World Electric Vehicle Journal

Article

Statistical Risk and Performance Analyses on Naturalistic Driving Trajectory Datasets for Traffic Modeling

Ruixue Zong [1], Ying Wang [2,*], Juan Ding [3,4,*] and Weiwen Deng [1]

[1] School of Transportation Science and Engineering, Beihang University, Beijing 100191, China; zongrx@buaa.edu.cn (R.Z.); wdeng@buaa.edu.cn (W.D.)
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[3] College of Mechanical and Electrical Engineering, Jiaxing Nanhu University, Jiaxing 314001, China
[4] PanoSim Technology Limited Company, Jiaxing 314000, China
* Correspondence: wangying_jlu@jlu.edu.cn (Y.W.); juan.ding@panosim.com (J.D.)

Abstract: The development of autonomous driving technology has made simulation testing one of the most important tools for evaluating system performance. However, there is a lack of systematic methods for analyzing and assessing naturalistic driving trajectory datasets. Specifically, there is a lack of comprehensive analyses on data diversity and balance in machine learning-oriented research. This study presents a comprehensive assessment of existing highway scenario datasets in the context of traffic modeling in autonomous driving simulation tests. In order to clarify the level of traffic risk, we design a systematic risk index and propose an index describing the degree of data scatter based on the principle of Euclidean distance quantization. By comparing several datasets, including NGSIM, highD, INTERACTION, CitySim, and our self-collected Highway dataset, we find that the proposed metrics can effectively quantify the risk level of the dataset while helping to gain insight into the diversity and balance differences of the dataset.

Keywords: naturalistic driving trajectory datasets; simulation tests; traffic modeling; risk

check for updates

Citation: Zong, R.; Wang, Y.; Ding, J.; Deng, W. Statistical Risk and Performance Analyses on Naturalistic Driving Trajectory Datasets for Traffic Modeling. *World Electr. Veh. J.* **2024**, *15*, 77. https://doi.org/10.3390/wevj15030077

Academic Editor: Joeri Van Mierlo

Received: 21 December 2023
Revised: 13 February 2024
Accepted: 17 February 2024
Published: 21 February 2024

1. Introduction

The virtual simulation testing on autonomous driving using digital twin technology to simulate the driving environment can effectively avoid the risk of collisions and injuries that may occur in the real world. With the increasing demand for fidelity in virtual driving environments, data-driven approaches are widely and rapidly adopted in traffic modeling and simulation. As a result, the acquisition and processing of traffic data becomes one of the most critical issues in data-driven traffic modeling.

With the advancement of sensing technology, especially image recognition technology, naturalistic driving data have been greatly improved in terms of quantity, variety, and quality by acquiring image or trajectory data from videos. Such progress serves as a robust foundation for capturing driving behavior characteristics, comprehending distinct traffic phenomena, and propelling the evolution of digital twin and autonomous driving technologies within the domains of traffic flow theory and autonomous driving.

Studies in traffic and driver behavior modeling rely on datasets of vehicle trajectory. The Next Generation Simulation (NGSIM) [1] dataset is one of the first open-source trajectory datasets collected and released in 2005 and is primarily used for traffic simulation. In recent years, this dataset has been widely used in a variety of different traffic studies at both macro and micro levels, supporting in-depth studies of traffic phenomena, including but not limited to the discovery or validation of macro and micro level phenomena such as traffic hysteresis, capacity degradation, asymmetric driving behavior [2], delayed effects of driving behavior [3], and relaxation phenomena of lane-changing behavior [4].

Naturalistic driving trajectory data plays a crucial role in the development and testing of autonomous driving technology, in helping autonomous driving systems better comprehend and navigate complex driving environments. Consequently, it is imperative that the dataset has attributes such as fine granularity, high accuracy, and diversity. In the quest for improved data accuracy and richness, a number of high-precision trajectory datasets have been introduced. The highD [5] dataset addresses the limitations of the NGSIM dataset, particularly in terms of trajectory accuracy, speed range, and sample duration. In particular, the use of drone-mounted cameras for data acquisition in highD, represents a novel approach to flexible data collection. Based on the same collection method, the highway merge-in and merge-out dataset exiD [6], the roundabout intersection dataset rounD [7], and the trajectory datasets in various scenarios with vulnerable road user dataset inD [8] have been successively launched to further enrich the trajectory data.

Furthermore, to fulfill the requirements for human-like behavior learning and critical scenario handling in autonomous driving, a number of road user interaction datasets, consisting of INTERACTION [9], OpenACC [10], OpenDD [11], AUTOMATUM [12], CitySim [13], pNEUMA [14], ZEN [15] and MAGIC [16] have been sequentially introduced.

Existing studies have summarized and analyzed the above datasets from different perspectives based on different research objectives. As these datasets are all formed with trajectory as time series in positions, errors in these data can cause shifts in kinematic variables (speed and acceleration, etc.) obtained from differentiation, thus resulting in disturbances of the vehicle dynamics and kinematics. NGSIM suffers from accuracy problems, such that there are more studies on the accuracy of NGSIM, including the study of the error analysis [17], the study of the distortion phenomenon of the trajectory [18], the data reconstruction techniques [19], and the smooth algorithm [20] to extract more realistic velocity and acceleration information from the positional data. Furthermore, the objective of characterizing the data is a key part of the process of understanding and applying the data. The viewpoints encompass the macro level (speed-density-flow relationship, etc.), the micro-level (speed and acceleration, etc.), the level of driving behavior [21] (following and changing lanes, etc.), and the scenario level [22]. Autonomous driving-related research has focused more on the interaction and risk characteristics of data. With respect to evaluating the degree of data interaction, existing studies have mainly focused on lane-changing behavior, where the description of risk or conflict is mainly based on the Time to Collision (TTC) [6,23–25]. The use of driving simulator-enhanced datasets to supplement specific interaction data [26] has also become one of the approaches to dataset construction.

However, the above datasets and analyses are not sufficient for virtual simulation testing. The driving environment in virtual simulation testing needs to reflect the diversity and riskiness of driving behaviors. Specifically, autonomous driving systems must be able to handle the diverse, complex, and risky driving behaviors that occur in the real world. If the behaviors of traffic vehicles in the virtual driving environment are highly homogeneous or low-risk, it will not be conducive to verifying the generalization performance of autonomous driving algorithms and evaluating the performance limits of autonomous driving systems. In data-driven approaches, the effectiveness of the traffic model is strongly influenced by the training data. The large size but high homogeneity of the data implies inefficiencies in data processing and model training and poor generalization performance of the models. In addition, there has been limited consideration of the differences between different datasets, resulting in poor model performance due to ignoring data differences and reuse of modeling methods.

