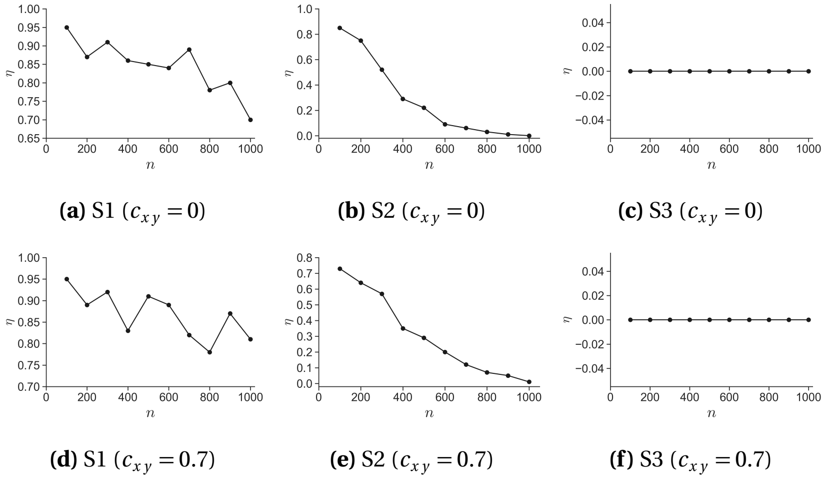
c. Using the two-sample KS test to determine whether the distribution of imputed y and the true marginal distribution of y, i.e., $\mathcal{N}(0,2)$, are significantly different (at the 0.05 significance level).

3. Compute the proportion of simulations with non-significant test, η.

5.2.4 Simulations for the KNN sample weighting method

For each target dataset:

1. Generate 10,000 samples from the marginal distribution of x.
2. Use the generated samples to set the sample weights for the three observed datasets using the KNN sample weighting method (see Section 5.5 for further information regarding the algorithm).
3. Use the weighted two-sample KS test to determine whether the



(a) S1 ($c_{xy} = 0$) **(b)** S2 ($c_{xy} = 0$) **(c)** S3 ($c_{xy} = 0$)

(d) S1 ($c_{xy} = 0.7$) **(e)** S2 ($c_{xy} = 0.7$) **(f)** S3 ($c_{xy} = 0.7$)

Figure 5.3: Performances of the KNN imputation method.

	Target dataset					
	$c_{xy} = 0$			$c_{xy} = 0.7$		
Observed dataset type	S1	S2	S3	S1	S2	S3
p-value	0.99	0.82	0.00	0.92	0.51	0.00
sample size	1,000	886	970	1,000	922	936
effective sample size	852	310	401	842	436	493

Table 5.2: Performances of the KNN sample weighting method.

weighted distribution of y and the true marginal distribution of y, i.e., $\mathcal{N}(0,2)$, are significantly different (at the 0.05 significance level).

5.3 Results

Figure 5.3 shows the simulation results for the KNN imputation method; Table 5.2 shows the KNN sample weighting method simulation results. Neither of the two methods can handle the observed dataset of the S3 type. The KNN imputation method's performance tends to decrease as the generated dataset's sample size increases. In contrast, the KNN sample weighting method produces results that are not significantly different for observed datasets of the S1 and S2 types for both target datasets. In addition, 20 various S0 and S1 datasets were added to test the robustness of KNN sample weighting. None of the results show a significant difference.

5.4 Conclusions

- The covariance c_{xy} does not strongly influence the simulations in the study.
- Neither of the two methods can handle the observed dataset of the S3 type (biased on both x and y).
- The KNN imputation method may produce a significantly different y distribution.
  - As the sample size increases, the possibility generally increases.
  - As the bias level of observed data increases, the possibility increases.
- The KNN sample weighting method does not produce any significantly different y distributions in any simulation for observed datasets of the S1 and S2 types. However, it would ignore observed data points that were never selected as the top k nearest neighbors for any raw sample, and the weighted data has an even smaller effective sample size.

5.5 The KNN sample weighting algorithm

Given:

- The observed dataset $\{\mathbf{X}_i | i \in [1, n]\}$, where $\mathbf{X}_i = [x_1^{(i)}, x_2^{(i)}, ..., x_K^{(i)}]^T$.
- The known target joint distribution $\tilde{\Phi}(x_1, x_2, ..., x_m)$ $(m < K)$.

