

Trajectory data collecting and sharing: Where are we? Where shall we go?

Christine Buisson

`christine.Buisson@univ-eiffel.fr`

Université Gustave Eiffel

Rim Slama

École Nationale des Travaux Publics de l'État

Zuduo Zheng

University of Queensland

Mohammad Ali Arman

KU Leuven

Benjamin Coifman

The Ohio State University

Biagio Ciuffo

Joint Research Centre

Derek Gloudemans

Vanderbilt University

Samer Hamdar

George Washington University

XiangWang Hu

Soochow University

Junyi Ji

Vanderbilt University

Rui Jiang

Beijing Jiaotong University

Abhijnan Maji

Indian Institute of Technology Roorkee

Michalis Makridis

ETH Zurich

Hani Mahmassani

Northwestern University

Marcello Montanino

University of Naples Federico II

Vincenzo Punzo

University of Naples Federico II

Chris Tampere

KU Leuven

Alireza Talebpour

University of Illinois Urbana-Champaign

Junfang Tian

Tianjin University

Dan Work

Vanderbilt University

Shi-Teng Zheng

Beijing Jiaotong University

Research Article

Keywords: trajectory datasets, data collection methods, traffic automation, reproducible research

Posted Date: October 27th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7497430/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Trajectory data collecting and sharing: Where are we? Where shall we go?

Abstract

Purpose: Many trajectories data sets have been openly provided by different teams worldwide during the last decade. The experimental design, data acquisition setup, sensor placement, data analysis and presentation is biased to the ultimate scope of research, financial support, and experimental constraints. While this is understandable, there is a clear need for the adoption of a common and generic template that will act as an umbrella for the presentation of any new dataset. This is important to ensure transparency regarding data acquisition and limitations, applicability in terms of research investigations, complementarity with existing available resources and the realization of benchmarks for different research topics.

This paper aims to contribute towards this cause.

Method: A common template is used to describe each of recent nine datasets, representative of different data collection methods and research purposes. For each dataset, a dedicated section identifies the primary sources of data errors or incompleteness. In addition, a methodological framework that systematically retraces the full process of trajectory dataset collection, from preparation to sharing, has been identified. This comprehensive approach made it possible to draw lessons and formulate proposals at each stage of the dataset life cycle. Finally, a collective discussion process among all the authors was conducted to define future data collection methods and how to improve data sharing.

Results: This paper produces a structured synthesis of lessons and key observations drawn from open-source trajectory datasets. It develops a taxonomy that retraces the full lifecycle of trajectory dataset collection, from preparation to sharing, supported by a critical discussion of current practices. Building on this framework, the paper delivers recommendations and discussion at two levels: (i) general recommendations on data collection and sharing, and (ii) recommendations for new types of trajectory datasets.

Discussion: Although the datasets presented and discussed here are only a subpart of the numerous undergoing efforts for trajectory data collection and sharing across the globe, their representativeness is sufficient for a clear definition of guidelines for future initiatives, helping the community to build richer, more interoperable datasets without repeating past limitations.

Keywords: trajectory datasets; data collection methods; traffic automation; reproducible research

1 Introduction

Understanding the microscopic behavior of traffic is key to developing efficient and safe traffic management systems, with or without autonomous vehicle technologies. This detailed understanding allows precise microscopic modeling of traffic dynamics, safety studies, and emission and energy consumption prediction. A fuel for those aims is the access to accurate and high-resolution positioning data: the vehicle trajectory data.

The term "trajectory data" is ambiguous, as it can refer to different types of information: for example, spatiotemporal information collected from mobile phones tracking the movements of a subset of travelers across an entire urban area, or data that records the path of a single vehicle. For recent examples, see papers of [1] for the first case of travelers data, and for individual cars see [2], in the case of taxis, and [3], where authors fuse many independent data collection efforts for individual autonomous vehicles trajectory data into a unique database.

Globally, trajectory data have, and could be, in the near future, used for many research tasks. To group them into broad categories, we can define the following clusters:

- Microscopic driving behavior studies (MDB), targeting the driver behavior, the instability creation and propagation, the platoon creation and the analysis of vehicle control;
- Traffic flow theory (TFT), where the observation of the fundamental diagram, hysteresis, and oscillations, as well as the definition of capacity, are the objectives;
- Safety and risk analysis (SR), with the analysis of accident occurrence and the estimation of surrogate metrics;
- Control and ITS (Ctrl), with a large set of researches devoted to improving the vehicle automation, the setup of urban intersections, including non-motorized modes, and the control of highways (ramp metering, HOV, variable speed limits, lane usage restrictions, etc.);
- Machine learning and forecasting (MLF), for which large datasets covering the complete diversity of scenarios and the camera image sources are valuable inputs;
- Network-level operations (NLO), with all the researches related to the understanding of the mode and the route choices of individuals traveling in a wide urban area;
- Environmental and energy impacts (EEI), with all studies related to understanding how the vehicle-driver pair consumes the energy, and what the related emissions of diverse pollutants, including noise, are. This request requires high-frequency data.

In the sequel of this paper, we will here limit ourselves to the case of trajectory datasets containing not only the precise positions and speeds of a subject vehicle, but also all information permitting to understand and thus model the vehicle-driver pair behavior: the trajectories of all vehicles immediately surrounding it and the infrastructure description. Indeed, this information contains the explanatory variables of the vehicle-driver pair decisions of acceleration, deceleration, and lane change decisions and executions.

The collection of trajectory data has been integral to traffic modeling since the earliest days of the field. The figure 1 presents broadly the three periods.

During the first period, the data collection methods were either manual or based on rudimentary devices. To the best of our knowledge, the earliest measurements of vehicle trajectory-like data were the car following studies in the late 1950's that collected speed and inter-vehicle distance, with the first study reported in [4]. In 1975, Gerlough and Huber, in their handbook *Traffic Flow Theory* [5], reported (p. 94) on the experiments described in [6, 7]: *"A number of car-following experiments were performed in the Holland and Lincoln Tunnels in New York and at the General Motors test track. Two vehicles were connected by a length of wire on a reel, permitting direct measurement of spacing and relative velocity simultaneously with measurements of the acceleration and velocity of the following vehicle. Different drivers were given instruction to follow the lead vehicle as they would under normal driving conditions."* Although [8] argue that the data collection system in [4] actually influenced the car-following behavior, those early methods did provide interesting insights—for example, showing that the reaction time T *"ranged from 1.0 to 2.2 sec."*

Treiterer and Myers (1974) [9] is another seminal work in traffic flow theory. Their data collection used high-speed aerial photography and manual data reduction to track hundreds of vehicles on southbound I-71 in Columbus, Ohio, USA, in the summer of 1967, using a helicopter moving with the traffic to create a moving surveillance region 3–5 km long.

It is worth noting that two distinct approaches were used to measure the trajectories of a subject vehicle and its leader: intra-vehicular or on-board data collection in the earlier studies, and external measurements via image recording in [9]. This two types of data collections are still present and will be illustrated later on. One can also underline the lack of data sharing before 2000.

To summarize, during this period (between 1950 and 2000), the manual nature of data collection at the time made precise measurement very difficult, precluding

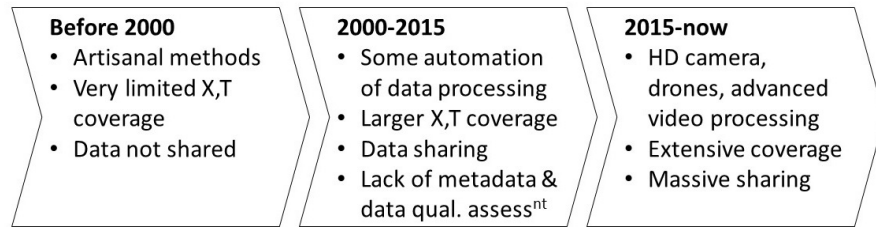


Figure 1: A synthesis of the three periods of trajectory data collection.

the collection of large datasets, the provision of high-quality data or even to enable an accurate quality assessment; nevertheless, the foundational observations essential for the development of traffic flow theory were made, and they are still valid more than 50 years later.

During the second period, between the late 1970s and 2010, the motivations for collecting more trajectory data stemmed from two main origins. On the one hand, the emergence of computer-based tools generated new needs for calibration studies. The development of microscopic simulation tools started in the late 1970s in the USA [10], [11], expanded through the 1980s, and reached a steady pace during the 1990s [12]. A similar trend occurred in Europe. According to [13], the first coding of the ancestor of VISSIM was done in 1983. The development of SUMO started in 2001 [14], while the earliest work on Aimsun can be traced back to Ferrer and Barcelo's presentation of AIMSUN2 [15]. Those tools, however, initially lacked proper microscopic calibration and validation.

On the other hand, the advent of moving object geolocalization technologies, along with advances in image collection (such as digital cameras with higher resolution) and video processing methods, made it possible to create larger datasets. A first attempt was conducted under the umbrella of the US Federal Highway Administration [16]. The camera was located on a airplane, and the data frequency acquisition was 1 Hz. A set of bottlenecks were observed, with lengths of hundreds of meters and one hour of recording per location. Although announced as shared in this report, the data is now difficult to access, more than 40 years after collection.

A landmark initiative was the launching of the Next Generation Simulation (NGSim) program in 2004, with the first release of its open access database in 2005 (now available here [17]), which had a tremendous impact on trajectory-based research. For NGSim data collection, a set of cameras was mounted on top of tall buildings, requiring the stitching of multiple images. This, as we will see in section 7, introduced measurement errors in both lateral and longitudinal positions

for the trajectories spanning over more than one image. Meanwhile, several other notable efforts were made to collect large volumes of vehicle trajectory data. For instance, the TU Delft team put the extensive efforts over many years [18–20], using a helicopter combined with a posterior image stabilization process to address the challenge of stitching multiple images. Similarly, the French Mocopo team also undertook a similar effort, though on a smaller scale [21].

Efforts were also made in on-board data collection. For example, [22, 23] reports that trajectories were collected in 2002-2003 with differential kinematic GPS for a platoon of four vehicles "along urban and extra-urban roads under different traffic conditions", in areas surrounding Naples, IT.

Over the last period, the trajectory data collection efforts in the traffic flow community have been dedicated to building much larger and openly shared datasets. This shift has been enabled by the decrease in the cost of data storage and processing, along with the rise of the impressive efficiency of convolutional neural networks (such as YOLO, which has undergone multiple improvements in recent years; for a pedagogical synthesis see [24]). The availability of such datasets has fueled research at multiple scales. For a comprehensive overview of the advances made possible through the systematic use of trajectory data, we refer the reader to [8]. Another relevant contribution is [25], which presents a comprehensive perspective on the research opportunities created by trajectory dataset, including simulator studies.

As evidenced by the studies reviewed, trajectory datasets have progressively evolved in both scale and complexity. However, important types of data still remain to be collected. This situation calls for a clearer understanding of what lessons can be learned from existing datasets, what kinds of data are still missing, and how future collections should be designed to address current limitations and meet emerging research and policy needs. At the same time, the diversity of existing approaches complicates systematic use and comparability, and underlines the need for clearer guidance

on how such datasets should be designed and disseminated so as to foster reproducibility and shared use within the research community.

To address these challenges, we gathered a set of open trajectory datasets reflecting different acquisition strategies and engaged in exchanges with their authors in order to assess their methodological choices and practical constraints. Building on the lessons learned from this representative sample, we give recommendations for future trajectory data collection and sharing.

The resulting contribution is not only a synthesis of recent practices, but also a set of actionable guidelines intended to enhance reproducibility, comparability, and broader usability of trajectory datasets within the research community and beyond.