To address the lack of evaluation methods for the quality of naturalistic driving trajectory datasets in existing studies, and based on the needs of traffic modeling with respect to the quality of naturalistic driving trajectory datasets, this paper proposes a method to evaluate the diversity and balance of the data based on systematic risk indices. To quantify the traffic system risk in the data rather than the risk of individual behaviors, we define and quantify the systematic risk indices of traffic segments. Based on systematic risk indices, a dispersion index is proposed, which adopts the Euclidean distance to quantify the similarity of the systematic risk vectors of each traffic segment, indirectly reflecting the diversity and balance of data through the dispersion of the risk vectors in Euclidean space. In the analysis stage, we compare and analyze the dataset quality of highway scenarios in NGSIM, highD, INTERACTION, CitySim, and a self-selected dataset, Highway, by adopting the above method. The results show that the risk and dispersion indices can effectively quantify the risk level of the datasets and distinguish the differences between datasets in terms of diversity and balance.

The rest of the paper is organized as follows: Section 2 presents the data requirements for virtual simulation test traffic modeling and an overview of existing natural driving trajectory datasets and preprocesses all the datasets; Section 3 describes micro-feature extraction of the trajectory datasets as well as extraction and quantification of the system risk metrics; Section 4 proposes metrics describing the data diversity and balance, and provides a comprehensive evaluation of NGSIM, highD, INTERACTION, CitySim, and Highway, and Section 5 presents the conclusions and outlook.

2. Traffic Modeling Datasets for Simulation Tests

2.1. Requirements

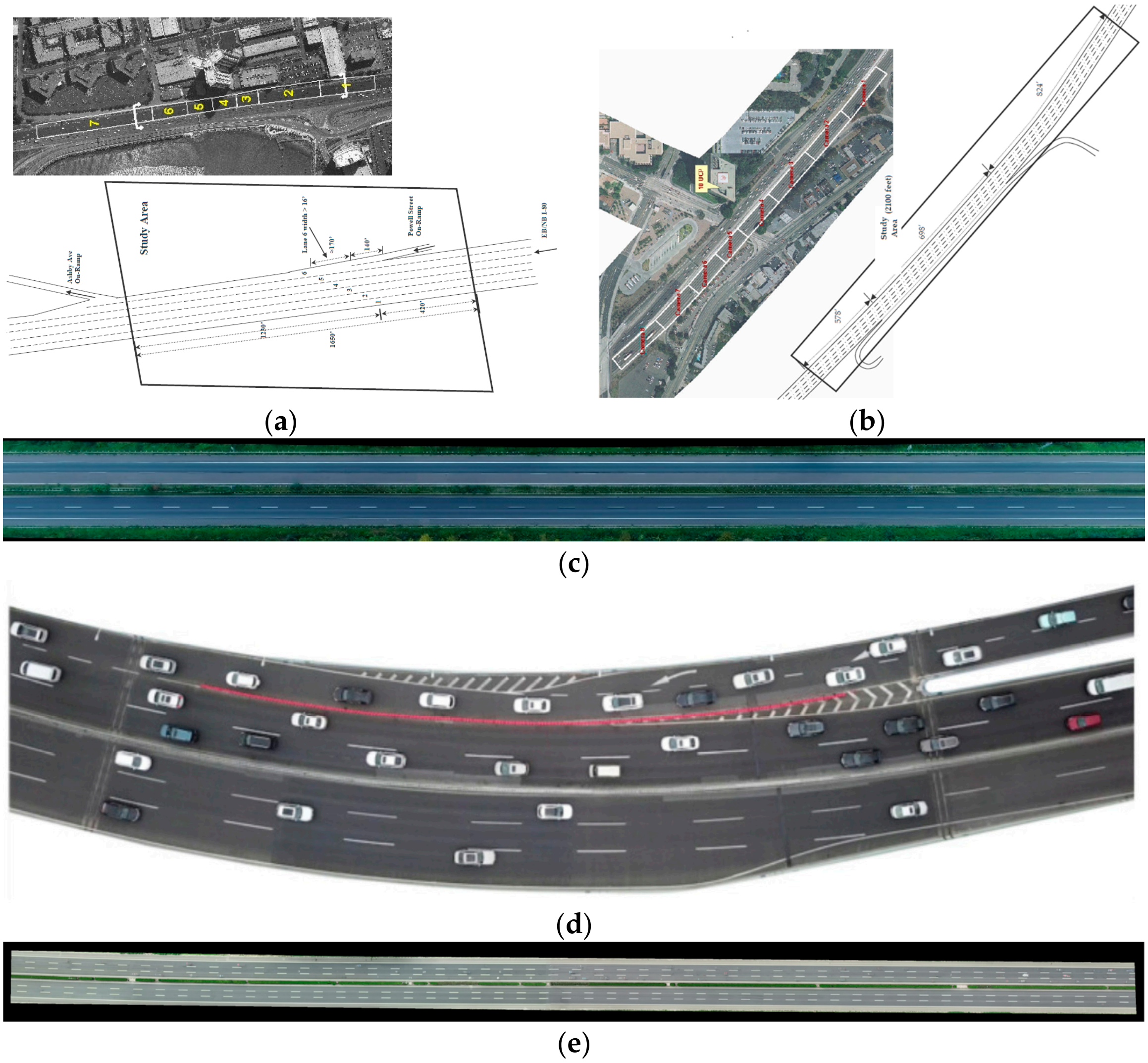
Before selecting or collecting a data set, one should first identify the requirements that must be met by a data set applicable to traffic modeling for simulation tests. Combined with the fidelity and complexity of the traffic model required for simulation tests, the data set requirements can be summarized as follows:

- To ensure that the traffic model accurately simulates real traffic behavior, the dataset should be collected without observational interference. That is, the road users are unaware of the observation in order to show the most natural and realistic movement and interaction state.
- To enable the traffic model to accurately understand the spatiotemporal interactions between road users, the dataset should be collected to ensure the completeness of the information. That is, the movement trajectories of all road users under the lane restriction should be completely recorded to ensure the completeness of the interaction information.
- To achieve a broader applicability and robustness of the traffic model, the data collection should be as diverse as possible, i.e., the data set should cover a wide range of densities, speeds, and risk levels.

2.2. Datasets

Based on the first two points of the above requirements, four highway subsets from natural driving trajectory datasets collected with a bird's eye view, which are from NGSIM [1], highD [5], INTERACTION [9], and CitySim [13], were selected for further analysis for the subsequent study. It is important to note that the subset from NGSIM and the subset from INTERACTION have almost no free-flow data due to significant traffic congestion and traffic fluctuation. The highD dataset and CitySim have been found to be less congested, with highD consisting mostly of free-flow data. Furthermore, we collected traffic trajectory data using LIDAR mounted on buildings throughout the day and meticulously organized it into Highway datasets. Table 1 shows the collection details for each of the datasets mentioned above, and Figure 1 shows the collection area of the subsets discussed in this paper. The accuracy and completeness in Table 1 are summarized from the findings of existing studies and the problems we found in processing the data.