Objective:

- Set sample weights for the observed data so that the weighted data maps the target joint distribution.

Algorithm:
As shown in Algorithm 1, the KNN sample weighting method contains four main steps.

Algorithm 1 KNN sample weighting algorithm.

```
Set $w_i = 0 \; \forall \; i \in [1, n]$
Generate $N$ samples from $\tilde{\Phi}(x_1, ..., x_m)$: $\{[\tilde{x}_1^{(j)}, ..., \tilde{x}_m^{(j)}]^T \mid j \in [1, N]\}$
For $j = 1$ to $N$:
    $d_j^{(i)} = \sqrt{\sum_{p=1}^{m} (\tilde{x}_p'^{(j)} - x_p'^{(i)})^2} \; \forall \; i \in [1, n]$
    $\omega_j^{(i)} = 1/d_j^{(i)}$ if all$(d_j^{(i)} > 0)$ else $I_{\{d_j^{(i)}=0\}} \; \forall \; i \in [1, n]$
    $H = \arg\max_i(\{\omega_j^{(i)} \mid i \in [1, n]\}, k)$
    $w_{h_l} \leftarrow w_{h_l} + \frac{\omega_j^{(h_l)}}{\sum_{l=1}^{k} \omega_j^{(h_l)}} \; \forall \; h_l \in H$
$w_i \leftarrow \frac{w_i}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} I_{\{w_i > 0\}} \; \forall \; i \in [1, n]$
```

1. Set the initial sample weight for each raw sample to zero: $w_i = 0 \; \forall \; i \in [1, n]$.

2. Sample N samples from the known reference distribution $\tilde{\Phi}(x_1, ..., x_m)$.

3. For any generated sample $\tilde{\mathbf{X}}_j$:

   a. Compute the Euclidean distance between $\tilde{\mathbf{X}}_j$ and $\mathbf{X}_i$, $d_j^{(i)}$, for all $i \in [1, n]$. ($\tilde{x}_p'^{(j)}$ and $x_p'^{(i)}$ are the standardized value of $\tilde{x}_p^{(j)}$ and $x_p^{(i)}$, respectively.)

   b. Compute the distributing weight of the raw sample $\mathbf{X}_i$ for $\tilde{\mathbf{X}}_j$, $\omega_j^{(i)}$, for all $i \in [1, n]$. (A smaller Euclidean distance correlates to a higher distributing weight.)

   c. Distribute a weight value of one among the top k raw samples with the highest distributing weights ($\{\mathbf{X}_{h_l} | h_l \in H\}$).

4. Scale the weights so that $\sum_{i=1}^{n} w_i = n$.

The value of k is determined by minimizing the loss, $\sum_{l=1}^{m} s_l^{(k)}$, where $s_l^{(k)}$ is the KS statistic for x_l conditioned on k computed with the weighted two-sample KS tests between the weighted x_l data and the reference data $\{\tilde{x}_l^{(j)} | j \in [1, N]\}$.

References

[1] L. Xiao and F. Gao (2010). A comprehensive review of the development of adaptive cruise control systems. *Vehicle system dynamics* **48** (10), pp. 1167–1192. DOI: 10.1080/00423110903365910.

[2] W. Chen, W. Wang, K. Wang, Z. Li, H. Li and S. Liu (2020). Lane departure warning systems and lane line detection methods based on image processing and semantic segmentation: a review. *Journal of traffic and transportation engineering (English edition)* **7** (6), pp. 748–774. DOI: 10.1016/j.jtte.2020.10.002.

[3] E. Coelingh, A. Eidehall and M. Bengtsson (2010). 'Collision warning with full auto brake and pedestrian detection-a practical example of automatic emergency braking'. In: *13th international ieee conference on intelligent transportation systems*. IEEE, pp. 155–160. DOI: 10.1109/itsc.2010.5625077.

[4] D. P. Bui, S. Balland, C. Giblin, A. M. Jung, S. Kramer, A. Peng, M. C. P. Aquino, S. Griffin, D. D. French, K. P. Porter et al. (2018). Interventions and controls to prevent emergency service vehicle incidents: a mixed methods review. *Accident Analysis & Prevention* **115**, pp. 189–201. DOI: 10.1016/j.aap.2018.01.006.