The remainder of the paper is organized as follows. Section 2 provides a systematic review and detailed analysis of the NGSim dataset. In several other sections, NGSIM is also used as a primary example to support our discussions. The relatively extensive coverage of NGSim in this paper is justified by three reasons. First, NGSim is one of the earliest and most widely recognized large-scale open trajectory datasets, which has inspired numerous subsequent data collection initiatives and continues to serve as a reference point. Second, the dataset's quality has been rigorously scrutinized, with many errors identified and significant correction efforts undertaken in the literature. These efforts not only produced multiple reconstructed versions of NGSim, as discussed later, but also informed more recent initiatives, helping them avoid similar issues. Finally, NGSim has played a catalytic role in advancing empirical understanding and theoretical modeling of driving behavior and traffic dynamics, as evidenced by the large number of studies that were based on it. We believe NGSim is important role and impact in the community will continue to grow as researchers increasingly turn to benchmarking datasets. Section 3 synthesizes the main approaches currently used for trajectory data collection, processing, and sharing, thereby providing a comparative framework across diverse practices. Section 4 distills the key lessons that can be drawn from these experiences and trade-offs between different design choices. Section 5 builds on these insights to propose a set of recommendations and perspectives for future design, processing, and dissemination of trajectory datasets. Finally, Section 6 concludes the paper.

2 NGSim Systematic reviews and analysis

This section provides a systematic review and analysis of research related to NGSim trajectory data. We begin by formulating a structured literature query to retrieve relevant publications, followed by an examination of publication trends.

2.1 Systematic Literature Query on NGSim Trajectory Research

Among the publicly available trajectory datasets, the NGSim dataset [26] stands as a central resource for studying vehicle interactions and motion patterns at a microscopic level. As an open-access dataset, it has been widely adopted in research, leading to significant advancements in traffic flow modeling, autonomous vehicle testing, and safety analysis [27].

Given its accessibility, the literature has extensively analyzed the impact of NGSim, highlighting its scientific contributions [27], lessons learned from data acquisition [28], and guidelines for future data collection and analysis [29].

To systematically assess the impact and application trends of this dataset, we formulated a targeted query to retrieve all research papers explicitly referencing these terms in their title or abstract. The query was performed on Scopus as follows:

```
(( TITLE-ABS-KEY ("NGSim" OR "next generation simulation" OR "next-generation simulation") AND TITLE-ABS-KEY ("trajectory" OR "trajectories")) AND (LIMIT-TO (LANGUAGE, "English")))
```

This query resulted in a total of 627 records as of March 16, 2025, capturing a comprehensive collection of studies in english leveraging NGSim and trajectory data from 2006 to 2025.

2.2 Analysis of NGSim and Trajectory Data Research Trends

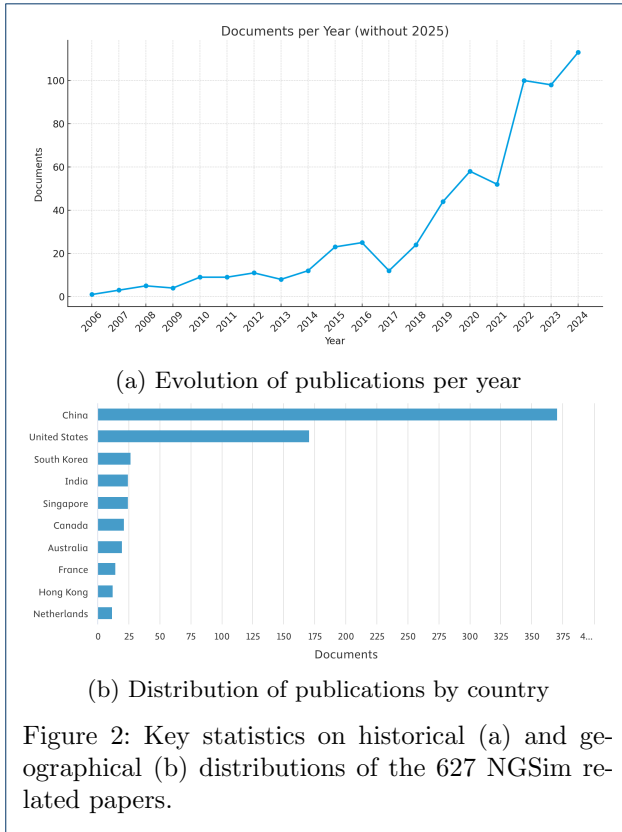
Since its publication in 2006, Figure 2a demonstrates the continued importance of the NGSim dataset in research. Its use has grown steadily, with a sharp rise after 2018 and peaks in 2022-2024, confirming its lasting relevance and its status as a benchmark and reference dataset in the field (The drop in 2025 is likely due to incomplete indexing). Figure 2b shows that, despite being collected in the U.S., the NGSim dataset is widely used worldwide, demonstrating its global impact on traffic and mobility research.

Using this database of 627 research papers retrieved from the query, we conducted an in-depth analysis using VOSviewer [30] to visualize keyword occurrences, clustering, and citation relationships. The analysis was performed using all keywords (author and index) as the unit of analysis, with a minimum keyword occurrence threshold of 15, resulting in 102 keywords meeting the threshold.

Figure 3 presents a network graph of keyword clustering, where closely related topics are grouped together and showing links between these items.

We can distinguish four clusters, each represented by a different color, corresponding to key research themes:

- *Red cluster*: Focuses on traffic control, traffic congestion, and adaptive cruise control, highlighting



studies on microscopic traffic modeling and vehicle interactions.

- *Green cluster*: Centers around lane-changing behavior, risk assessment, behavioral research, and safety studies, reflecting human driver modeling, decision-making analysis, and traffic safety evaluations.
- *Blue cluster*: Encompasses trajectory prediction, recurrent neural networks, and forecasting, emphasizing AI-driven applications for autonomous vehicles and predictive modeling.
- *Yellow/Purple nodes*: Represent cross-disciplinary concepts linking different clusters, including intelligent vehicle systems and advanced driver assistance systems (ADAS), which bridge traffic control, driver behavior analysis, and predictive modeling.

Figure 4 clearly highlights the current research trend towards the integration of artificial intelligence (AI) in trajectory and traffic analysis as well as predictive modeling. More specifically, recent deep learning technologies, such as encoder-decoder architectures, recurrent neural networks (RNNs), and attention mechanisms, are increasingly utilized.

Taken together, this analysis shows both the opportunities and the limitations of relying on a single

benchmark dataset. While NGSIM has enabled reproducibility and comparability across studies worldwide, its age, scope, and known quality issues also underline the need for new datasets designed with broader objectives and improved data collection practices. This motivates the broader synthesis of trajectory datasets presented in the following section.

3 Synthesis of trajectory data collection, processing, and sharing

This section is devoted to presenting the various methods of data collection, as well as the steps of data processing and correction. The results presented here are mainly based on the material provided in the annex. Although those datasets are a subset of the set of all the openly available datasets at the time we write this paper, we gather the contributions of volunteering teams, and consider that the resulting set is sufficient to represent the diversity of the complete set of trajectory data. Here is the list of the datasets presented in the annex.

Four on-board sensors datasets:

- Waymo [31, 32] and Ohio Instrumented Probe Vehicle experiments [33, 34]. These two datasets are based on the perception of uniquely equipped vehicles: the autonomous Waymo vehicle in the first case, and a specially designed vehicle from Ohio University in the second, accompanied by detailed and extensive loop detector data.
- OpenACC [35] and TOD [36] datasets, which are, respectively, collected by the JRC Research Institute and by the Beijing Jiaotong University, containing data on vehicle platoons (ACC vehicles in the former and human-driven vehicles in the latter).

Five external-sensors-based datasets:

- Flanders dataset [37], collected with a dense set of loops providing individual and precise measurements that enable vehicles identification,
- Two extensive camera-based datasets: Jata-yu, from the Indian Institute of Technology, Roorkee, where different road conditions and urban configurations, as well as many different vehicle types, are present; and the I24 Motion dataset, collected by a team led by the Vanderbilt University along 6 km of the two directions of a highly circulated freeway in Tennessee, USA. Although we expect those two trajectory datasets to be released soon, we have access, for the moment, only to some chunks of individual trajectories and to the whole images collection.
- Two camera-based datasets provided with the trajectories: Zen Traffic Data, originally provided by

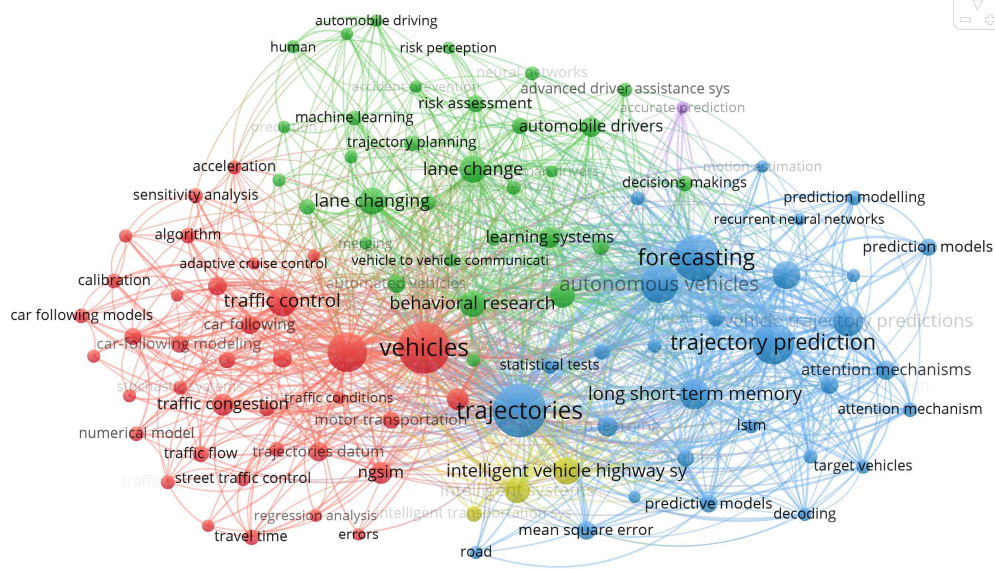


Figure 3: Keyword clustering based on co-occurrence in research papers on NGSIM and trajectory data. Nodes represent keywords, with larger nodes indicating higher occurrence, while edges illustrate their relationships.

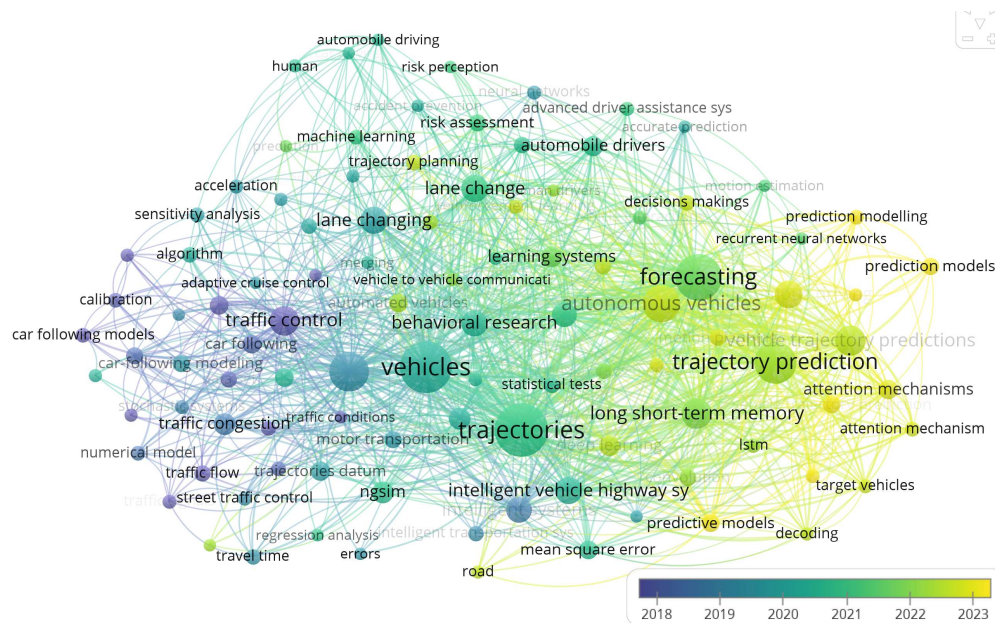


Figure 4: Keyword co-occurrence clustering with a temporal gradient (2006-2025). Colors from blue to yellow illustrate the evolution of research interests over time.

the traffic management center of the Hanshin expressways, in Japan, near Osaka, and presented here by a team from University of Queensland, and the TGSIM dataset, collected by a team of various US universities, led by the late Hani Mahamassani.