Table 2 presents detailed information on the collection range, collection duration, and data volume of the highway part of the above dataset. Considering that highway driving involves both car-following and lane-changing (including merge-in and merge-out), Table 2 also provides the percentage of lane-changing vehicles in each dataset. The information in Table 2 is further explained below.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

Figure 1. Collection Scenarios of open-source datasets (**a**) NGSIM, I-80; (**b**) NGSIM US-101; (**c**) highD; (**d**) INTERACTION, DR_CHN_Merging_ZS; (**e**) CitySim, Freeway B.

Table 1. Comparison of bird's eye view naturalistic driving trajectory datasets.

Dataset	Country	Road Type	Collection	Frequency	Unit	Accuracy and Completeness
NGSIM	USA	Highway, Intersection,	Camera mounted on building,	10 fps	ft	False positive trajectory collisions and physically illogical vehicle speeds and accelerations [5,20].
highD	Germany	Highway	Drone	25 fps	m	Acceleration anomalies exist.
INTERACTON	International	Highway, Intersection, Roundabouts	Drone, Camera mounted on building	10 fps–30 fps	m	Acceleration anomalies exist.
CitySim	International	Highway, Intersection, Roundabouts	Drone	30 fps	ft	Acceleration anomalies exist.
Highway	China	Highway	LIDAR mounted on building	10 Hz	m	A few trajectories have missing location points. due to obscuration between vehicles.

Table 2. Comparison of highway datasets.

Dataset	Tracks	Lanes	Range (Meters)	Duration (Hours)	Lane-Changing/ Merge in/out Tracks	Ratio of Lane-Changing/ Merge in/out (%)
NGSIM	9207	6	602	1.35	2678	28.24
highD	10,9769	3 × 2	404	82.81	11,717	10.19
INTERACTON	4104	2	130	1.48	243	5.90
CitySim	6542	3 × 2	680	3.40	1562	23.88
Highway	21,191	3 × 2	110	6.49	3838	21.89

2.2.1. NGSIM

The highway sub-datasets in NGSIM are I-80 and US-101, both collected by cameras mounted on the tops of tall buildings. The collected road for the I-80 dataset is shown in Figure 1a. The roadway is a section of highway with a length of approximately 500 m and contains six one-way lanes, and the total collection time is 45 min. The US-101 trajectory dataset was collected from a section of highway with a length of approximately 640 m and contains six lanes (Figure 1b), and the data was collected for a total of 45 min. The above two datasets contain a total of 9207 vehicle trajectories, of which the number of lane-changing vehicles is 2678, representing 28.24% of the total number of vehicles. It is worth noting that both I-80 and US-101 have significant traffic congestion and traffic fluctuation.

2.2.2. highD

The highD dataset was released in 2018 and covered the trajectories of approximately 110,000 vehicles on German highways. The dataset records vehicle trajectories with a drone-mounted camera, which allows for much more flexible data collection compared to the NGSIM. The whole dataset contains 60 sub-datasets from different road sections, each covering a road section of about 420 m with 3 lanes in both directions (Figure 1c), and the collection time is up to 82 h. However, the speeds in the highD dataset are generally high, the driver behavior is simple, and the congestion level is low. The percentage of lane-changing vehicles is relatively low at 10.19%.

2.2.3. INTERACTION

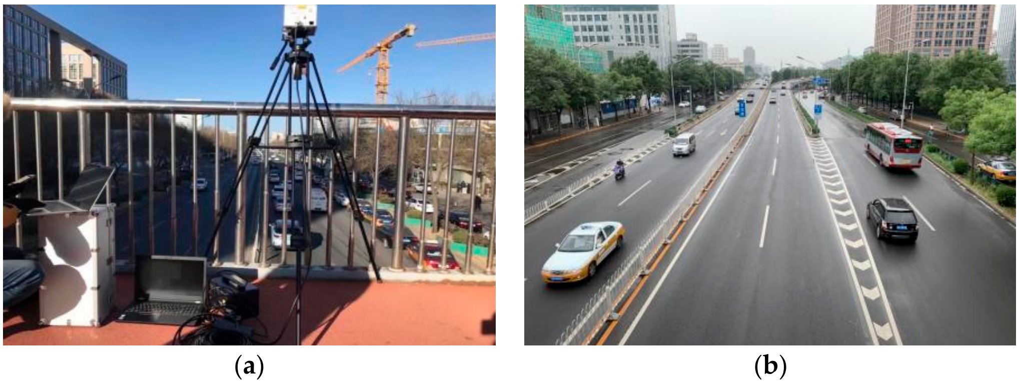
The INTERACTION dataset is a comprehensive, multi-country, multi-scenario dataset that focuses on describing the interaction behavior of road participants. The dataset includes data from scenarios such as signalized and unsignalized intersections, roundabouts, and highways. One of the sub-sets, DR_CHN_Merging_ZS, was obtained from a highway in China with a section length of about 130 m. The data on two lanes in the DR_CHN_Merging_ZS scenario were selected for the following study, and there were noticeably low-speed vehicles in the selected portion of the dataset. The data selected included 4104 vehicles, of which about 5.90% had lane-changing behavior. The duration of data collection was 1.48 h.

2.2.4. CitySim

The CitySim dataset is aimed at supporting traffic safety research and applications. The dataset collects aggressive and high-density vehicle interaction data in a wider range of scenarios, such as intersections and highways, by integrating images captured by multiple drones. For a more accurate record of the microscopic risk behavior of vehicles, the CitySim dataset provides more fine-grained vehicle boundary information. For this research, we choose the Freeway B sub-set of the CitySim dataset. It has vehicle trajectories from a section of the freeway that is approximately 680 m in length, a collection time of 3.40 h, and a total of 6542 vehicles, of which 23.88% have observed lane-changing behavior.

2.2.5. Highway

The Highway is a self-collected dataset using LiDAR on a section of a highway in Beijing with a length of about 100 m. The LiDAR was attached to a building during collection, and the data collection equipment and collection scenario are shown in Figure 2. The collected Highway dataset includes both peak and off-peak hours, with a total observation time of 6.5 h, and contains a total of 21,191 trajectories, of which 21.89% are identified as lane-changing trajectories. The high frequency of lane-changing is due to the fact that the observed road section contains both merge-in and merge-out intersections, and during peak hours, there are more vehicles merging in or out, resulting in a higher frequency of lane-changing. The Highway dataset is in the process of applying for its open-source license and will be released in the near future.



(**a**) (**b**)

Figure 2. Data Collection for The Highway dataset. (**a**) LiDAR equipment; (**b**) Collection scenario.

2.3. Preprocessing

Before analyzing the data, we first process the data in a unified manner. This included denoising, coordinate conversion, and unifying the units and sampling frequency of all datasets. The purpose of this is to avoid the influence of noise, road curvature, and differing units and sampling frequencies on the fairness of the evaluation results.