[5] G. Bathla, K. Bhadane, R. K. Singh, R. Kumar, R. Aluvalu, R. Krishnamurthi, A. Kumar, R. Thakur and S. Basheer (2022). Autonomous vehicles and intelligent automation: applications, challenges, and opportunities. *Mobile Information Systems* **2022**. DOI: 10.1155/2022/7632892.

[6] S. Abbasi and A. M. Rahmani (2023). Artificial intelligence and software modeling approaches in autonomous vehicles for

safety management: a systematic review. *Information* **14** (10), p. 555. DOI: 10.3390/info14100555.

[7] A. Giannaros, A. Karras, L. Theodorakopoulos, C. Karras, P. Kranias, N. Schizas, G. Kalogeratos and D. Tsolis (2023). Autonomous vehicles: sophisticated attacks, safety issues, challenges, open topics, blockchain, and future directions. *Journal of Cybersecurity and Privacy* **3** (3), pp. 493–543. DOI: 10.3390/jcp3030025.

[8] A. Kashevnik, R. Shchedrin, C. Kaiser and A. Stocker (2021). Driver distraction detection methods: a literature review and framework. *IEEE Access* **9**, pp. 60063–60076. DOI: 10.1109/access.2021.3073599.

[9] J. F. May and C. L. Baldwin (2009). Driver fatigue: the importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transportation research part F: traffic psychology and behaviour* **12** (3), pp. 218–224. DOI: 10.1016/j.trf.2008.11.005.

[10] R. B. Voas and J. C. Fell (2011). Preventing impaired driving opportunities and problems. *Alcohol Research & Health* **34** (2), p. 225.

[11] M. E. Dean and L. E. Riexinger (2022). *Estimating the real-world benefits of lane departure warning and lane keeping assist.* Tech. rep. SAE Technical Paper. DOI: 10.4271/2022-01-0816.

[12] H. Tan, F. Zhao, H. Hao and Z. Liu (2020a). Estimate of safety impact of lane keeping assistant system on fatalities and injuries reduction for china: scenarios through 2030. *Traffic injury prevention* **21** (2), pp. 156–162. DOI: 10.1080/15389588.2020.1711518.

[13] H. Tan, F. Zhao, H. Hao, Z. Liu, A. A. Amer and H. Babiker (2020b). Automatic emergency braking (aeb) system impact on fatality and injury reduction in china. *International journal of environmental research and public health* **17** (3), p. 917. DOI: 10.3390/ijerph17030917.

[14] Y. Zhao, D. Ito and K. Mizuno (2019). Aeb effectiveness evaluation based on car-to-cyclist accident reconstructions using

video of drive recorder. *Traffic injury prevention* **20** (1), pp. 100–106. DOI: 10.1080/15389588.2018.1533247.

[15] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick and F. Diermeyer (2020). Survey on scenario-based safety assessment of automated vehicles. *IEEE access* **8**, pp. 87456–87477. DOI: 10.1109/access.2020.2993730.

[16] E. de Gelder and J.-P. Paardekooper (2017). 'Assessment of automated driving systems using real-life scenarios'. In: *2017 ieee intelligent vehicles symposium (iv)*. IEEE, pp. 589–594. DOI: 10.1109/ivs.2017.7995782.

[17] S. Feng, Y. Feng, X. Yan, S. Shen, S. Xu and H. X. Liu (2020). Safety assessment of highly automated driving systems in test tracks: a new framework. *Accident Analysis & Prevention* **144**, p. 105664. DOI: 10.1016/j.aap.2020.105664.

[18] R. Donà and B. Ciuffo (2022). Virtual testing of automated driving systems. a survey on validation methods. *IEEE Access* **10**, pp. 24349–24367. DOI: 10.1109/access.2022.3153722.

[19] J. Cai, W. Deng, H. Guang, Y. Wang, J. Li and J. Ding (2022). A survey on data-driven scenario generation for automated vehicle testing. *Machines* **10** (11), p. 1101. DOI: 10.3390/machines10111101.

[20] Z. Szalay (2023). Critical scenario identification concept: the role of the scenario-in-the-loop approach in future automotive testing. *IEEE Access*. DOI: 10.1109/access.2023.3298875.