To gather information from collaborators regarding each dataset a common template was proposed. For each dataset, we requested details on the data collection perimeter (vehicle types, location, traffic conditions, spatial and temporal coverage), experimental setup (sensors, test plan, protocol), trajectory char-

acteristics (format, frequency, quality), and data processing (filtering, missing data handling, outlier detection). We also gathered information on metadata (e.g., weather, incidents), network description, data accessibility (open, on request, or private), availability of preprocessed and curvilinear data, and any deviations from the initial plan. Finally, contributors were asked to report challenges in the employed methodology and suggest possible improvements or future work. Once gathered, the information was reviewed and refined to create a consistent and comprehensive section. All the detailed information provided by collaborators for each dataset is available in the appendix (Sec 7).

The next table 1, presents those datasets in a synthetic manner, along with details on where they can be accessed or downloaded. One can see that the extension of the datasets in terms of the hours / number of trajectories collected, type of networks, traffic conditions, etc... is highly variable.

Conducting a trajectory dataset collecting and sharing project requires organization and a clear vision of the objectives and available resources. Four steps can be identified: the preparation of the data collection and sharing, the data collection, the data processing, and the data sharing.

The preparation should not be taken as an easy part. Unfortunately, this part of the work, although important in how it impacts the final result, is not well documented, as far as we know. We did not access, for example, any Data Management Plan (DMP) for any trajectory datasets listed in the annex. The data collection phase can thus begin. Collecting data also involves collecting metadata, where a precise description of the data collection process and the conditions under which it takes place are given. The processing step can thus begin, comprising many different sub-tasks. Finally, the fourth step, sharing the data, again needs resources for storage space and human resources, as documenting, maintaining, and answering external dataset users requires some time.

The literature review about the nine datasets and other data sets collected in the past (like NGSim) shows us that most papers mainly focus on the third step: processing the initial sensor data to produce a shared dataset. Figure 5 presents the different tasks that must be conducted and lists the initial, intermediate, and final datasets that are produced and can be openly shared.

3.1 Initial signal measurement

As stated in the introduction, in this paper, we define trajectory data as a sequence of positions and/or speeds of individual vehicles following one another. Such trajectory data are particularly valuable for microscopic traffic observation and modeling, and even

safety studies. These intended applications impose a set of requirements:

- The information related to each individual vehicle should be complemented with the information of the surrounding vehicles, enabling the computation of critical input variables for microscopic traffic flow modeling (e.g., spacing, relative speed, etc.) and safety surrogate measures (e.g., time to collision);
- The position measurements must be expressed in a universal reference frame;
- The observation frequency of positions should be at least 1 Hz, and preferably higher;
- The measurement errors in positions should not exceed one meter, and preferably lower.

Consequently, all data collection processes are designed to ensure these quality requirements. Trajectories can be collected either:

- **Externally**, by observing vehicles as they travel on the road.
- **Internally**, using sensors installed inside a vehicle (probe vehicle: PV) or a fleet of vehicles.

The main difference between the two data collection approaches lies in driver awareness. In external observation, drivers are typically unaware that they are being monitored, whereas in internal collection using on-board sensors, drivers give explicit consent at some point for their cars to be equipped. Depending on the country and methodology, this consent can be provided either specifically within the context of the data collection experiment or more broadly when agreeing to share personal data. This type of data can be collected via devices mounted in the vehicle or geo-localization of a passenger's smartphone.

External data collection techniques are based on video cameras or dense arrays of road-embedded sensors, usually magnetic loop detectors. In experimental settings, probe vehicles can be equipped not only with position and speed sensors but also with LiDAR, radar, and cameras to capture distances and relative speeds of surrounding vehicles. These heterogeneous data streams must then be synchronized within a common reference frame.

Regarding externally collected data, the historical example of the trajectory data collection led by Treiterer and Meyers [9] is interesting. The original data collection occurred on the southbound lanes of the I-71 freeway in Columbus, Ohio, USA, in the summer of 1967. Their data collection used high-speed aerial photography and manual data reduction. The data set includes 210 vehicles, spans 4.8 kilometers and lasts 4 minutes. The data were collected from a helicopter moving with the traffic, yielding a moving surveillance region that is up to one kilometer long. In spite of the

Table 1: Overview of the studied datasets and their specificities. Abbreviations are used for conciseness: (*) Location of main sensors : O = On-board, E = External to the vehicle. (**) Research objectives: MDB = Microscopic driving behavior studies, TFT = Traffic flow theory, SR = Safety and risk analysis, Ctrl = Control and ITS, MLF = Machine learning and forecasting, NLO = Network-level operations, EEI = Environmental and energy impacts.

Dataset Name	Open Access	Described in annex of this paper	Location of main sensors *	Year	Data Collection Method	Location	Total Duration	Network Type	Format	Research objectives **
OpenACC [35]	JRC Repository	✓	O	2021	GNSS + on-board	Europe	Varies	Urban, Hwy	CSV	MDB, TFT, SR, EEI
TOD [38]	Under request	✓	O	2014–	GNSS	China	Varies	Urban, Hwy	CSV	MDB, TFT, SR
Waymo [31]	Waymo Dataset	✓	O	2020	GNSS + Li-dar + Cam	USA (cities)	(6 574h)	Urban	TfRecord	MDB, TFT, Ctrl
Ohio IPV [33, 39]	Ohio	✓	O	–	GNSS on-board + loops	USA (Ohio)	90 km/trip	Hwy	Matlab	MDB, TFT, SR, Ctrl
Flanders [37]	To be shared	✓	E	2018–21	Loops GNSS	Belgium	Many days	Hwy	–	MDB, TFT, SR, Ctrl
Jatayu [40]	GitHub	✓	E	–	UAV camera	India (cities)	(6 33h)	Urban, Semi-urban	YOLO +video	MDB, TFT, Ctrl, MLF
I24 Motion [41]	I24 Data	✓	E	–	Fixed cameras	USA (TN)	234h	Hwy	Video	MDB, TFT, SR, Ctrl, MLF
Zen Traffic Data [42, 43]	Zen Dataset	✓	E	–	Fixed cameras	Japan	Various	Hwy	CSV	MDB, TFT, SR, Ctrl, MLF
TGSim [44]	TGSim	✓	E	2024	UAV camera	USA	Unspecified	Hwy	CSV	MDB, TFT, SR, Ctrl, MLF
pNeuma [45]	pNeuma Dataset		E	2023	UAV cameras	Greece (CBD)	12.5h	Urban	CSV	MDB, TFT, SR, Ctrl, MLF
LevellXdata [46]	LevellXdata		E	2018	UAV camera	Germany	16.5 hours	Hwy	CSV	MDB, TFT, SR, Ctrl

herculean collection effort, the trajectory dataset was never released and has since been lost. Fortunately, the plots survive and Coifman et al. (2018) [47] re-extracted the vehicle trajectory data from the time-space diagrams published in Treiterer (1969) [48]. The trajectory data set is openly available for download at [49].

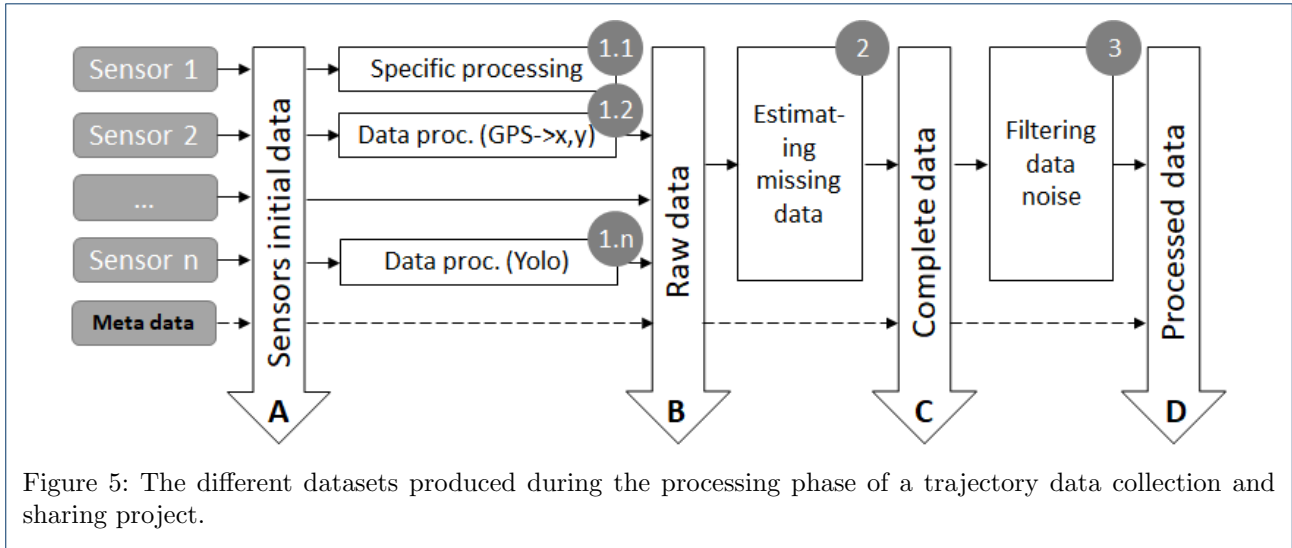
3.2 From initial signal to raw data

From the initial signal up to the raw data, the data processing consists of several steps depending on the type of initial signal collected. Each step in the process introduces specific errors which, when combined, lead to inaccuracies in the final raw dataset.

Given its celebrated status in the community, NGSIM dataset is used as an example to demonstrate errors that typically happen at this stage. Collected using a network of cameras, the dataset required stitching together images from multiple sources—an operation that introduced significant errors in position measurements and other derived measurements. For instance, studies have shown that the NGSIM data contain numerous severe tracking errors. In particular, the original dataset displays piecewise-constant speeds, which compromise any application that relies on acceleration. If left uncorrected, such errors can cause significant issues for microscopic modeling and simulation. Consequently, substantial efforts have been devoted in

the literature to detecting and correcting these errors, as elaborated below.

For NGSIM dataset, vehicle trajectories were obtained by tracking vehicles across multiple, synchronized cameras. However, object-recognition errors and algorithmic shortcomings often led to miss tracking, producing substantial errors in estimated positions, speeds, and accelerations. These issues were identified by [50] and examined in detail by [51], who also proposed a comprehensive methodology to assess trajectory accuracy based on jerk analysis, spectral analysis, and internal/platoon consistency. Building on these insights, [52] developed a traffic-informed, multi-step procedure to restore the physical plausibility and platoon integrity of vehicle trajectories. To overcome the limitations of conventional filtering/smoothing - namely, the trade-off between removing large errors and avoiding unrealistic smoothing of speed and acceleration profiles - the method performs local reconstruction by replacing severely biased trajectory fragments. The replacements yield synthetic positions that are (i) physically compatible i.e., yielding plausible speeds and accelerations, (ii) consistent with the distance traveled within the reconstruction window i.e., internally consistent, and (iii) preserving inter-vehicle spacing i.e., platoon consistent (the concept of "platoon consistency" was first introduced in XX and then clarified in Punzo et al., 2011).



The reconstruction methodology was validated against traffic macroscopic patterns emerging from the trajectory data. The procedure was applied to the entire NGSIM I-80-1 dataset, and the corrected trajectories were subsequently released by the FHWA alongside the original data.

Subsequent research revealed that many of the errors in the original NGSIM dataset could not be corrected through simple post-processing. Consequently, Coifman and colleagues manually re-extracted all passing vehicles from the original video of a single camera in the I-80 NGSIM subset, using the tools later published in [53]. Special attention was given to obtaining reliable acceleration data. They also revealed systematic positioning errors near the start and end of the camera's field of view. These boundary errors motivated a more extensive investigation by Coifman and Li (2024) [49], who conducted a more in-depth study across all seven cameras used in the NGSIM I-80 data collection. Focusing again on the first time period, all vehicles were manually tracked in each camera. This study developed tools that leveraged trajectories across successive camera views to identify overlaps or gaps between adjacent cameras, as well as potential scaling errors. The analysis demonstrated that the original NGSIM data extraction had erroneously introduced small overlaps between all camera pairs at this site, which led to systematic speed errors as vehicles crossed camera boundaries. To establish ground truth, they used a newly identified aerial photo of the site, captured within one year of the original data collection. The re-extracted dataset not only corrected these issues but also tracked vehicles over a 14% longer distance and added 23% more vehicles compared to the original NGSIM dataset. Both manually re-extracted

trajectory datasets are openly available for download [49].