2.3.1. Denoising

In the process of vehicle trajectory data extraction, due to the measurement error of video or point cloud image, abnormal fluctuations of speed can easily occur, which in turn leads to local amplification when calculating acceleration. We adopt the wavelet denoising [27] method to solve the above problems. The method is based on the characteristics that the wavelet decomposition coefficients of the noise and the signal in different frequency bands have different intensities, the wavelet coefficients corresponding to the noise in each frequency band are removed, the wavelet decomposition coefficients of the original signal are retained, and then the wavelet reconstruction of the processed coefficients is performed to obtain the pure signal. In this paper, we apply the noise reduction to the speed, and after obtaining the smooth speed profile, the trajectory and acceleration profiles are obtained by integral and differential operations of the speed profile, respectively.

2.3.2. Frenet Coordinate

Due to road curvature, the Cartesian coordinate system is unable to represent the distance traveled by a vehicle and its offset relative to the lane centerline. To accurately determine the position of the vehicle relative to the lane centerline, the Frenet coordinate system (also known as the SL coordinate system) is introduced. The Frenet coordinate system is usually used in trajectory tracking [28], path following, trajectory planning, and prediction [29], and its basic principle is based on a series of reference points on the reference curve to find the nearest reference point for all the trajectory points of the vehicle and to calculate the relative position of the trajectory points with respect to the reference point. In this paper, by analyzing the principle of Frenet coordinate system transformation, it is considered that the method is also applicable to solving the problem of road type unification. Specifically, the depicted roads with curvature in Figure 1d,e are transformed into roads without curvature using the Frenet coordinate system. This results in vehicle motion being separated into longitudinal and lateral motion, facilitating the handling of vehicle-lane and vehicle-vehicle relative relationships and simplifying subsequent analysis and application.

The sampling frequency of the data was standardized to 10 Hz, which means the sampling interval was 0.1 s after denoising and coordinate conversion.

3. Risk

In this section, three risk indices based on trajectory datasets are introduced in detail. We first summarize the microscopic features utilized for analysis and define subsequent risk characteristics, including position, speed, and distance. The physical meaning and calculation of the microscopic features are shown in Table 3. It is worth noting that the trajectory of the vehicle is decoupled into longitudinal and lateral motions by the Frenet coordinate system transformation so that both velocity and distance in the microscopic features are decomposed into longitudinal and lateral directions.

Table 3. Microscopic features.

Symbol	Meaning	Calculation
l_i	Length of vehicle i	/
x_i	Longitudinal position of vehicle i	/
y_i	Lateral position of vehicle i	/
$v_{x,i}$	Longitudinal speed of vehicle i	$v_{x,i} = \frac{dx_i}{dt}$
$v_{y,i}$	Lateral speed of vehicle i	$v_{y,i} = \frac{dy_i}{dt}$
a	Acceleration of vehicle	$a_i = \frac{dv_{x,i}}{dt}$
d_i	Headway between vehicle i and the preceding vehicle $i-1$	$d_i = x_{i-1} - x_i$
th_i	Time headway between vehicle i and the preceding vehicle $i-1$ (TH)	$th_i = \frac{d_i}{v_{x,i}}$
ttc_i	Time to collision between vehicle i and the preceding vehicle $i-1$ (TTC)	$ttc_i = \frac{d_i - l_{i-1}}{v_{x,i} - v_{x,i-1}}$

To comprehensively evaluate the risk level of vehicle movements in the dataset, several risk indices are defined and quantified in this subsection. While TH and TTC and their variations have been widely used in risk assessment studies, TH and TTC are mostly used for transient evaluation of individual vehicle movements and are not sufficient for understanding the risk level of traffic. Therefore, this study expands on the development of indices for assessing traffic risk using TTC, Integrated Time to Collision (TIT) [30], and Deceleration rate to avoid crash (DRAC) [31], which now includes the Modified Integrated Time to Collision (MTIT), Modified Crash Potential Index (MCPI), and Modified Minimum Difference of Time-to-Conflict (MMDT). The indices above quantify collision risk in traffic, the intensity of actions taken to avoid collision risk, and the conflict intensity of lane-changing behavior.

3.1. Modified Integrated Time to Collision

The TTC is used to indicate the probability of a collision. Specifically, at the moment t when the speed of the vehicle i is greater than the speed of the preceding vehicle $i-1$ and the speed difference is fixed, the value of TTC represents the time before the collision between the two vehicles occurs. The smaller the value of TTC, the greater the risk of collision. TIT considers the length of the risk, defined as the integral of the difference between the TTC and the threshold value over the period when the TTC is less than the threshold value. A larger value of MTIT indicates a higher risk of collision; that is, both a short period of high risk and a long period of low risk must be considered. TIT is specified in this paper by the calculation method as follows:

$$\begin{gathered} tit_i^* = \sum_{t=0}^{T_i} (ttc^* - ttc_i(t)) \cdot \delta_i(t) \\ \delta_i(t) = \begin{cases} 1, & 0 < ttc_i(t) < ttc^* \\ 0, & else \end{cases} \end{gathered} \tag{1}$$

where tit_i^* denotes the TIT of the vehicle i under the threshold ttc^*. In this paper, $ttc^* = 20$ s. T_i is the duration of the vehicle i. The unit of TIT is s.

Considering that the interactions between vehicles that lead to risk occur over a certain duration and road, the transient metrics that consider only two vehicle interactions have limitations in evaluating complex traffic scenarios. Therefore, in this paper, we divide each processed dataset into traffic segments of duration T and consider the traffic segments as a system to extend the concept of transient risk and thus realize the risk assessment of the traffic segments.

On the basis of the definition of TIT, crash risk for a traffic segment Θ is in the following normalized form:

$$MTIT_\Theta = \frac{\sum_{i=1}^{n_{veh}} tit_i}{L_{lane} \cdot T \cdot n_{lane}} \tag{2}$$

where $MTIT_\Theta$ is the collision risk of traffic segment Θ that can be considered as the sum of the collision risks generated by all vehicles in the traffic segment over the duration T. In this paper $T = 10$ s. n_{veh} denotes the number of vehicles. Due to the different lane lengths and number of lanes in each dataset, m is adopted in the normalized form to avoid the different roads affecting the subsequent comparisons between different datasets, L_{lane} is the length of the lanes and n_{lane} is the number of the lanes. The unit of $MTIT_\Theta$ is s.