[21] P. Wimmer, O. Op_Den_Camp, H. Weber, H. Chajmowicz, M. Wagner, J. L. Mallada, F. Fahrenkrog and F. Denk (2023). 'Harmonized approaches for baseline creation in prospective safety performance assessment of driving automation systems'. In: *27th international technical conference on the enhanced safety of vehicles (esv), yokohama, japan*, pp. 3–6.

[22] N. Kauffmann, F. Fahrenkrog, L. Drees and F. Raisch (2022). Positive risk balance: a comprehensive framework to ensure vehicle safety. *Ethics and Information Technology* **24** (1), p. 15. DOI: 10.1007/s10676-022-09625-2.

[23] A. Fries, F. Fahrenkrog, K. Donauer, M. Mai and F. Raisch (2022). 'Driver behavior model for the safety assessment of automated

driving'. In: *2022 ieee intelligent vehicles symposium (iv)*. IEEE, pp. 1669–1674. DOI: 10.1109/iv51971.2022.9827404.

[24] S. Shah, D. Dey, C. Lovett and A. Kapoor (2018). 'Airsim: high-fidelity visual and physical simulation for autonomous vehicles'. In: *Field and service robotics: results of the 11th international conference*. Springer, pp. 621–635. DOI: 10.1007/978-3-319-67361-5_40.

[25] W. Baron, C. Sippl, K.-S. Hielscher and R. German (2020). 'Repeatable simulation for highly automated driving development and testing'. In: *2020 ieee 91st vehicular technology conference (vtc2020-spring)*. IEEE, pp. 1–7. DOI: 10.1109/vtc2020-spring48590.2020.9129208.

[26] S. Feng, X. Yan, H. Sun, Y. Feng and H. X. Liu (2021). Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications* **12** (1), p. 748. DOI: 10.1038/s41467-021-21007-8.

[27] H. Hamdane, T. Serre, C. Masson and R. Anderson (2015). Issues and challenges for pedestrian active safety systems based on real world accidents. *Accident Analysis & Prevention* **82**, pp. 53–60. DOI: 10.1016/j.aap.2015.05.014.

[28] J. Bärgman, C.-N. Boda and M. Dozza (2017). Counterfactual simulations applied to shrp2 crashes: the effect of driver behavior models on safety benefit estimations of intelligent safety systems. *Accident Analysis & Prevention* **102**, pp. 165–180. DOI: 10.1016/j.aap.2017.03.003.

[29] M. Bareiss, J. Scanlon, R. Sherony and H. C. Gabler (2019). Crash and injury prevention estimates for intersection driver assistance systems in left turn across path/opposite direction crashes in the united states. *Traffic injury prevention* **20** (sup1), S133–S138. DOI: 10.1080/15389588.2019.1610945.

[30] S. H. Haus and H. C. Gabler (2019). The potential for active safety mitigation of us vehicle-bicycle crashes. *Future Active Safety Technology Towards Zero Traffic Accidents (FAST-Zero)*.

[31] R. Utriainen (2020). The potential impacts of automated vehicles on pedestrian safety in a four-season country. *Journal*

of Intelligent Transportation Systems **25** (2), pp. 188–196. DOI: 10.1080/15472450.2020.1845671.

[32] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle and T. Victor (2021). Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. *Accident Analysis & Prevention* **163**, p. 106454. DOI: 10.1016/j.aap.2021.106454.

[33] W. Li, C. Pan, R. Zhang, J. Ren, Y. Ma, J. Fang, F. Yan, Q. Geng, X. Huang, H. Gong et al. (2019). Aads: augmented autonomous driving simulation using data-driven algorithms. *Science robotics* **4** (28), eaaw0863. DOI: 10.1126/scirobotics.aaw0863.

[34] R. Arvin, A. J. Khattak, M. Kamrani and J. Rio-Torres (2020). Safety evaluation of connected and automated vehicles in mixed traffic with conventional vehicles at intersections. *Journal of Intelligent Transportation Systems* **25** (2), pp. 170–187. DOI: 10.1080/15472450.2020.1834392.

[35] J. Wu, C. Flannagan, U. Sander and J. Bärgman (2024). Modeling lead-vehicle kinematics for rear-end crash scenario generation. *IEEE Transactions on Intelligent Transportation Systems.* DOI: 10.1109/TITS.2024.3369097.