3.3 From raw data to postprocessed data

Raw data are inherently noisy, with errors arising both from sensor inaccuracies and from the various stages of initial signal processing. In addition, missing data are common. Thus, there is a rich body of literature focusing on denoising and enhancing the quality of vehicle trajectory datasets, many of them included in this paper. Describing those methods in detail is beyond the scope of this paper. Instead, a brief summary is provided below. In [50], one of the first papers identifying and proposing postprocessing methods to improve NGSIM data quality, the speeds and acceleration of the raw data. Later, Zheng et al. developed wavelet-energy-based methods to denoise NGSIM trajectories and detect interesting events (singular points) both for car-following and lane-changing modeling, and this approach was later refined through the use of wavelet modulus maxima.

Hu et al. (2022) comprehensively processed and evaluated the Waymo Open Dataset, with a particular focus on car-following paired trajectories. Data quality was assessed in terms of internal consistency, jerk values, and trajectory completeness. The results indicate that, although all extracted trajectories are incomplete, they generally exhibit higher quality than those in the NGSIM dataset.

Mahajan et al. (2023) applied a combination of smoothing filters and Extreme Gradient Boosting (XGBoost) with adaptive regularization to process speed and acceleration profiles in the pNEUMA dataset, aiming to remove noise and anomalies. Their approach effectively detected anomalies in the form

of unrealistic transient peaks. Adaptive regularization adjusts dynamically based on the maximum acceleration value, thereby simplifying anomaly detection by reducing the number of tunable parameters.

Li et al. (2024) proposed a four-step car-following (CF) extraction algorithm—vehicle grouping, elimination of false overtaking, vehicle sorting, and CF pair matching—applied to the Zen Traffic Data, yielding 246 CF pairs and three derived datasets. Their quality analysis showed inconsistencies across data fields, with the geography-based dataset least affected by noise. Later, Li et al. (2025) introduced an adaptive multi-layer trajectory reconstruction method combining window alignment, adaptive window sizing, and an XGBoost-based ground truth estimation module. A jerk-based reconstruction approach, outperforming acceleration-based methods, was used to remove outliers in the Zen Traffic Data.

4 Lessons learned for future trajectory data collection processing and sharing

As shown in the elements gathered from presented datasets (see Table 1), certain trends emerge, along with essential information that appears consistently across all datasets despite diverse methodological approaches.

4.1 Cross-Dataset Insights

Beyond the simple classification of datasets in the appendix 7, several cross-cutting observations emerge. First, the datasets illustrate a clear trade-off between controlled accuracy and naturalistic representativeness: platoon experiments (OpenACC, TOD) offer high-quality, homogeneous measurements but under constrained conditions, while large-scale naturalistic collections (Waymo, I24 Motion, Zen, Jatayu) provide breadth and realism at the cost of noise, occlusions, and shorter trajectory spans. Loop-based and hybrid approaches (Flanders, Ohio IPV) occupy an intermediate position, ensuring long-term coverage and infrastructure integration but at lower spatial resolution. Second, data processing methods vary considerably, from cubic spline interpolation (OpenACC) and wavelet-based denoising (Waymo) to manual re-extraction (Ohio, NGSim), each striking a balance between smoothing, noise removal, and preservation of realistic variability. Third, metadata availability remains uneven: while some datasets provide detailed road geometry, traffic conditions, and even environmental descriptors (Waymo, OpenACC), others share raw trajectories with limited contextual information (TOD, Jatayu). Finally, accessibility and interoperability are heterogeneous, ranging from fully open repositories in standard CSV or YOLO formats

(OpenACC, Jatayu) to proprietary storage formats (tfrecord for Waymo, Matlab for Ohio IPV), or access upon request (TOD, Flanders). This diversity highlights both the richness of current resources and the need for harmonization in trajectory data collection, processing, and sharing, particularly regarding open metadata standards and reproducible processing pipelines.

Here are some key observations:

- **Trends in Trajectory Data Collection:** Trajectory data collection has evolved from traditional methods such as probe vehicles, GNSS-based tracking, and loop detectors (e.g., OpenACC, Ohio Instrumented Probe Vehicle) to multi-sensor systems integrating LIDAR, HD cameras, and UAVs (e.g., Waymo Open Dataset, I24 Motion, Jatayu).
- **Data Processing:** Datasets increasingly include preprocessing pipelines to enhance data quality. Advanced denoising techniques, such as wavelet-based filtering, help remove outliers and improve trajectory consistency. Interpolation methods are commonly applied to reconstruct missing trajectories caused by occlusions or sensor limitations. Spline-based interpolation is frequently used to filter GNSS noise, ensuring smoother trajectory estimation. Additionally, post-processing tools enhance vehicle tracking by refining cross-camera association and addressing inconsistencies in multi-sensor datasets. Not all of datasets provide code repositories used to process the data.
- **Meta data in open formats:** Standardized open formats such as CSV are more used than proprietary ones (e.g., Matlab). Their use is encouraged by the improvement of data interoperability and accessibility.
- **Metadata and Context shared information** Rich metadata significantly increases the value of trajectory datasets by providing road geometry, vehicle attributes, and environmental descriptors that enable more refined analyses. Some datasets include extensive contextual layers (e.g., Waymo, I24 Motion), while others mainly provide raw trajectories with limited additional information. Differences in metadata availability may come from the heterogeneity of data collection protocols and dissemination policies.
- **Extensive and Homogeneous data columns:** The structure and content of data columns differ substantially across the various studied datasets. To define how we may commonly present our dataset and what should be the data attributes, one could refer to the unified datasets proposed in the literature recently. With the aim of facilitating

the benchmarking of car-following models, the authors of [54] built a dataset with for all 15 seconds or more events of car-following in NGSim, HighD, SPMD (Safety Pilot Model Deployment, an experiment designed by NHTSA of US DoT to test V2V communication), Waymo and Lyft datasets. The common data format was, for all observation instants of each car-following events: spacing, subject vehicle speed, relative speed, leading vehicle speed. There was no information about the vehicles types and lengths, the road characteristics, the weather conditions,... A second and more extensive example is [3], and its companion dataset [55]. There, a synthesis database is proposed for 13 experiments collecting trajectory for autonomous vehicles. The dataset contains 12 attributes, in addition to trajectory ID and time index: the ID of the lead vehicle, its type (AV or not), its position in Frenet coordinates, its speed and acceleration, the same information for the subject automated vehicles, the bump to bump distance between the two vehicles, the distance between their center of gravity, and their speed difference. This data format is adequate for car-following studies, but will not be suitable for any lane changing studies of multi-anticipation studies. Also, the fact of using Frenet coordinates (if we understand correctly, the longitudinal coordinate in the curvilinear coordinates system, is directly linked with the choice of concentrating on car-following studies.

- Common Practices in Data Collection:** Datasets increasingly follow common practices to ensure data quality and usability. High-frequency sampling, typically at rates of 10 Hz or higher, allows better trajectory tracking and movement analysis. Metadata inclusion, such as road attributes and environmental conditions, provides essential context for interpreting vehicle trajectories. Finally, datasets now cover a different range of driving conditions, including urban and highway environments, as well as congestion and free-flow scenarios, ensuring their applicability across diverse traffic studies. These details are sometimes provided on the dataset's website, in associated research papers, or in separate metadata files. To improve accessibility and reproducibility, a standardized metadata file including all these essential elements should be systematically provided alongside collected data.
- Global Insights on Speed and Acceleration Behavior:** The acquisition duration varies significantly across datasets, reflecting diverse research objectives. While some, like Waymo Open Dataset (574 hours) and I24 MOTION (234 hours), offer

extensive recordings for large-scale analysis, others, such as Traffic Open Data and Jatayu (33 hours). Additionally, datasets like Ohio Instrumented Probe Vehicle quantify acquisition by distance rather than time, highlighting differences in data collection approaches. However, a notable gap across these datasets is the absence of detailed statistics on acquisition time distribution and speed variations within the collected data. Such information is crucial for assessing dataset representativeness, identifying traffic flow patterns, and understanding vehicle dynamics.

- Accessibility:** Although these datasets are open-access, they are distributed through various platforms and adopt different access approaches. Some datasets are freely available for direct download, while others require users to request access through specific hosting platforms. Additionally, certain datasets are stored on dedicated repositories, whereas others are accessible via research institutions' platforms or cloud-based hosting services. In some cases, access is granted immediately, while in others, users must complete a formal request process.

5 Where shall we go?

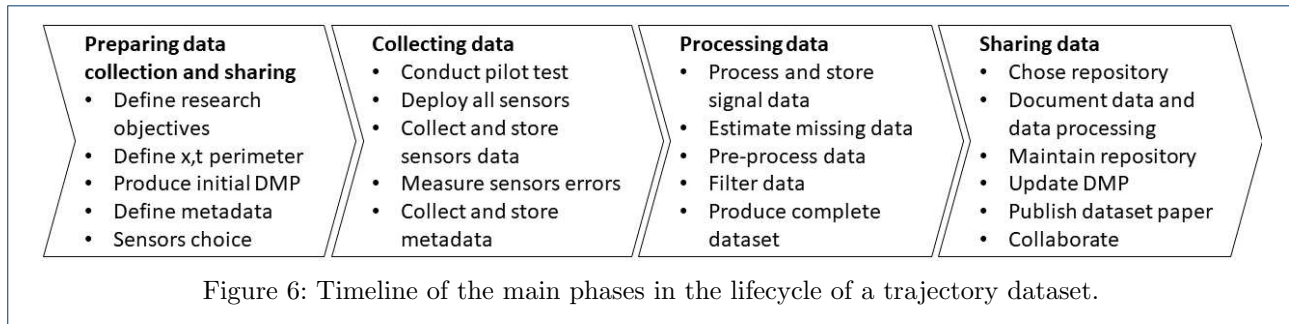
Building on insights from existing datasets, the lessons learned in the previous section, recent trends highlighted in the literature, and the authors' collective experience in trajectory data collection and analysis, this section presents practical recommendations for the future development of trajectory data practices. The recommendations identify key priorities to ensure datasets are more comprehensive, standardized, and accessible; responsive to emerging mobility technologies; supportive of cross-study benchmarking; and conducive to long-term comparability and usability in traffic behavior research. The key recommendations are categorized into two groups: general recommendations on data collection and sharing; and specific recommendations to fulfill new data needs with respects to the enlarging of trajectory datasets usages. They are presented below, respectively.

5.1 General recommendations on data collection and sharing

We have seen earlier, in part 3 that one can distinguish four steps in any project for trajectory datasets collecting, and sharing. These steps are illustrated in Figure 6.

5.1.1 Preparing data collection and sharing

- Research gap, perimeter, and data collection settings definitions:** the design of new trajectory data collection campaigns should be informed by the systematic gap analysis of existing



datasets provided by the previous step. Considering the resource limitations of our research field, we encourage readers to conduct a careful literature review before defining the data collection project. Updating it throughout the preparation phases of the project permits the adaptation of the experimental settings if needed. Rather than reproducing experimental configurations that are already extensively covered, efforts should target scenarios, environments, and traffic conditions that remain underrepresented in the literature. Such configurations may include specific road geometries, rare or extreme traffic states, adverse weather conditions, nighttime operations, or regions with heterogeneous driving behaviors. It is also important to define a data collection perimeter, in term of space and duration and period of data collection suitable to the available resources. Being too ambitious is a common drawback.

- **Data management plan:** We recommend to adopt a robust Data Management Plan (DMP) tailored to trajectory dataset collection and sharing, covering ethics, anonymization of personally identifiable information (e.g., license plates, faces), licensing conditions, storage horizon, and rigorous ground truth documentation. European commission provided recently a template to fill that can be taken as an example to follow. The DMP must be defined before the data collection itself, and updated all along the project (see [ref](#)). The DMP should ensure that data and metadata comply with the FAIR principles (Findable, Accessible, Interoperable, Reusable) as recommended by the European Commission, thereby facilitating long-term accessibility, interoperability, and reproducibility for traffic flow modeling, autonomous vehicle development, and mobility research [56, 57].