3.2. Modified Crash Potential Index

In addition to the risk of collision, the potential risk of traffic is also related to risk avoidance. On the one hand, the intensity of risk avoidance maneuvers reflects the severity of the conflict. On the other hand, risk avoidance maneuvers affect traffic stability and are one of the sources of traffic fluctuations. DRAC quantifies the minimum deceleration required by a vehicle to avoid a collision with the preceding vehicle, defined as,

$$drac_i(t) = \begin{cases} \frac{\left(v_{x,i}(t) - v_{x,i-1}(t)\right)^2}{2(d_i(t) - l_n)}, & v_{x,i}(t) > v_{x,i-1}(t) \\ 0, & else \end{cases} \tag{3}$$

The unit of DRAC is m/s^2. When the DRAC value exceeds the dynamic limits of the vehicle, it means there is a greater likelihood of a collision. Except for the different physical meanings, DRAC is highly correlated with TTC, which is also a transient risk index. Therefore, inspired by the Crash Potential Index (CPI) [32], a crash avoidance index MCPI is proposed by combining the actual acceleration and DRAC of the vehicle. Specifically, the MCPI denotes the integral in time of the difference between the actual deceleration and the minimum deceleration required to avoid a collision, is calculated as,

$$\begin{gathered} mcpi_i = \sum_{t=0}^{T_i} (d_i(t) - drac_i(t)) \cdot \eta_i(t) \\ \eta_i(t) = \begin{cases} 1, & v_{x,i}(t) > v_{x,i-1}(t) \\ 0, & else \end{cases} \end{gathered} \tag{4}$$

where $d_i(t)$ denotes the actual deceleration of vehicle i at time t. From the Equation (4), if the deceleration $d_i(t) < 0$, means that the vehicle i is accelerating, therefore, the smaller the value of the MCPI, the more drastic the braking action required by the vehicle to avoid a collision. The unit of MCPI is $\mathrm{m/s^2}$. The total avoidance risk of the traffic segment Θ is defined as:

$$MCPI_{\Theta} = \frac{\sum_{i=1}^{n_{veh}} mcpi_i}{L_{lane} \cdot T \cdot n_{lane}} \tag{5}$$

The unit of $MCPI_{\Theta}$ is $\mathrm{m/s^2}$.

3.3. Modified Minimum Difference of Time-to-Conflict

MMDT is a risk index defined for lane-changing behavior, which is a variant of the Minimum Difference of Time to Conflict Point (MDTTC) [23] between the lane-changing vehicle and the rear vehicle in the target lane. The smaller the MDTTC value, the higher the probability of a collision. The MDTTC is calculated as follows,

$$mdttc_i = \min|ttcp_i(t) - ttcp_{tr}(t)| \tag{6}$$

where $ttcp_i$ denotes the TTC to conflict point of the lane-changing vehicle i at time t, $ttcp_{tr}$ is the TTC to conflict point of the rear vehicle tr on target lane, $mdttc_i$ is the minimum difference of TTC of vehicle i. The unit of MDTTC is s. As a variant of MDTTC, the MMDT is calculated as follows,

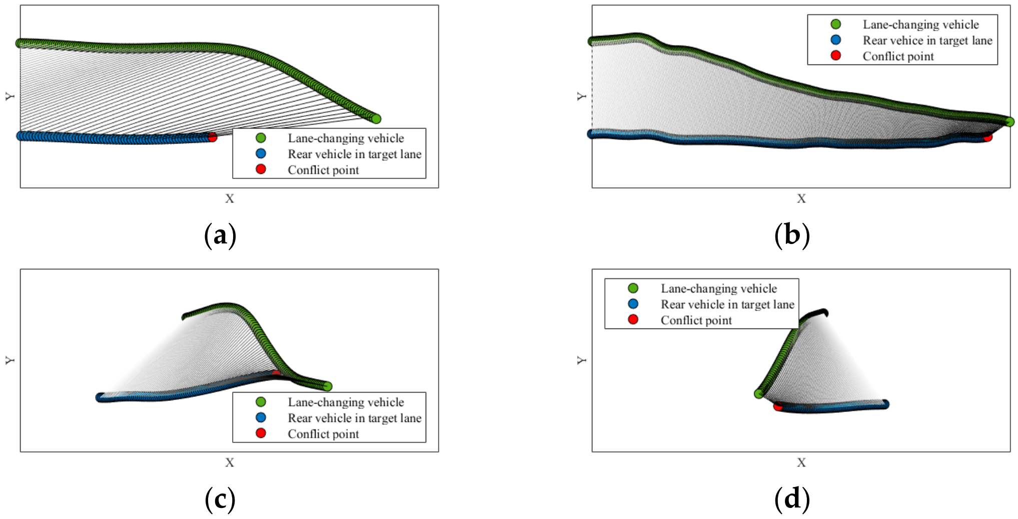
$$mmdt_i = \frac{1}{(mdttc_i)^{1/\alpha}} \tag{7}$$

where α is an adjustment parameter, and $\alpha = 4$ in this paper.

Since the trajectories in the dataset are irregular, the conflict point in this paper is defined as the intersection of the trajectory of the lane-changing vehicle and the trajectory of the rear vehicle in the target lane or the point with the closest distance between the trajectories. Specifically, the conflict point is defined directly as the position point at t^* on the trajectory of vehicle tr, and the conflict point is $p_c = (x_{tr}(t^*), y_{tr}(t^*))$. t^* is determined as follows,

$$t^* = arg \min_q \sqrt{(x_i(t) - x_{tr}(q))^2 + (y_i(t) - y_{tr}(q))^2} \tag{8}$$

In order to be a more intuitive way to show the correspondence between vehicle lane-changing trajectories and MMDT values, four representative high-speed and low-speed lane-changing scenarios from the four open-source datasets mentioned in the above section are selected as examples in Figure 3. Two high-speed lane-change scenarios from highD and CitySim are shown in Figure 3a,b, and two low-speed lane-change scenarios from NGSIM and INTERACTION are shown in Figure 3c,d. The longitudinal range of the figure is 200 m, and the lateral range is 7 m. The two points connected by solid lines indicate the positions of the lane-changing vehicle and the rear vehicle of the target lane at the same moment, and the red point is the conflict point. By observing the connection relationship of the solid lines, it can be seen that the lane-changing vehicle from highD maintains a larger longitudinal distance from the rear vehicle of the target lane at the later stage of the lane-changing process and has a larger time difference for the conflict, corresponding to an MMDT value of 0.7503. In the lane-change scenario from CitySim, with the lateral distance between the two vehicles reduced, the distance in the longitudinal direction is not increased, resulting in a greater possibility of collision, and the MMDT value is 5.6962. In the low-speed scenario, the change in MMDT value also follows the above trend. In the low-speed scenario, when the longitudinal distance between the two vehicles near the conflict point is approximated to that in the high-speed scenario (Figure 3b,d), the corresponding MMDT decreases, which means that the collision risk for lane-changing vehicles is generally lower in low-speed conditions, which is in the same way as the real-world law.