[36] A. Leledakis, M. Lindman, J. Östh, L. Wågström, J. Davidsson and L. Jakobsson (2021). A method for predicting crash configurations using counterfactual simulations and real-world data. *Accident Analysis & Prevention* **150**, p. 105932. DOI: 10.1016/j.aap.2020.105932.

[37] E. de Gelder, J. Hof, E. Cator, J.-P. Paardekooper, O. O. den Camp, J. Ploeg and B. De Schutter (2022). Scenario parameter generation method and scenario representativeness metric for scenario-based assessment of automated vehicles. *IEEE Transactions on Intelligent Transportation Systems* **23** (10), pp. 18794–18807. DOI: 10.1109/tits.2022.3154774.

[38] P. Olleja, J. Bärgman and N. Lubbe (2022). 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? *Journal of safety research* **83**, pp. 139–151. DOI: 10.1016/j.jsr.2022.08.011.

[39] I. R. Jenkins, L. O. Gee, A. Knauss, H. Yin and J. Schroeder (2018). 'Accident scenario generation with recurrent neural networks'. In: *2018 21st international conference on intelligent transportation systems (itsc)*. IEEE, pp. 3340–3345. DOI: 10.1109/itsc.2018.8569661.

[40] X. Wang, Y. Peng, T. Xu, Q. Xu, X. Wu, G. Xiang, S. Yi and H. Wang (2022). Autonomous driving testing scenario generation based on in-depth vehicle-to-powered two-wheeler crash data in china. *Accident Analysis & Prevention* **176**, p. 106812. DOI: 10.1016/j.aap.2022.106812.

[41] F. Zhang, E. Y. Noh, R. Subramanian and C.-L. Chen (2019). *Crash investigation sampling system: sample design and weighting*. Tech. rep.

[42] A. Gambi, T. Huynh and G. Fraser (2019). 'Generating effective test cases for self-driving cars from police reports'. In: *Proceedings of the 2019 27th acm joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pp. 257–267. DOI: 10.1145/3338906.3338942.

[43] A. Z. Zambom and D. Ronaldo (2013). A review of kernel density estimation with applications to econometrics. *International Econometric Review* **5** (1), pp. 20–42.

[44] 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. URL: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813375.

[45] J. M. Hankey, M. A. Perez and J. A. McClafferty (2016). *Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets*. Tech. rep. Virginia Tech Transportation Institute.

[46] A. Schubert, H. Liers and M. Petzold (2017). The gidas pre-crash-matrix 2016: innovations for standardized pre-crash-scenarios on the basis of the vufo simulation model vast.

[47] J. Schoner, R. Sanders and T. Goddard (2023). Effects of advanced driver assistance systems on impact velocity and

injury severity: an exploration of data from the crash investigation sampling system. *Transportation Research Record*, p. 03611981231189740. DOI: `10.1177/03611981231189740`.

[48] D. I. Swedler, B. Ali, R. Hoffman, J. Leonardo, E. Romano and T. R. Miller (2024). Injury and fatality risks for child pedestrians and cyclists on public roads. *Injury epidemiology* **11** (1), pp. 1–11. DOI: `10.1186/s40621-024-00497-2`.

[49] K. Böhm, T. Kubjatko, D. Paula and H.-G. Schweiger (2020). New developments on edr (event data recorder) for automated vehicles. *Open Engineering* **10** (1), pp. 140–146. DOI: `10.1515/eng-2020-0007`.

[50] D. Otte, M. Jänsch and C. Haasper (2012). Injury protection and accident causation parameters for vulnerable road users based on german in-depth accident study gidas. *Accident Analysis & Prevention* **44** (1), pp. 149–153. DOI: `10.1016/j.aap.2010.12.006`.

[51] H. Johannsen, D. Otte and M. Urban (2015). 'Pre-crash analysis of accidents involving turning trucks and bicyclists'. In: *Ircobi council (hg.): 2015 ircobi conference proceedings. ircobi*, pp. 09–11.

[52] F. Char and T. Serre (2020). Analysis of pre-crash characteristics of passenger car to cyclist accidents for the development of advanced drivers assistance systems. *Accident Analysis & Prevention* **136**, p. 105408. DOI: `10.1016/j.aap.2019.105408`.