5.1.2 Collecting data

- **Pilot testing in experimental design:** it is recommended to conduct preliminary test campaigns before full-scale data collection to detect

and resolve synchronization issues, sensor calibration errors, or coverage gaps. Use these tests to optimize sensor placement, validate acquisition parameters, and confirm that the planned scenarios address gaps in existing datasets.

- **Metadata collection and storage, in the perspective of sharing:** in the context of traffic flow modeling and traffic evolution analysis, it is essential to collect detailed road environment data, such as curvature and slope conditions, and meteorological details (sunlight conditions...). It is also necessary to provide information about the road quality (marking visibility, asphalt state, ...). This information provides a more holistic understanding of how external factors influence driving behavior. Detailed geographic information is also needed to permit the precise positioning of the vehicles with respect to the road geometry. Indeed longitudinal and lateral GPS coordinates are not sufficient to identify rapidly the lane positioning. All those information constitute the metadata and must be carefully stored. This will permit later on to share them with the trajectory datasets users.
- **Sensor Accuracy, Redundancy, and Human Factors Capture:** prioritizing state-of-the-art sensing technologies such as high precision GNSS, high resolution lidar and advanced inertial measurement units (IMUs) is essential to reduce measurement noise and improve trajectory fidelity, particularly in complex environments. To further enhance robustness, datasets should combine multiple complementary sensors including GNSS, lidar, cameras and IMUs. This redundancy supports cross validation of measurements, facilitates the detection of anomalies and increases overall data reliability.

Systematic quantification of measurement errors is also necessary, with explicit reporting of error bounds, sensor uncertainty specifications and validation results against ground truth to ensure transparency and reproducibility.

Future trajectory data collection can evolve toward the comprehensive integration of vehicle tra-

jectory data such as speed and position with driver status and psychophysiological information including electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), electrodermal activity (EDA) and eye tracking. Incorporating these measurements provides valuable insights into cognitive load, emotional state and fatigue levels, enabling a deeper understanding of human factors and supporting the development of interpretable driver behavior models.

5.1.3 Processing data

- **Documenting any step of data processing:** with the experience of using NGSim data, we can guaranty that the NGSim team’s silence about the coding hypothesis posed during the filtering was the cause of hours of extra work for NGSim data users.
- **Adapt the filtering method to the specific sensors noises:** the method to process the data is strongly dependent on the initial quality of the sensors data [21]. The filtering method must be suited for the precise dataset one wants to filter and may not be adapted to all data subsets. When the observed vehicles follow a lane discipline, it is important that one of the processing permits to access to trajectories placed in the road reference and not on (x, y) or a GNSS plane. For this, a precise localization of the landmarks of the road in link with the reference system of vehicles trajectories is needed.
- **Fostering collaboration about the data processing:** to help the community to benchmark the missing data correction and the many filtering methods, it is needed that the teams authoring the datasets
 - share the code they used for the various processing phases. This is illustrated in figure 5, with the numbers on the top left of the processes.
 - share also various datasets, namely the sensor initial dataset (A), the intermediate datasets (B and C), and not only the processed data resulting of the complete pipeline, as it is commonly the case.

With this complete share, international comparison could be undertaken, contribution to a truly reproducible research.

5.1.4 Sharing data

- **Data preprocessing documentation** the code developed and used to extract, fuse initial sensors data and define the raw trajectories must be shared with the community. We encourage our

readers to follow the recommendations for level 5 of repository quality indicated in [58]. Any preprocessing on existing dataset should be mentioned and added to the reference repository of initial dataset.

When trajectory data is derived from video analysis, it is essential to clearly identify portions of the dataset that have undergone manual annotation or correction. Such information is crucial for assessing the reliability of the resulting data and for enabling its reuse in other research contexts. Whenever possible, the associated postprocessing code, annotation protocols, and source material should be openly shared with appropriate documentation. In particular, manually labeled data can serve as valuable ground truth for benchmarking and validating algorithms in subsequent studies.

- **Standardization of Trajectory Data Formats:** current trajectory datasets come in various formats (e.g., CSV, JSON, XML), which complicates integration across sources and hinders collaboration, benchmarking, and large-scale model validation. To address this, we recommend the adoption of a standardized structure combining (i) a trajectory file stored in Parquet format, ensuring efficient, scalable, and non-proprietary storage of large spatio-temporal datasets, and (ii) a companion metadata file in JSON or YAML, capturing essential contextual information such as acquisition setup, sampling frequency, coordinate system, road geometry, weather conditions, and traffic states.

In cases where data cannot be natively provided in the agreed format, dataset providers should supply conversion scripts or APIs to transform the original files into the standardized structure

- **Versioning and traceability of datasets:** trajectory datasets should adopt explicit version control mechanisms, comparable to those used in software development. Each release must be accompanied by a unique version identifier and a detailed changelog documenting updates, corrections, and preprocessing operations .
- **Dataset sharing and collaboration:** beyond common formats and annotation standards, collaboration can be strengthened by allowing contributors to enrich existing datasets through shared pipelines, for instance by providing conversion scripts, preprocessing routines, or additional experiments. Such practices foster interoperability, continuous improvement, and collective validation of trajectory data.

Data sharing recommendations: current practices vary from fully open datasets to request-based or restricted access, limiting reproducibility. We recommend FAIR-compliant repositories (e.g., Zenodo, Dataverse) with persistent identifiers, versioning, and standardized metadata. At present, trajectory datasets are often scattered across heterogeneous platforms (personal websites, project pages, institutional repositories), which hampers discoverability and reuse. To address this, datasets should be systematically deposited in those recognized repositories that provide persistent identifiers, metadata integration, and open access. Beyond storage, repositories should also support community-driven extensions, allowing researchers (after validation) to contribute pre-processing scripts, conversion pipelines, or even new experimental data, thus fostering continuous improvement. To maximize usability and transparency, the metadata should be stored in accessible, non-proprietary formats, and be subject to regular updates to reflect corrections or evolutions of the underlying database.

- **Publish a dataset paper:** this type of paper is, to the authors' opinion, the best way to evidence and document the work done during the project. Recent examples for trajectory data sharing are the following:
 - A team from the Institute for Transport Planning and Systems, ETH Zurich, Switzerland, provides in [59] a complete comparison of different methods for trajectory reconstruction taken from a camera footage.
 - A team from the Institute of Transport and Economics, in TU Dresden, Germany, describes in [60] the dataset collected using six drones over a 900 m section of the A50 urban freeway in Milan, Italy.

Those two papers are examples in the precision and the clarity of the information provided. This eliminates for the future users of those datasets tedious questions that arise too often when using other datasets, like, for example NGSim.

5.2 Recommendations for new types of trajectory datasets

Despite the growing number of published trajectory datasets, the research community still lacks other types of datasets that would enable addressing specific applications, exploiting the latest sensing technologies, and supporting interdisciplinary studies. There is a pressing need for datasets that adapt to technological evolution while remaining relevant for a broad range of research and practical challenges.

- **Real-Time Data Sharing and Accessibility for Traffic Management and Research:** Most trajectory data are collected and analyzed retrospectively, limiting their usefulness for real-time traffic management, incident detection, and dynamic control strategies. There is a need for more efficient ways to share trajectory data in real time for both operational and research purposes.

The development of Open Access platforms in collaboration with cities around the globe would exponentially accelerate the research in traffic engineering and relevant topics and would facilitate the assessment of real-world applicability of any novel solutions. In coordination with the first point, practitioners and scientists would have the opportunity to assess their methods on a real-world environment overcoming strong real-world obstacles that may relate to sensor accuracy, data gaps, cloud computing delays, synchronization issues, communication lags and computational complexity.

Pilot projects, international collaboration and the establishment between university and authorities (cities) is the key for the initiation of such actions.

- **The need for creating a benchmarking data repository:** In domains such as image processing, computer vision, signal processing, and computer science, widely adopted benchmark datasets have played a pivotal role in advancing research — for example, ImageNet in image processing and vision [61], and the UCI Machine Learning Repository [62] in broader machine learning applications. In contrast, although the number of trajectory datasets has been rapidly increasing, the field still lacks a widely recognized benchmark repository for systematic evaluation and comparison of algorithms and models. There is a pressing need to establish a standardized benchmark dataset specifically designed for vehicle trajectory-based research, with particular relevance to microscopic traffic modeling (e.g., car-following and lane-changing behavior) and autonomous vehicle testing. The absence of such a benchmark dataset limits transparency, impedes reproducibility, and ultimately constrains the cumulative progress of traffic research.

- **The need for creating safety-oriented datasets:** Road safety has long been a primary concern in both traffic research and practice. Researchers have sought to develop microscopic models capable of accurately predicting and reproducing observed collisions; however, progress has been limited. One key challenge is the scarcity of trajectory datasets that include a sufficiently rich sample of

vehicle collisions, as collisions are inherently rare events.

The need for safety-oriented datasets is becoming increasingly urgent with the rapid advancement of autonomous driving, which demands rigorous testing across a wide range of hazardous and safety-critical scenarios. Such testing is most effectively conducted in virtual environments that integrate both human drivers and automated vehicles into the loop. Researchers are increasingly recognizing that the most effective strategy for evaluating AV safety is to simulate human errors and safety-critical situations at scale before deployment on public roads [63].

Importantly, a safety-oriented dataset does not necessarily require a large number of actual collisions. Instead, it should encompass a diverse set of safety-critical events and near misses, which can be systematically characterized using surrogate safety measures such as Time to Collision (TTC), Deceleration Rate to Avoid Collision (DRAC), and related indicators [64].

- **The need for creating micromobility-oriented datasets:**

Micromobility offers a promising means to mitigate road transport externalities such as congestion, emissions, and energy use. Since their commercial launch in 2016, e-scooters have expanded into a USD 33.18 billion global market by 2024, growing at 10% annually and adopted in over 600 cities worldwide [65].

Yet, widespread adoption faces persistent barriers. Safety concerns, conflicts with vehicles and pedestrians, and infrastructure limitations remain critical challenges. In the United States, e-scooter injuries rose from 8,566 in 2017 to 56,847 in 2022 [66]. In Germany, accidents increased 14.1% between 2022 and 2023, with fatalities doubling to 22 [67]. Paris banned rental e-scooters in 2023 after recording 459 accidents the previous year [67]. Beyond safety, limited understanding of interactions between micromobility and other modes hinders integration. Developing multimodal modeling and simulation platforms capable of capturing interactions among conventional, connected, and autonomous vehicles and micromobility users—is essential.

Besides, Bicycles, including electric bikes, are the most established micromobility mode, and trajectory datasets for cycling trips are beginning to emerge. Some recent efforts [68, 69] highlight this trend, but more comprehensive data are still needed to support safety analysis, multimodal integration, and sustainable transport planning.

Central to address these issues is the development of trajectory datasets with substantial representation of micromobility trips to support safe and efficient integration into urban transport systems.

- **The need for creating sustainability-oriented datasets:**

As the 2nd-largest consumer of energy and the largest consumer of petroleum products, the transportation sector is a major source of urban air pollution worldwide. Within the sector, road transport is the largest source of greenhouse gas emissions, responsible for nearly 75% of transport-related CO₂ emissions [70].

Although numerous trajectory datasets exist, to the best of our knowledge there is no large-scale dataset that integrates high-resolution vehicle trajectories with corresponding energy consumption and emissions. An exception is the Arizona Ring Experimental Dataset (ARED) [71], which includes 160 vehicle trajectories from eight platoons under both stable and oscillatory traffic conditions. However, these data were collected in controlled experiments rather than on open roads.

There is thus a critical need for large-scale datasets that couple detailed vehicle movement with corresponding energy consumption and emissions. Such datasets are essential for developing accurate microscopic traffic models capable of capturing traffic dynamics and their environmental consequences. Without them, the design of effective traffic operation and control strategies (e.g., eco-driving [72]) to promote sustainability is hindered, as is the assessment of the energy and environmental impacts of emerging technologies such as connectivity and automation.