(**a**) (**b**) (**c**) (**d**)

Figure 3. Lane-changing Scenarios (**a**) highD, MMDT = 0.7503; (**b**) CitySim, MMDT = 5.6962; (**c**) NGSIM, MMDT = 0.7909; (**d**) INTERACTION, MMDT = 2.9790.

Based on the definition of MMDT, the risk to the traffic segment Θ due to lane-changing behavior is,

$$MMDT_{\Theta} = \frac{\sum_{i=1}^{n_{lc}} mmdt_i}{L_{lane} \cdot T \cdot n_{lane}} \tag{9}$$

where n_{lc} denotes the number of lane-changing vehicles.

4. Comprehensive Assessment of Trajectory Data from Multiple Perspective

To address the third requirement of traffic modeling data for simulation testing mentioned in Section 2.1, we further analyzed the diversity and balance of risk levels in the data sets.

Since traffic has time-varying attributes and each dataset is collected at different lengths, a fixed duration is used to segment the dataset in this paper in order to impartially represent the heterogeneity of traffic risk levels. Specifically, a single dataset is divided into 10-s segments by using a sliding window. At the same time, in order to avoid the effect of the difference in data volume of different datasets on the distribution of risk indices, we perform a uniform random sampling for each dataset and select 370 traffic segments for further analysis.

Figure 4 shows the correlation and distribution plots of the risk indices of the traffic segments. The three sub-plots on the diagonal shows the distribution of $MTIT_{\Theta}$, $MCPI_{\Theta}$ and $MMDT_{\Theta}$ from top left to bottom right, while the scatter plot below the diagonal show the correlation of $MTIT_{\Theta}$ with $MCPI_{\Theta}$, $MTIT_{\Theta}$ with $MMDT_{\Theta}$ and $MCPI_{\Theta}$ with $MMDT_{\Theta}$ from top to bottom and left to right, respectively. From the distribution plots, it can be seen that almost all risk indices are distributed around 0. HighD and Highway are relatively obvious, with distributions showing high and narrow peaks. The correlation plot also shows the unbalanced distribution of the individual data sets with respect to the risk indices, i.e., most of the risk points are concentrated in a certain range. In addition, the correlation plot shows that $MTIT_{\Theta}$ is linearly correlated with $MCPI_{\Theta}$, but the correlation level varies among the datasets. And there is no correlation between $MMDT_{\Theta}$ and the other two indices.

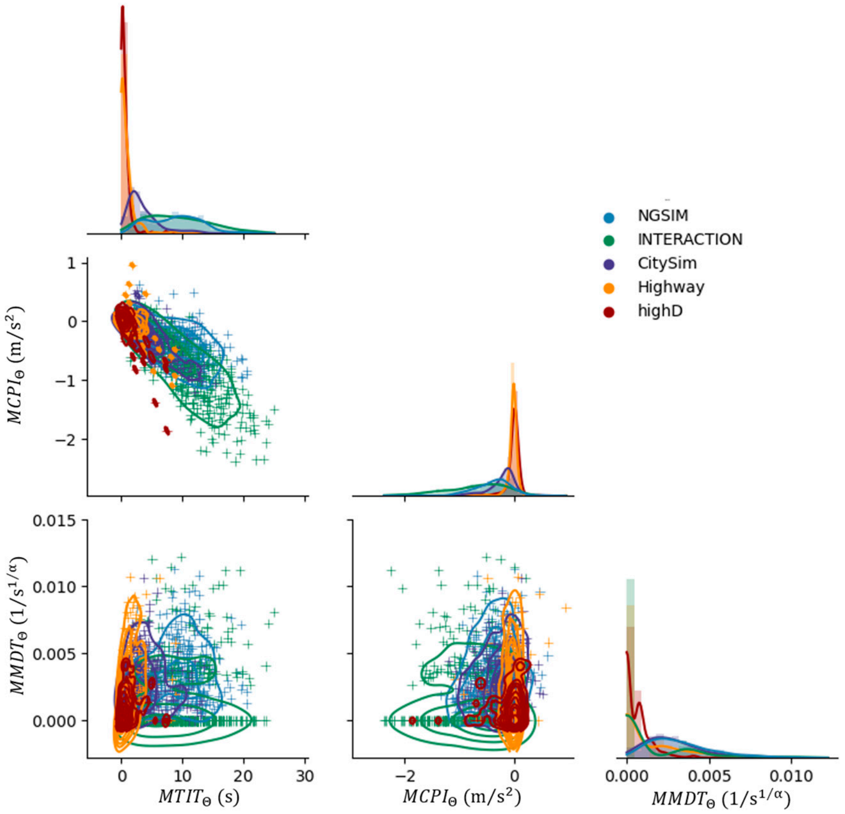


Figure 4. The correlation and distribution plots of the risk indices.

For diversity, we use basic statistics such as Mean (MA), Standard Deviation (SD), and Range (RNG) to statistically analyze individual risk indices. These statistics help synthesize the overall distribution and variation of individual risk indices in datasets. RNG describes the extended range of the variable, which is obtained by subtracting the minimum value from the maximum value of the variable:

$$rng = \max(I) - \min(I) \tag{10}$$

where I denotes one of the risk indices. Then, the Average Minimum Euclidean Distance (AMED) is introduced to quantify the risk level differences in the whole data set. The definition of AMED is based on the quantification of vector dissimilarity by Euclidean distance. The risk vectors consisting of $MTIT_\Theta$, $MCPI_\Theta$ and $MMDT_\Theta$ are considered as points in a three-dimensional space, and as the risk level similarity between two points increases, the two points become closer. The specific quantification method of AMED is to find another traffic segment with the most similar risk level for each traffic segment in the data set and calculate their Euclidean distances as the Minimum Euclidean Distances (MED) of that traffic segment, and the AMED is the average value of the minimum Euclidean distances of all traffic segments. In the specific calculation, the risk indices should be first normalized before calculating the AMED because the range differences of the three risk indices are different:

$$I_{norm} = \frac{I - \min(I)}{\max(I) - \min(I)} \tag{11}$$

After normalization, AMED is calculated as follows,

$$AMED = \frac{\sum_{i=1}^{N} \underset{j}{argmin}\left(Euclidean\left(R_{\Theta_i}, R_{\Theta_j}\right)\right)}{N}, j \in [1, N], j \neq i \tag{12}$$

where N is the number of traffic segments, $R_{\Theta_i} = \left[MTIT_{\Theta_i}, MCPI_{\Theta_i}, MMDT_{\Theta_i}\right]$ is the risk vector of traffic segment Θ_i, $Euclidean(a, b)$ is the Euclidean distance calculation of vector a and vector b. The result and comparison of risk diversity is shown in Table 4. In Table 4, bold numbers are maximum values and underlined numbers are minimum values.

Table 4. Compare risk diversity.