[53] E. Rosen (2013). 'Autonomous emergency braking for vulnerable road users'. In: *Proceedings of ircobi conference*, pp. 618–627.

[54] U. Sander (2018). *Predicting safety benefits of automated emergency braking at intersections-virtual simulations based on real-world accident data.* Chalmers University of Technology.

[55] F. Char, T. Serre, S. Compigne and P. Puente Guillen (2022). Car-to-cyclist forward collision warning effectiveness evaluation: a parametric analysis on reconstructed real accident cases. *International journal of crashworthiness* **27** (1), pp. 34–43. DOI: `10.1080/13588265.2020.1773740`.

[56] R. Putter, A. Neubohn, A. Leschke and R. Lachmayer (2023). Predictive vehicle safety—validation strategy of a perception-based crash severity prediction function. *Applied Sciences* **13** (11), p. 6750. DOI: 10.3390/app13116750.

[57] T. A. Gennarelli and E. Wodzin (2006). Ais 2005: a contemporary injury scale. *Injury* **37** (12), pp. 1083–1091. DOI: 10.1016/j.injury.2006.07.009.

[58] C. Wang, F. Guo, R. Yu, L. Wang and Y. Zhang (2023). The application of driver models in the safety assessment of autonomous vehicles: a survey. *arXiv preprint arXiv:2303.14779*.

[59] P. G. Gipps (1981). A behavioural car-following model for computer simulation. *Transportation research part B: methodological* **15** (2), pp. 105–111. DOI: 10.1016/0191-2615(81)90037-0.

[60] M. Brackstone and M. McDonald (1999). Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour* **2** (4), pp. 181–196. DOI: 10.1016/s1369-8478(00)00005-x.

[61] M. Treiber, A. Hennecke and D. Helbing (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical review E* **62** (2), p. 1805. DOI: 10.1103/physreve.62.1805.

[62] Y. Li and D. Sun (2012). Microscopic car-following model for the traffic flow: the state of the art. *Journal of Control Theory and Applications* **10** (2), pp. 133–143. DOI: 10.1007/s11768-012-9221-z.

[63] O. Derbel, T. Peter, H. Zebiri, B. Mourllion and M. Basset (2013). Modified intelligent driver model for driver safety and traffic stability improvement. *IFAC Proceedings Volumes* **46** (21), pp. 744–749. DOI: 10.3182/20130904-4-jp-2042.00132.

[64] V. A. Butakov and P. Ioannou (2014). Personalized driver/vehicle lane change models for adas. *IEEE Transactions on Vehicular Technology* **64** (10), pp. 4422–4431. DOI: 10.1109/tvt.2014.2369522.

[65] S. Moridpour, M. Sarvi and G. Rose (2010). Lane changing models: a critical review. *Transportation letters* **2** (3), pp. 157–173. DOI: 10.3328/tl.2010.02.03.157-173.

[66] B. Zhou, Y. Wang, G. Yu and X. Wu (2017). A lane-change trajectory model from drivers' vision view. *Transportation Research Part C: Emerging Technologies* **85**, pp. 609–627. DOI: 10.1016/j.trc.2017.10.013.

[67] D. D. Salvucci (2006). Modeling driver behavior in a cognitive architecture. *Human factors* **48** (2), pp. 362–380. DOI: 10.1518/001872006777724417.

[68] M. Baumann and J. F. Krems (2007). 'Situation awareness and driving: a cognitive model'. In: *Modelling driver behaviour in automotive environments: critical issues in driver interactions with intelligent transport systems.* Springer, pp. 253–265. DOI: 10.1007/978-1-84628-618-6_14.

[69] J. D. Lee, D. V. McGehee, T. L. Brown and M. L. Reyes (2002). Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. *Human factors* **44** (2), pp. 314–334. DOI: 10.1518/0018720024497844.

[70] G. Li, W. Wang, S. E. Li, B. Cheng and P. Green (2014). Effectiveness of flashing brake and hazard systems in avoiding rear-end crashes. *Advances in Mechanical Engineering* **6**, p. 792670. DOI: 10.1155/2014/792670.

[71] X. Wang, M. Zhu, M. Chen and P. Tremont (2016). Drivers' rear end collision avoidance behaviors under different levels of situational urgency. *Transportation research part C: emerging technologies* **71**, pp. 419–433. DOI: 10.1016/j.trc.2016.08.014.