Overall, current trajectory datasets still fall short of supporting several critical research and practical needs which have been identified earlier. Future efforts should prioritize the development of open, standardized, and specialized datasets. We provided guidelines to these aims.

6 Conclusion

This paper has reviewed a representative set of trajectory datasets, analyzing their collection, processing, and sharing practices through a unified descriptive template. By comparing diverse methodologies, it has highlighted both the strengths and the recurring shortcomings of current datasets, particularly in terms of experimental scope and objectives, tools and sensors employed, and the design of data collection plans, as well as issues related to standardization and reproducibility. Drawing on these observations and the collective experience of dataset authors, we first derived key lessons and then outlined recommendations

for future initiatives, with an emphasis on improved metadata management, sensor integration, error quantification, and standardized sharing formats. We also highlighted several important data needs. While the datasets examined here provide valuable resources for traffic flow modeling, intelligent transportation systems, and autonomous vehicle research, significant opportunities remain to expand coverage, enhance quality, and ensure long-term usability. Continued collaborative efforts, guided by shared specifications and reproducibility principles, will be essential to building richer, more interoperable datasets that meet the evolving needs of the transportation research community.

7 Annexes

7.1 Waymo Open Dataset (WOD)

7.1.1 *What was the research question targeted by the team? What can be considered as a potential use?*

The Waymo Open Dataset (WOD) consists of two parts: Waymo Open Perception Dataset (WOPD)[31] and Waymo Open Motion Dataset (WOMD)[32]. The WOPD mainly focuses on the task of 3D semantic segmentation and object tracking, while the WOMD is primarily aimed at real-time trajectory prediction. The WOD can support a range of driving behaviors research under various scenarios, such as car following [73], lane changing [74], and unprotected left turn [75]. Particularly, the WOD is most valuable in comparing the behavioral differences between automated vehicles and human-driven vehicles.

7.1.2 *Data collection perimeter definition*

The WOD was collected by a fleet of automated vehicles with multiple onboard cameras and lidar scanners. For the WOPD, the sensor combination is composed of 1 mid-range lidar, 4 short-range lidars and 5 cameras (front and sides). The WOMD also has 1 mid-range lidar, 4 short-range lidars, but three more camera sensors were added to achieve 360-degree coverage. The geographic locations include 6 cities in the U.S.: San Francisco, Phoenix, Mountain View, Los Angeles, Detroit, and Seattle. Additionally, the 3D road map data was also provided, including lane centers, lane boundaries, road boundaries, crosswalks, speed bumps, stop signs, and entrances to driveways. In terms of data volume, the WOPD includes 11.3 hours of sensor data and the corresponding ground truth labels in 2030 scenarios, while the WOMD includes 574 hours of trajectory data under 103,354 interactive scenarios (merges, intersections etc.). In each scenario, the 3D bounding box of road users (vehicles, pedestrians, cyclists) were classified and continuously tracked by the moving subject automated vehicle for 20 seconds, where the detection range is around 80 meters. In general, there is only one automated vehicle in each scenario, but the number of objects detected varies.

7.1.3 *Experimental design*

The initial data collection plan was structured around recording real-world driving interactions of automated vehicles in diverse environments. The subject vehicles, equipped with a combination of lidars and cameras, continuously captured data on surrounding road users. In each scenario, there was typically one automated vehicle, with varying numbers of observed surrounding vehicles. The test environments included urban intersections, highways, and merging zones. The

vehicles were outfitted with GPS, lidar, and video sensors, enabling comprehensive scene reconstruction. The dataset also provides curvilinear data, allowing a better understanding of driving trajectories in relation to road geometry.

7.1.4 *Metadata*

The WOD also covers a wide variety of traffic situations, i.e., freeway and urban roads under congested/uncongested traffic flow and different weather conditions.

7.1.5 *Data*

The original dataset is stored in the format of “.tfrecord”, which is further transformed to a more user-friendly tabular structure with 25 attributes. These attributes are related to frame context information (IDs of each segment and each frame, time of day, location, weather, number vehicles detected by lidar), object characteristics (object type, object ID, length, width, height) and object trajectory information (time stamp, 3D positions of object center, speed, acceleration etc.). Specifically, the trajectories are collected in a frequency of 10Hz, and the trajectories can be curvilinear (3 dimensional) as road geometry changes. Several conclusions can be drawn by data quality assessment: (1) consistency analysis shows that the dataset itself is not internally consistent; (2) jerk analysis reveals that a large proportion of anomalies exist; (3) driving regime completeness analysis suggests that the trajectories in the WOD are all incomplete due to short length.

7.1.6 *Data processing*

Since some anomalies and inconsistencies still exist in the WOD, the dataset has been further enhanced by using an optimization-based outlier removal method and a wavelet denoising method. Basically, the outlier removal method is able to regenerate the trajectories in a small time window where the outlier lies in while constraining the border point to be the same as the original trajectory point. The wavelet denoising method can reduce the proportion of anomaly jerk sign inversion to a reasonable level. For more methodological details, please refer to this publication[76]. The processed trajectory data (organized by car-following pairs) and the related codes can be openly accessed from this: [website](#).

7.1.7 *Challenges in the methodology employed/ future work / improvements;*

As one of the largest and also high-resolution datasets that contain real-world deployed automated vehicles, the WOD has great potential in supporting research

such as microscopic driving behavior analysis, trajectory prediction and motion planning under various complex driving scenarios. Nevertheless, to better support traffic flow related studies in the future, several improvements can still be made: (1) the trajectory length should be significantly prolonged; (2) the detection range of the shared data should be expanded to the maximum capability of the lidar sensors; (3) scenarios with multiple interacting automated vehicles should be considered.

7.2 Ohio Instrumented Probe Vehicle Studies

7.2.1 *What was the research question targeted by the team? What can be considered as a potential use?*

This study seeks to investigate the detailed interactions between vehicles traveling through a 22.5 km long freeway corridor (45 km round trip) with several recurring bottlenecks. Thereby providing a glimpse of conditions both upstream and downstream of active bottlenecks. The data set was collected from an instrumented probe vehicle (IPV) equipped with localization and perception sensors to monitor ambient traffic. The probe vehicle traveled a pre-defined route on freeways I-70 and I-71 in Columbus, Ohio, USA, between 2008 and 2011.

So far one day has been processed (two round trips, or roughly 90 km of travel) and shared with the research community. The data is split into three components. The first portion of data is focused on tracking the ambient vehicles [33]. The associated data set includes forward and rearward target tracking with a 180° planar LIDAR field of view in front and rear and a 12° radar field of view in front. The second portion of data is focused on synchronizing the IPV data with roughly 40 loop detector stations along the route [34]. The loop detector stations recorded individual vehicle actuations and the actuations corresponding to the IPV passages were found. These data were used to quantify the limits of what the IPV perception sensors could "see". A third portion of data is still being processed that uses side-view, vertically scanning planar LIDAR sensors to collect point clouds of vehicles that pass the IPV on the side and used these for shape based vehicle classification.

The one day of ambient vehicle tracking and concurrent loop detector data are openly available for download at [49]. The side-view point clouds will be added as soon as that work is finished. In the long run a total of up to 62 days (48 with concurrent loop detector data) will be processed and added to the data set as time and resources allow.

7.2.2 *Data collection perimeter definition*

Details about the study corridor are summarized in the respective papers [33, 34]. The data sets include detailed schematics of the freeway corridor including interchanges and loop detector station locations. The IPV data includes linear distance along and across the corridor as well as curvilinear Easting and Northing coordinates. The IPV data were collected from a moving vehicle while the loop detector data were collected from fixed locations along the route of the IPV.

7.2.3 *Experimental design*

The IPV provided a moving sensing platform that traveled over the 45 km round trip route while the loop detectors provided concurrent individual vehicle actuations along the corridor. The IPV perception sensors included forward and rearward planar scanning LIDAR sensors, each providing a 180° field of view and a forward facing commercial grade adaptive cruise control radar sensor providing a 12° field of view in front. The IPV localization sensors included DGPS and inertial navigation derived from the OBD speed and distance information combined with high resolution gyroscopic heading measurements. Low resolution video was also collected for validation with up to four cameras (forward, rearward, driver's side, and passenger's side)

For the one day that is currently available, over 1100 ambient vehicles are included in the IPV data set (though most are only seen briefly) and the full 24 hours of loop detector data on the given day are included.

7.2.4 *Metadata*

Weather data is not provided, but weather and lighting conditions can be assessed from the included validation video. Upstream and downstream conditions are not provided but many of these can be distilled from the time-space diagrams of speeds from the loop detector data (code included).

7.2.5 *Data*

The IPV data share a common reference clock but are sampled at different rates, mostly in the range 4Hz to 40Hz. The original sampling was preserved to reduce data loss. In the LIDAR data vehicles were first tracked using an automated process and then manually inspected and cleaned up. The LIDAR detected most targets out to 40m and then detection drops off, but the radar extends the forward coverage out to 60m. The data sets include detailed schematics of the freeway corridor including interchanges and loop detector station locations. The IPV data includes linear distance along and across the corridor as well as curvilinear Easting and Northing coordinates.

7.2.6 Data processing

The data are stored in Matlab format and include sample code for generating various plots. The sample code demonstrates ways of working with the data that goes beyond the documentation. For the sample day the raw LIDAR data and video are included.

7.2.7 Challenges in the methodology employed/ future work / improvements;

In the long run a total of up to 62 days (48 with concurrent loop detector data) will be processed and added to the data set as time and resources allow.

7.3 OpenACC

7.3.1 What was the research question targeted by the team? What can be considered as a potential use?

Road transport networks are poised for significant transformation in the coming decades, driven primarily by the introduction of Connected and Automated Vehicles (CAVs). The integration of connectivity and automation is occurring gradually, with distinct levels of automation beginning with single-function systems, such as Adaptive Cruise Control (ACC) for regulating longitudinal vehicle behavior.

The socio-economic and environmental impacts of these interrelated trends could be highly beneficial, potentially reducing accidents, congestion, and emissions. However, without a well-planned transition, unintended negative consequences may arise.

To support European policies on vehicle automation and connectivity, the Joint Research Centre of the European Commission launched an exploratory research project. The objective was to assess the effects of automated driving systems, such as ACC, on traffic congestion, safety, emissions, and driver behavior through empirical observations of commercial vehicles.

As part of this initiative, a series of diverse experimental campaigns was designed to address these key issues. The OpenACC dataset, developed under the Joint Research Centre's open-access policy, comprises car-following experiments involving both human-driven and ACC-equipped vehicles on public roads, test tracks, and the JRC's private urban network. The dataset includes measurements from various vehicle models and brands, utilizing a range of data acquisition systems, from low-cost U-blox and OBD devices to high-precision Differential GNSS systems. The openACC has been widely used in various domains of transportation research. It has supported the development of car-following (CF) development [77], as well as their calibration [?]. Additionally, it has enabled studies comparing autonomous versus human driving behaviors [78], analysis of reaction times [79], and time

gap estimations [80]. The dataset has also contributed to research on traffic control system design, including vehicle controllers [81], as well as studies on energy consumption and emissions [82]. Furthermore, it has been used for safety evaluations [83], traffic flow modeling [84], and investigations into string stability [85].

The OpenACC dataset was collected in a controlled environment with free-flow traffic conditions around the platoons. Both spatial and temporal coverage were ensured, with data gathered from both fixed and moving perspectives. The dataset includes a complete description of the road network, featuring precise lane marking positions and GPS coordinates defining the road boundaries. This comprehensive information provides valuable insights for autonomous vehicle research and traffic flow analysis.

7.3.2 Data collection perimeter definition

The OpenACC dataset was collected through structured experimental campaigns designed to study vehicle behavior in controlled and real-world environments. The key aspects of the experimental setup include:

- 27 vehicle models from 17 different vehicle brands equipped with Adaptive Cruise Control systems for autonomous longitudinal vehicle movement.
- Experiments conducted on test track locations, open roads, and at the Joint Research Centre.
- Similar driving cycles (speed profiles) with and without automation.
- Data acquisition using different devices varying in noise levels, accuracy, and cost.
- Single-lane experiments (no focus on lane changing).
- Road grade information included.
- Experiments conducted with different powertrains, including electric, internal combustion engines, and hybrid vehicles.
- Platoons ranging from two to eleven vehicles in length.