Dataset	Index	MA	STD	RNG	AMED
NGSIM	$MTIT_\Theta$	8.7503	4.1605	18.8991	**0.0347**
	$MCPI_\Theta$	−0.3738	0.3113	1.9687	
	$MMDT_\Theta$	**0.0033**	**0.0026**	0.0113	
highD	$MTIT_\Theta$	0.5761	0.8493	7.4554	0.0077
	$MCPI_\Theta$	−0.0129	0.1586	2.1176	
	$MMDT_\Theta$	0.0006	0.0008	0.0042	
INTERACTON	$MTIT_\Theta$	**9.2416**	**5.2374**	**24.9199**	0.0329
	$MCPI_\Theta$	**−0.7030**	**0.5525**	**2.6264**	
	$MMDT_\Theta$	0.0014	0.0024	**0.0122**	
CitySim	$MTIT_\Theta$	3.5680	2.7132	15.0046	0.0262
	$MCPI_\Theta$	−0.2043	0.2478	1.7290	
	$MMDT_\Theta$	0.0028	0.0017	0.0106	
Highway	$MTIT_\Theta$	0.8225	1.1902	8.7480	0.0123
	$MCPI_\Theta$	−0.0157	0.1258	2.0478	
	$MMDT_\Theta$	0.0014	0.0022	0.0107	

Through the MA, STD, and RNG in Table 4, we find that all traffic segments in the INTERACTION dataset are generally characterized by a high level of risk and a wide range of risk, followed by NGSIM. This indicates that when congestion is present in the traffic scenario, the corresponding dataset tends to contain more higher-interaction and higher-risk driving behaviors and also outperforms the other datasets in terms of diversity. In contrast, the risk level of all the traffic segments in the highD dataset is much lower, which may be related to it collected from high-speed and low-density scenarios. The AMED values in Table 4 further confirm the aforementioned viewpoint, which indicates that NGSIM and INTERACTION have an advantage in terms of diversity from the perspective of the similarity of risk vectors between traffic segments. NGSIM has an AMED value of 0.0347, while INTERACTION is slightly lower at 0.0329. Among all datasets, highD has the lowest level of diversity, at 0.0077.

We analyze the balance of the data using MED values. The MED of all traffic segments in a given data set is sorted from low to high to obtain the curve shown in Figure 5. From the definition of MED, if the curve is far from the horizontal axis and tends to be horizontal, it indicates that all traffic segments are widely and evenly distributed in space. On the contrary, when the curve is closer to the horizontal axis, and the slope increases rapidly, it means that most of the traffic segments in the dataset are clustered together to form a category with a higher degree of similarity, while there are traffic segments with a greater difference in similarity wandering away from the category. The dataset exhibits an imbalance. From Figure 5, it can be seen that the MED curves of NGSIM, CitySim, and INTERACTION have similar trends, with NGSIM having the best data balance. In contrast, the highD and Highway datasets are unbalanced. In addition, AMED is flawed as an average form, and when there is a very large MED, AMED deviates from the actual diversity. However, from what we see in Figure 5, none of the curves have abnormally large MED values, so we believe that AMED is consistent with the actual diversity in this paper.

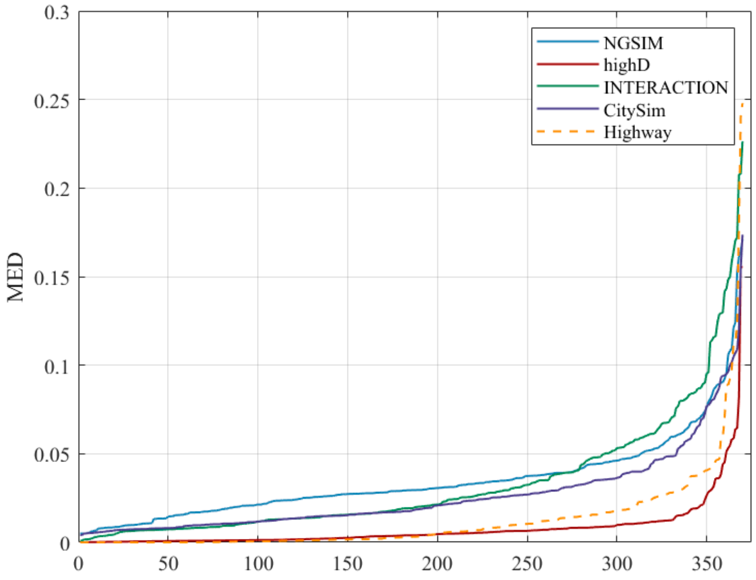


Figure 5. Med curves.

5. Conclusions

This paper provides a detailed analysis of natural driving trajectory datasets from highways for traffic simulation in autonomous driving tests. We quantify the traffic risk levels of highway scenarios in NGSIM, highD, INTERACTION, CitySim, and a self-collected dataset named Highway by introducing systematic risk indices and comprehensively assessing the diversity and balance of the datasets using scatter indices. Through the above indices, we comprehensively evaluated and recognized the differences in risk level, diversity, and balance of highway scenarios among different datasets, which to some extent provided a powerful preliminary study for the construction of high-risk traffic scenarios and the establishment of traffic models with strong generalization ability in the virtual simulation test of autonomous driving.

Based on the evaluation results, we have summarized some of the ideas and flaws of this paper. Further verification can be conducted in future research. Firstly, we believe that for virtual testing of autonomous driving, when the proportion of congested flows in the data set is higher, the traffic has a higher level of risk and diversity and may be more suitable to be used for testing the upper limit of the capability of the autonomous driving system. Secondly, when collecting and using data, it is important to consider not only the total amount of data but also the potential for homogenization. In this paper, we have discovered that a high percentage of free flow may be a contributing factor to data homogenization. Future studies should aim to develop more efficient methods for data collection and analysis. Finally, this paper only considers highway scenarios. In future studies, we will explore data assessment methods for urban road scenarios.

Author Contributions: Conceptualization, Y.W. and W.D.; methodology, R.Z., Y.W. and J.D.; software, R.Z.; validation, R.Z., Y.W. and J.D.; formal analysis, W.D.; investigation, R.Z.; resources, Y.W. and J.D.; data curation, R.Z.; writing—original draft preparation, R.Z.; writing—review and editing, W.D.; visualization, R.Z.; supervision, Y.W.; project administration, J.D.; funding acquisition, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by “Lingyan” R&D Program of Zhejiang Province (ID: 2023C01238) and “Jianbing” R&D Program of Zhejiang Province (ID: 2023C01133).

Data Availability Statement: The original contributions presented in the study are included in the article material, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: Juan Ding is an employee of PanoSim Technology Limited Company. The paper reflects the views of the scientists and not the company.