[72] M. Svärd, G. Markkula, J. Bärgman and T. Victor (2021). Computational modeling of driver pre-crash brake response, with and without off-road glances: parameterization using real-world crashes and near-crashes. *Accident Analysis & Prevention* **163**, p. 106433. DOI: 10.31234/osf.io/6nkgv.

[73] D. N. Lee (1976). A theory of visual control of braking based on information about time-to-collision. *Perception* **5** (4), pp. 437–459. DOI: 10.1068/p050437.

[74] O. Kempthorne (2005). *Design and analysis of experiments: advanced experimental design.* Wiley-Interscience.

[75] S. M. Ross (2014). *Introduction to probability models.* Academic press. DOI: 10.1016/b978-0-12-407948-9.00012-8.

[76] S. Wolfram (1991). *Mathematica: a system for doing mathematics by computer.* Addison Wesley Longman Publishing Co., Inc.

[77] G. Ridder and R. Moffitt (2007). The econometrics of data combination. *Handbook of econometrics* **6**, pp. 5469–5547. DOI: 10.1016/s1573-4412(07)06075-8.

[78] D. Holt and T. F. Smith (1979). Post stratification. *Journal of the Royal Statistical Society Series A: Statistics in Society* **142** (1), pp. 33–46.

[79] D. Sloane and S. P. Morgan (1996). An introduction to categorical data analysis. *Annual review of sociology* **22** (1), pp. 351–375.

[80] R. Malarvizhi and A. S. Thanamani (2012). K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev* **5** (1), pp. 5–7.

[81] T. W. Anderson and I. Olkin (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear algebra and its applications* **70**, pp. 147–171. DOI: 10.1016/0024-3795(85)90049-7.

[82] A. G. Stephenson (2009). High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics* **51** (1), pp. 77–88. DOI: 10.1111/j.1467-842x.2008.00528.x.

[83] A. Kottas, P. Müller and F. Quintana (2005). Nonparametric bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics* **14** (3), pp. 610–625. DOI: 10.1198/106186005x63185.

[84] N. Yang, Y. Huang, D. Hou, S. Liu, D. Ye, B. Dong and Y. Fan (2019). Adaptive nonparametric kernel density estimation ap-

proach for joint probability density function modeling of multiple wind farms. *Energies* **12** (7), p. 1356. DOI: 10.3390/en12071356.

[85] A. Sancetta and S. Satchell (2004). The bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric theory* **20** (3), pp. 535–562. DOI: 10.1017/s026646660420305x.

[86] R. B. Nelsen (2006). *An introduction to copulas*. Springer. DOI: 10.1007/0-387-28678-0.

[87] I. Kojadinovic and J. Yan (2010). Modeling multivariate distributions with continuous margins using the copula r package. *Journal of Statistical Software* **34**, pp. 1–20. DOI: 10.18637/jss.v034.i09.

[88] H. Kazianka and J. Pilz (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic environmental research and risk assessment* **24**, pp. 661–673. DOI: 10.1007/s00477-009-0353-8.

[89] Y.-C. Chen (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology* **1** (1), pp. 161–187. DOI: 10.1080/24709360.2017.1396742.

[90] J. Orava (2011). K-nearest neighbour kernel density estimation, the choice of optimal k. *Tatra Mountains Mathematical Publications* **50** (1), pp. 39–50. DOI: 10.2478/v10127-011-0035-z.

[91] V. Vapnik (2013). *The nature of statistical learning theory*. Springer science & business media.

[92] R. G. Easterling (2010). Passion-driven statistics. *The American Statistician* **64** (1), pp. 1–5. DOI: 10.1198/tast.2010.09180.

[93] J. Warmbrod (2001). Calculating, interpreting, and reporting estimates of "effect size" (magnitude of an effect or the strength of a relationship). *Texas Tech. University, Texas*, pp. 1–22.

[94] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang and I. Cohen (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pp. 1–4.

[95] Y. M. Bishop, S. E. Fienberg and P. W. Holland (2007). *Discrete multivariate analysis: theory and practice.* Springer Science & Business Media. DOI: 10.1007/978-0-387-72806-3.