The data collection campaigns covered both on-road and proving-ground environments, with experiments conducted in human-driven and ACC-enabled modes. Individual campaigns lasted between about 3 and 51 hours, covering distances from just under 200 km to over 1,600 km, for a total of several thousand kilometers and dozens of driving hours across all scenarios.

Single-lane experiments with and without Adaptive Cruise Control enabled and no participating trucks (only passenger vehicles). The experiments involved a varied fleet of diesel, gasoline, hybrid, plug-in hybrid, and fully electric vehicles from various models of major brands. Engine sizes ranged from about 1.2 L to 2.5 L, with battery capacities up to 79 kWh for electric models, and top speeds between 170 km/h and 238

km/h. The sample included both compact cars and SUVs from model years 2010 to 2019.

The first version of the OpenACC database includes four experimental campaigns conducted at different locations, two on public roads and two on proving grounds. Each campaign featured a designated leading vehicle (either human-driven or ACC-enabled) followed by multiple vehicles operating in both human and ACC modes. The vehicle fleets varied in size and composition, and tests were performed with different time gap settings, ranging from short to long, and in some cases mixed configurations.

The experimental campaigns were conducted on dedicated test tracks designed for vehicle testing and evaluation. The following are the test sites, including their geographic context and road characteristics: ZalaZone [86] and AstaZero [87].

7.3.3 Metadata

The initial data collection plan is not publicly accessible. Compared to the original plan, fewer scenarios were conducted due to time limitations. Regarding the subject vehicle, no surrounding vehicles were observed (see Table 3 above). The experiments took place at AstaZero and ZalaZone proving grounds, with professional drivers provided by the facilities, following a safety-first experimental protocol. The dataset was collected using onboard sensors, including Ublox devices, Differential GPS, OBD, and CAN systems. No cameras were used, and no additional information on video reconstruction is available. Furthermore, curvilinear data is not provided.

No weather incidents, roadworks, or disruptions were reported along the road stretch or in upstream / downstream sections. A detailed description of the road coordinates is not available. The initial data collection plan is also not publicly accessible. Regarding trajectory characteristics, the dataset provides trajectory data at a fixed frequency of 10 Hz.

Since satellite reception capabilities are not uniform on public road experiments, at a final stage, splines data interpolation was applied to ensure a constant sampling frequency of 10 Hz. A piecewise cubic polynomial being twice continuously differentiable on subintervals of two points was used to interpolate the data. GNSS data contain significant noise levels especially when studying vehicle dynamics with regard to vehicles' acceleration or even jerk.

7.3.4 Data processing

Position and speed vehicle data were acquired with different acquisition systems for the four experimental campaigns. For the first two campaigns that took place on public roads, data acquisition was obtained

in binary format from the U-blox M8 devices, one device installed per vehicle. U-blox M8 devices are commercial systems equipped with accelerometers, gyroscopes and GNSS receivers. OBD and CAN data were also available for some vehicles. The average horizontal accuracy reported by the receivers is less than 50 cm and the sampling frequency is of about 3-5 Hz. Taking into account the difficulties of conducting experiments on public roads, manual cleaning was applied to identify problematic trajectory parts and apply corrections in the OpenACC database. OBD data provide information of vehicle dynamics such as speed, acceleration, engine speed. OBD data are often incomplete, of low frequency and, in most cases, they do not provide enough accuracy for trajectory/speed series reconstruction. However, they were useful in automatically identifying trajectory parts, where the ACC was activated due to recorded acceleration pedal positions (OBD) or driving assistance information (provided by the CAN). This information is available in the OpenACC database. Since the experiment was coordinated among the vehicles' driver, detection of an enabled ACC system in the platoon was used as indication that all vehicles were performing with their ACC systems (if they had this capability). At a final stage, visual inspection of speed series was performed in order to detect noisy parts in the speeds and accelerations. These parts have been removed and are not included in the final dataset. For the first two campaigns, the trajectories corresponding to urban conditions were removed as, for those parts, we were unable to detect per vehicle when the ACC was enabled or not due to the technical problems with the data acquisition devices. Specific instances of cut-ins or lane changes are not removed up to now from the database. An automated detection process is ongoing work for this task. Since satellite reception capabilities are not uniform on public road experiments, at a final stage, splines data interpolation was applied to ensure a constant sampling frequency of 10 Hz. A piecewise cubic polynomial being twice continuously differentiable on subintervals of two points was used to interpolate the data. GNSS data contain significant noise levels especially when studying vehicle dynamics with regard to vehicles' acceleration or even jerk. Some works in the literature use linear interpolation in order to bring acceleration values to acceptable levels, but we have found that this introduces step function pattern in the jerk domain. This makes the data unsuitable for some topics, e.g. for studying the safety response of the vehicles. Therefore, in this work the piecewise cubic polynomial was used. Inevitably, any post-processing introduces errors that counter-balance existing ones. In this case, for an insignificantly small part of the data,

the post-processed speed and acceleration values were found negatives (usually when the vehicles start from or arrive at a stop). In those cases, the smoothed values were bounded to zero. For the AstaZero test track experiment, data acquisition was performed with an inertial navigation system, the RT-Range S multiple target ADAS measurements solution by Oxford Technical Solutions Company, with a differential GNSS accuracy, with precision of 2 cm/s in the speed and 2 cm in the positioning measurements with a sampling rate around 100 Hz. In the final database output, down sampling to 10 Hz was applied as this frequency is enough for capturing the vehicle dynamics in the platoon. Since the test took place in a protective environment, there is no interference from other users and output data ensure the absence of cut-ins or other types of noise.

7.3.5 Challenges in the methodology employed/ future work / improvements

The following factors introduce potential biases and challenges in the experimental data collection, impacting the reliability and consistency of the results across different testing environments:

- Bias due to safety requirements and human driver knowledge on the scenarios in experiments in test tracks.
- Time limitations in execution of multiple scenarios.
- Different noise levels on GPS signals in public road tests.
- Cut-ins from non-participating vehicles on open roads for long headways.

7.4 Traffic Open Data (TOD) experimental dataset

7.4.1 What was the research question targeted by the team? What can be considered as a potential use?

The phenomena in traffic flow arise from the site-specific empirical observations that include various confounding factors, such as bottleneck type, bottleneck strength, vehicle type, and lane number. These complexities make it challenging to isolate the role of each factor in shaping traffic flow evolution. To overcome this challenge, controlled traffic flow experiments provide an alternative to empirical observations by enabling the isolation of multiple influencing factors, and allows for a deeper understanding of traffic dynamics by pinpointing the true origins of puzzling traffic flow phenomena.

To this end, a series of diverse experimental campaigns have been conducted to collect real-world vehicle trajectory data and promote research in this field. The following mentioned experimental studies have been published, with more forthcoming. Researchers

interested in accessing relevant datasets can submit data requests via this [website](#). Recent experiments are listed below.

- 25-car-platoon experiment [88, 89]: This experiment is designed to investigate whether traffic oscillations still occur in a platoon unaffected by external perturbances (i.e., with a leading vehicle maintaining a constant speed). It examines how traffic oscillations evolve and grow along the platoon and how the speed-spacing relation is.
- 51-car-platoon experiment [90]: Building on the 25-car-platoon experiment, this experiment is designed to investigate how traffic oscillations evolve and grow along a longer platoon.
- 11-car-platoon highway experiment [91]: While the 25- and 51-car platoon experiments were conducted at speeds below 60 km/h, this experiment is designed to investigate how traffic oscillations evolve and grow at higher speeds (above 60 km/h).
- Lightly congested circular track experiment [92]: In addition to platoon experiments, another type of car-following experiment has been conducted on a circular road track. This experiment is designed to investigate the spatiotemporal characteristics and mechanisms of lightly congested flow under periodic boundary conditions.
- Fixed bottleneck experiment [36]: In platoon experiments, traffic flow evolution is induced by a leading vehicle moving at a speed lower than the desired speed of the following vehicles, with no overtaking allowed. Thus, the leading vehicle can be regarded as a moving bottleneck with zero permeability. This experiment is designed to investigate how traffic oscillations evolve and grow along the road upstream of fixed bottlenecks stand still on roads.

These experiments were conducted in controlled environments, either on open roads or on artificial constructed tracks in closed test fields. The absence of external vehicle interference ensures high-quality data for testing traffic models and theories. The insights gained from these studies contribute to a better understanding of traffic flow behavior.

7.4.2 Data collection perimeter definition

All vehicles used in the experiment were passenger gasoline cars. The TOD experiments took place on urban roads, highways, closed tracks, and a test circuit, with vehicle GPS trajectories collected and visual documentation from each scenario. The experiments were conducted on a single-lane track under car-following conditions, with overtaking prohibited. The traffic situations are provided as follows.

- On the open road, traffic was very light, and the selected road section had no traffic lights, ensuring an uninterrupted vehicle platoon. In the platoon experiments, the leading vehicle was controlled to maintain a constant speed for a while with the help of cruise control system or human control. The tested traffic speed ranges from 5 to 80 km/h. The spatial coverage is dynamic.
- In the circular track experiments, traffic flow conditions were regulated by adjusting the number of vehicles (resulting in different global densities in the absence of bottlenecks) or by modifying bottleneck intensity by imposed different speed limits. The spatial coverage is fixed.

The TOD experimental campaigns covered both platoon (11–51 vehicles on open roads, closed tracks, and highways) and circular driving scenarios (30–40 vehicles on an 812 m circuit), with durations ranging from 0.9 h to 5.5 h and total distances between 680 km-veh and 3780 km-veh.

7.4.3 Experimental design

The data collection plan for each experiment is detailed in corresponding literature and is publicly accessible. Due to space constraints, these plans are not provided here. The experimental designs are as follows:

- 25-car-platoon experiment: The driver of the leading car is required to maintain a constant speed at pre-determined values. After a set period, the driver will receive instructions to adjust to the next designated speed. Other drivers are required to drive their cars as they normally do, but overtaking is not allowed. The tested speeds include 7, 15, 20, 25, 30, 35, 40, 45, 50, 55, and 60 km/h.
- 51-car-platoon experiment: This experiment follows the same setup as the 25-car-platoon experiment. The tested speeds include 5, 12, 25, 30, 40, and 50 km/h.
- 11-car-platoon highway experiment: This experiment uses the same setup as the 25- and 51-car platoon experiments, but the platoon is tested at higher speeds of 60, 70, and 80 km/h.
- Lightly congested circular track experiment: The circular road track, marked by traffic cones, has a circumference of 812 meters. The number of cars used in the experiment is 40, 35, and 30, corresponding to vehicle densities of 0.049, 0.043 and 0.037 m^{-1} , respectively. Drivers are instructed to “Drive as if you were on rush hour expressways. Follow the vehicle ahead closely, whenever safety permits”.
- Fixed bottleneck experiment: This experiment uses the same circular track as the lightly congested experiment. A 30-meter speed limit bottleneck area is created by placing two pillars at

each end of the speed limit zone, making the bottleneck physically recognizable for drivers. The speed limit in the bottleneck area is set to 10 and 20 km/h for two independent runs. Drivers are instructed to “Drive as if you were on rush hour free-ways. Follow the vehicle ahead closely, whenever safety permits. Strictly observe the speed limit.”

7.4.4 Metadata

No weather data was collected for the experiments. However, weather and lighting conditions can be inferred from the experimental snapshots. No incidents occurred during the experiments. The upstream and downstream conditions, as well as the road coordinates, are not relevant to the objectives of these experiments, as the data is not derived from site-specific empirical observations.

7.4.5 Data

In the experiment, vehicle trajectory data was collected using GPS devices installed on each car with a sampling frequency of 10 Hz. The trajectory characteristics include time, speed, and x- and y-coordinates in the local plane. The measurement errors of the GPS devices are within ± 1 meter for location and within ± 1 km/h for velocity.

7.4.6 Data processing

Data loss primarily occurred due to brief interruptions in GPS signals. However, this missing data can be negligible and does not impact the overall analysis. Additionally, missing data can be estimated and reconstructed using information from adjacent vehicles and from before and after the gap.