References

1. Traffic Analysis Tools: Next Generation Simulation—FHWA Operations. Available online: https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm (accessed on 10 August 2023).
2. Yeo, H. *Asymmetric Microscopic Driving Behavior Theory*; University of California: Berkeley, CA, USA, 2008.
3. Zheng, J.; Suzuki, K.; Fujita, M. Car-following behavior with instantaneous driver–vehicle reaction delay: A neural-network-based methodology. *Transp. Res. Part C Emerg. Technol.* **2013**, *36*, 339–351. [CrossRef]
4. Leclercq, L.; Chiabaut, N.; Laval, J.; Buisson, C. Relaxation phenomenon after lane changing: Experimental validation with NGSIM data set. *Transp. Res. Rec.* **2007**, *1999*, 79–85. [CrossRef]
5. Krajewski, R.; Bock, J.; Kloeker, L.; Eckstein, L. The Highd Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2118–2125.
6. Moers, T.; Vater, L.; Krajewski, R.; Bock, J.; Zlocki, A.; Eckstein, L. The exiD Dataset: A Real-World Trajectory Dataset of Highly Interactive Highway Scenarios in Germany. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 958–964.
7. Krajewski, R.; Moers, T.; Bock, J.; Vater, L.; Eckstein, L. The Round Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
8. Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; Eckstein, L. The Ind Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1929–1934.
9. Zhan, W.; Sun, L.; Wang, D.; Shi, H.; Clausse, A.; Naumann, M.; Kummerle, J.; Konigshof, H.; Stiller, C.; de La Fortelle, A.; et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv* **2019**, arXiv:1910.03088.
10. Makridis, M.; Mattas, K.; Anesiadou, A.; Ciuffo, B. OpenACC. An open database of car-following experiments to study the properties of commercial ACC systems. *Transp. Res. Part C Emerg. Technol.* **2021**, *125*, 103047. [CrossRef]
11. Breuer, A.; Termöhlen, J.A.; Homoceanu, S.; Fingscheidt, T. openDD: A Large-Scale Roundabout Drone Dataset. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
12. Spannaus, P.; Zechel, P.; Lenz, K. Automatum Data: Drone-Based Highway Dataset for the Development and Validation of Automated Driving Software for Research and Commercial Applications. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1372–1377.
13. Zheng, O.; Abdel-Aty, M.; Yue, L.; Abdelraouf, A.; Wang, Z.; Mahmoud, N. CitySim: A drone-based vehicle trajectory dataset for safety oriented research and digital twins. *arXiv* **2022**, arXiv:2208.11036. [CrossRef]
14. Barmpounakis, E.; Geroliminis, N. On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment. *Transp. Res. Part C Emerg. Technol.* **2020**, *111*, 50–71. [CrossRef]
15. Seo, T.; Tago, Y.; Shinkai, N.; Nakanishi, M.; Tanabe, J.; Ushirogochi, D.; Kanamori, S.; Abe, A.; Kodama, T.; Yoshimura, S.; et al. Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: Zen Traffic Data. In Proceedings of the 2020 International Symposium on Transportation Data and Modelling, Ann Arbor, MI, USA, 24–26 June 2020.
16. Ma, W.; Zhong, H.; Wang, L.; Jiang, L.; Abdel-Aty, M. MAGIC dataset: Multiple conditions unmanned aerial vehicle group-based high-fidelity comprehensive vehicle trajectory dataset. *Transp. Res. Rec.* **2022**, *2676*, 793–805. [CrossRef]
17. Punzo, V.; Borzacchiello, M.T.; Ciuffo, B. On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transp. Res. Part C Emerg. Technol.* **2011**, *19*, 1243–1262. [CrossRef]
18. Coifman, B.; Li, L. A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset. *Transp. Res. Part B Methodol.* **2017**, *105*, 362–377. [CrossRef]
19. Montanino, M.; Punzo, V. Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. *Transp. Res. Part B Methodol.* **2015**, *80*, 82–106. [CrossRef]
20. Thiemann, C.; Treiber, M.; Kesting, A. Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data. *Transp. Res. Rec.* **2008**, *2088*, 90–101. [CrossRef]
21. Schneider, P.; Butz, M.; Heinzemann, C.; Oehlerking, J.; Woehrle, M. Scenario-Based Threat Metric Evaluation Based on the Highd Dataset. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 213–218.
22. Ponn, T.; Breitfuß, M.; Yu, X.; Diermeyer, F. Identification of Challenging Highway-Scenarios for the Safety Validation of Automated Vehicles Based on Real Driving Data. In Proceedings of the 2020 Fifteenth International Conference on Ecological Vehicles and Renewable Energies (EVER), Monte-Carlo, Monaco, 10–12 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–10.
23. Zhan, W.; Sun, L.; Wang, D.; Jin, Y.; Tomizuka, M. Constructing a highly interactive vehicle motion dataset. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 6415–6420.

24. Zhang, J.; Lee, J.; Abdel-Aty, M.; Zheng, O.; Xiao, G. Enhanced index of risk assessment of lane change on expressway weaving segments: A case study of an expressway in China. *Accid. Anal. Prev.* **2023**, *180*, 106909. [CrossRef] [PubMed]
25. Zhang, Y.; Chen, Y.; Gu, X.; Sze, N.N.; Huang, J. A proactive crash risk prediction framework for lane-changing behavior incorporating individual driving styles. *Accid. Anal. Prev.* **2023**, *188*, 107072. [CrossRef]
26. Srinivasan, A.R.; Schumann, J.; Wang, Y.; Lin, Y.S.; Daly, M.; Solernou, A.; Zgonnikov, A.; Leonetti, M.; Billington, J.; Markkula, G. The COMMOTIONS Urban Interactions Driving Simulator Study Dataset. *arXiv* **2023**, arXiv:2305.11909.
27. Baleanu, D. (Ed.) *Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology*; InTech: London, UK, 2012. [CrossRef]
28. Sun, Y.; Ren, D.; Lian, S.; Fu, S.; Teng, X.; Fan, M. Robust path planner for autonomous vehicles on roads with large curvature. *IEEE Robot. Autom. Lett.* **2022**, *7*, 2503–2510. [CrossRef]
29. Vogl, C.; Sackmann, M.; Kürzinger, L.; Hofmann, U. Frenet coordinate based driving maneuver prediction at roundabouts using LSTM networks. In Proceedings of the 4th ACM Computer Science in Cars Symposium, Feldkirchen, Germany, 2 December 2020; pp. 1–9.
30. Minderhoud, M.M.; Bovy, P.H. Extended time-to-collision measures for road traffic safety assessment. *Accid. Anal. Prev.* **2001**, *33*, 89–97. [CrossRef] [PubMed]
31. Archer, J. Indexes for Traffic Safety Assessment and Prediction and Their Application in Micro-Simulation Modelling: A Study of Urban and Suburban Intersections. Ph.D. Thesis, KTH, Stockholm, Sweden, 2005.
32. Cunto, F.J.C.; Saccomanno, F.F. Microlevel Traffic Simulation Method for Assessing Crash Potential at Intersections. In Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 21–25 January 2007.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.