[96] S. Kolenikov (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal* **14** (1), pp. 22–59. DOI: 10.1177/1536867x1401400104.

[97] A.-A. Choupani and A. R. Mamdoohi (2016). Population synthesis using iterative proportional fitting (ipf): a review and future research. *Transportation Research Procedia* **17**, pp. 223–233. DOI: 10.1016/j.trpro.2016.11.078.

[98] P. Norman (1999). Putting iterative proportional fitting on the researcher's desk.

[99] H. W. Lilliefors (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association* **62** (318), pp. 399–402. DOI: 10.2307/2283970.

[100] A. Demetriou, H. Alfsvåg, S. Rahrovani and M. Haghir Chehreghani (2023). A deep learning framework for generation and analysis of driving scenario trajectories. *SN Computer Science* **4** (3), p. 251. DOI: 10.1007/s42979-023-01714-3.

[101] L. Van der Maaten and G. Hinton (2008). Visualizing data using t-sne. *Journal of machine learning research* **9** (11).

[102] L. McInnes, J. Healy and J. Melville (2018). Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426.*

[103] N. M. Razali, Y. B. Wah et al. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics* **2** (1), pp. 21–33.

[104] N. A. Heckert, J. J. Filliben, C. M. Croarkin, B. Hembree, W. F. Guthrie, P. Tobias and J. Prinz (2002). Handbook 151: nist/sematech e-handbook of statistical methods.

[105] M. A. Stephens (2017). 'Tests based on edf statistics'. In: *Goodness-of-fit-techniques.* Routledge, pp. 97–194. DOI: 10.1201/9780203753064-4.

[106] L. Lerche (2012). *Quantitative methods*. Elsevier.

[107] G. Fasano and A. Franceschini (1987). A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society* **225** (1), pp. 155–170. DOI: `10.1093/mnras/225.1.155`.

[108] A. Hagen, S. Jackson, J. Kahn, J. Strube, I. Haide, K. Pazdernik and C. Hainje (2021). Accelerated computation of a high dimensional kolmogorov-smirnov distance. *arXiv preprint arXiv:2106.13706*.

[109] H. Roux de Bezieux (2021). *Ecume: equality of 2 (or k) continuous univariate and multivariate distributions*. URL: `https://CRAN.R-project.org/package=Ecume`.

[110] J. Bärgman, V. Lisovskaja, T. Victor, C. Flannagan and M. Dozza (2015). How does glance behavior influence crash and injury risk? a 'what-if'counterfactual simulation using crashes and near-crashes from shrp2. *Transportation Research Part F: Traffic Psychology and Behaviour* **35**, pp. 152–169. DOI: `10.1016/j.trf.2015.10.011`.

[111] E. Hargittai (2020). Potential biases in big data: omitted voices on social media. *Social Science Computer Review* **38** (1), pp. 10–24. DOI: `10.1177/0894439318788322`.

[112] J. Zhu and B. Salimi (2024). Overcoming data biases: towards enhanced accuracy and reliability in machine learning. *IEEE Data Eng. Bull.* **47** (1), pp. 18–35.

[113] M. Ankerst, S. Berchtold and D. A. Keim (1998). 'Similarity clustering of dimensions for an enhanced visualization of multidimensional data'. In: *Proceedings ieee symposium on information visualization (cat. no. 98tb100258)*. IEEE, pp. 52–60. DOI: `10.1109/infvis.1998.729559`.

[114] J. Heer and B. Shneiderman (2012). Interactive dynamics for visual analysis: a taxonomy of tools that support the fluent and flexible use of visualizations. *Queue* **10** (2), pp. 30–55. DOI: `10.1145/2133416.2146416`.

[115] J. Wu (2024). *QUADRIS project pre-crash/near-crash dataset*. URL: `https://github.com/JianWu09/QUADRIS-project-Pre-crash-near-crash-database`.

[116] N. Gibbs (2013). *Errors in the interpretation of 'no statistically significant difference'*. DOI: 10.1177/0310057x1304100203.

[117] P. Schwaferts and T. Augustin (2020). Bayesian decisions using regions of practical equivalence (rope): foundations.

[118] J. Kruschke (2014). Doing bayesian data analysis: a tutorial with r, jags, and stan.

[119] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**, pp. 2825–2830.