7.4.7 Challenges in the methodology employed/ future work / improvements

The artificially designed experimental plan is limited and cannot cover all possible scenarios due to time constraints. As the experiment goes on, drivers may become more familiar with the driving conditions and get accustomed with the setup and the situation. Additionally, the experimental instructions may influence driving behavior, which could in turn influence the experimental results[93].

In future work, larger-scale experiments on longer road sections, with larger platoon sizes and more diverse experimental instructions, are expected to be conducted.

7.5 Flanders dataset

7.5.1 What was the research question targeted by the team? What can be considered as a potential use?

KU Leuven university and the MOW ministry of public works of Flanders want to investigate the role of

lateral maneuvers (lane changes) in the efficiency and safety of traffic operations in complex motorway corridors, consisting of a rapid succession of on/off ramps and interchanges where frequent merging, diverging and weaving occurs. In Flanders, such busy corridors are typically equipped with double inductive loop detectors per lane at spatial intervals of 400-800m. Individual vehicle data (IVD) as well as minute-aggregate data is permanently being logged for over 10 years and counting. Through a traffic flow crystallization algorithm [94], the IVD recordings of individual vehicles over successive detector locations are matched, herewith inferring the driving lane (and thus: lane changes) at discrete spatial locations. The busier traffic, the more trajectories constrain each other (less free driving and more platooning), such that a spatially continuous reconstruction of the vehicle paths can be inferred both longitudinally and with approximate lane change locations between detector stations.

While such discrete and reconstructed trajectories are less accurate than continuously observed paths, it has the advantage that lane changes and trajectories can be reconstructed for any segment and time for which the raw IVD have been systematically logged. For example: the data warehouse of the MOW Flanders stores IVD of billions of vehicles over several years for more than 300km of motorway; in principle, each of these vehicles' trajectory can be reconstructed. This can be used for a variety of studies like for example: variation of tactical lane change behavior in weaving or other bottleneck sections with traffic composition, traffic conditions, weather; longitudinal studies of lane choice and consecutive lane change decisions of the same vehicle tracked over several kilometers; complete reconstruction of crash scenarios for historic crash databases collected over extended periods (multiple years); systematic monitoring of changes in car following, platoon structures, lane changes etc. during introduction and uptake of autonomous driving technology; before-during-after impact evaluation of infrastructural or traffic management measures (including retrospective studies for which no evaluation had been planned).

7.5.2 Data collection perimeter definition

In contrast to direct trajectory observations, the availability of reconstructed trajectories only depends on the availability of the raw individual vehicle data (IVD) collected at sufficiently close spatial intervals. The Flemish road operator stores these IVD permanently since more than 10 years; detector spacing is close enough for the reconstruction in any motorway segment where the RSS lane management system has been installed (approx. 300km, see RSS locations

posted on [95] or as blue markers on the live map of [96] by selecting "RSS border"). Knowing that some of these sections process more than 100,000 veh per day, the trajectories can be reconstructed for several billions of vehicles in all conceivable weather and traffic conditions, and road configurations.

In publications using these data [37, 94, 97–99], a maximally challenging selection of trajectories on a 3.3km long weaving section of 5 lanes during 5 days was reconstructed from raw IVD. A subset of 11,360 of them were validated using CCTV observations, showing a 97-99% accuracy of the matching process in all traffic conditions. In addition, cross validation with a GPS high-resolution trajectory dataset (2% sample of all traffic) and a single, highly accurate differential GPS probe vehicle further demonstrated validity of the trajectory reconstructions.

7.5.3 Experimental design

There was no specific experimental design, given that no new trajectory data was collected, rather reconstructed from already available IVD data.

7.5.4 Metadata

While the IVD data (source for the trajectory reconstruction) is not publicly available, there exists an open-data stream of 1-minute aggregate traffic data of the same detector set on the same road network. Metadata on the locations of the detectors and characteristics of the segments are published along with these open data [100, 101]. Incident and road works data are systematically collected by the Flemish road operator but are not openly available. They can be requested for non-commercial use under mild license conditions (see contact persons listed on the cited open-data websites). Local weather data of the site of interest can be collected through dedicated (commercial and open) online services.

7.5.5 Data

A data sample of reconstructed trajectories is available upon request with the author team of KU Leuven. For other and/or larger data sets, a license on the raw IVD should be negotiated with the Flemish road operator. Efforts to publish the Traffic Flow Crystallization (TFC) algorithm as open-source toolkit are ongoing (summer 2025). Validation on a very challenging case (5 lanes with dense, frequently lane-changing traffic) showed that reidentification of the same vehicle on successive detectors was successful in 99.15% and 96.97% of the cases in free and congested traffic respectively. Lane change rates are underestimated by 2%. For more details on validation and error statistics, see [94].

7.5.6 Data processing

An open-source version of the TFC algorithm including an unprocessed and processed data sample is in preparation (summer 2025). This code will be documented such that all processing can be reproduced.

7.5.7 Challenges in the methodology employed/ future work / improvements

As discussed, discrete trajectories need to be reconstructed from raw IVD data. The TFC algorithm requires some configuration to the local road and detector layout and for calibration of some empirical distributions that need to be retrieved from the raw data. While the algorithm has been validated under challenging conditions, there has been no validation on transferability of the reported error statistics to other situations.

In less challenging road configurations and/or traffic conditions, the algorithm may perform far better still. On the other hand, the detector spacing on the tested site was favorable, with average detector distance of 470m. It is to be expected that matching performance will decrease with increasing distance; but it was not tested so is not known how important this decrease can be. Also, even if matching would be perfect, there might be more missed lane changes in between detectors as detector spacing grows. To conclude: this alternative type of trajectory data has interesting characteristics (especially the, in theory, unlimited spatiotemporal coverage) that are complementary to video-collected trajectories; but experience is, so far, limited.

7.6 Jatayu dataset

7.6.1 What was the research question targeted by the team? What can be considered as a potential use?

The study targets the development of a comprehensive dataset for vehicle detection and tracking in mixed-traffic scenarios prevalent in India and Southeast Asia. The primary objective is to address the gaps in existing UAV datasets, which often lack the diversity and real-world complexity required for effective Intelligent Transportation Systems (ITS).

The dataset can be utilized for various applications:

- Demonstrator application: traffic surveillance, vehicle detection, and tracking using UAV imagery for urban road networks. The dataset enables the training and validation of object detection models like YOLO for improved accuracy in vehicle identification.
- Potential applications are: traffic safety evaluation, congestion analysis, vehicle trajectory tracking, detection of traffic violations (e.g., wrong-way driving, improper parking), Enhancing ITS frameworks for smart cities

7.6.2 Data collection perimeter definition

The dataset includes a diverse mix of vehicle types reflecting Indian traffic scenarios. It incorporates motorized and non-motorized vehicles such as: Auto-rickshaws; Bicycles; Buses; Cars; Cycle-rickshaws; E-rickshaws; Motorcycles; Pedestrians; Tractor-trolleys; Trucks

Data was collected from six Indian cities representing urban and semi-urban environments.

The dataset captures a range of traffic conditions: free flow, congestion, daytime and nighttime traffic (under streetlights). The dataset covers both spatial and temporal aspects, with :

- Spatial: approximately 250 meters of road length coverage achieved by flying the drone over a maximum altitude of 120 meters.
- Temporal: 33 hours of video, including 2.5 hours at night under street lighting conditions.

The dataset includes various infrastructure types such as roundabouts, intersections, highways, and bus terminals. It also covers diverse terrains to ensure comprehensive geographic diversity.

The dataset encompasses:

- Total vehicle coverage: Over 215,000 labeled objects.
- Temporal span: equivalent to 33 hours of traffic data.
- Trajectory data: as the dataset does not directly provide trajectories, it is difficult to estimate veh*km*minutes . Instead, it provides labeled images for vehicle detection and tracking to extract the trajectories.

7.6.3 Experimental design

The study planned to use UAVs for high-resolution video capture across multiple urban and semi-urban locations in India. Data collection included diverse lighting conditions and traffic scenarios. The dataset is accessible through the aforementioned GitHub link.

The data collection closely followed the initial plan, with drone flights conducted over 33 hours across six cities, capturing varied traffic conditions, including day and night scenarios. Minor adjustments were made for weather conditions and ensuring clear visibility during nighttime recording.

The dataset includes 10 observed vehicle classes, featuring unique types such as auto-rickshaws, cycle-rickshaws, e-rickshaws, and tractor-trolleys. Mixed traffic conditions were recorded, with high vehicle density in urban areas and moderate density in semi-urban settings.

Data was collected from various sites, including roundabouts, intersections, highways, and bus terminals. Aerial observations were performed without spe-

cific driver recruitment. The experimental protocol involved predefined flight paths and altitudes to ensure consistent data collection.

The DJI Mavic Mini 3 Pro UAV was used, equipped with a 48 MP camera capable of 4K-60fps resolution. A 3-axis mechanical gimbal provided stabilization for clear video capture. No additional sensors, such as GPS or LIDAR, were employed.

Video frames were extracted and annotated using tools such as Label Studio and CVAT. Annotations were in YOLO bounding box format with normalized coordinates. Data augmentation techniques, including rotation, scaling, and translation, were applied to enhance detection capabilities.

Drone specifications ensured consistent flight performance, with a maximum flight time of 47 minutes per session and data collection conducted at an altitude of 120 meters.

7.6.4 Metadata

The related information is presented in the previous sections.

7.6.5 Data

The related information is presented in the previous sections.

7.6.6 Data processing

The related information is presented in the previous sections.

7.6.7 Challenges in the methodology employed/ future work / improvements;

Methodological limitations and data collection constraints include:

- Limited flight time: The drone's maximum flight duration of 47 minutes per session required frequent landings and take-offs, which slightly disrupted the continuity of data collection.
- Under-represented vehicle classes: Certain vehicle types, such as bicycles and pedestrians, were less frequent in the dataset, impacting model performance for these categories.
- Occlusion: tree coverage and dense traffic caused partial occlusion of vehicles, complicating annotation and detection tasks.
- Nighttime data collection: ensuring clarity and quality of nighttime video under varying street-light conditions was challenging.

Efforts will focus on:

- Enhanced dataset diversity: incorporate data from different altitudes, camera angles, and additional geographic locations. Expand the dataset with more images of under-represented vehicle classes.

- Integration of additional sensors: use of GPS, LIDAR, or thermal cameras to complement video data and provide richer annotations.
- Algorithmic Improvements: development of improved methods to handle occlusion and enhance detection accuracy for small or partially visible objects.
- Trajectory analysis: addition of curvilinear trajectory data for deeper insights into vehicle movement patterns.

Further improvements may include:

- Augmented data collection: extend recording sessions to capture diverse weather and traffic conditions, including adverse scenarios like rain or fog.
- Collaborative efforts: partner with other research institutions to standardize UAV datasets and share resources for annotation and model training.
- Real-Time applications: focus on enabling real-time traffic monitoring and integrating the dataset into ITS solutions for smart cities.

These enhancements aim to make the Jatayu dataset more robust and versatile for various applications in traffic monitoring and ITS.

7.7 I-24 MOTION Video dataset

7.7.1 What was the research question targeted by the team? What can be considered as a potential use?

Long-distance complete trajectories offer valuable opportunities for microscopic modeling in traffic flow theory. The research question targeted by this dataset [102] is multi-object long-distance vehicle trajectory tracking across hundreds of cameras. Users can utilize the raw video dataset collected from 234 HD-cameras to develop new vehicle detection and tracking algorithms with benchmarks to compare. While the quality of the vehicle trajectory data can be improved with better computer vision algorithms, the size of the dataset remains valuable for traffic flow modeling, covering approximately 6.75 km during 1 hour of congested traffic.

7.7.2 Data collection perimeter definition

I-24 MOTION video dataset is a multi-camera tracking dataset consisting of 234 hours of video data recorded concurrently from 234 overlapping HD cameras covering a 6.75 km stretch of 8-10 lane interstate highway near Nashville, TN. GPS trajectories from 270 vehicle passes through the scene are manually corrected in the video data to provide a set of ground-truth trajectories for recall-oriented tracking metrics [102].

7.7.3 Experimental design

The data is collected during a period of morning peak hour with stop-and-go traffic.

7.7.4 Metadata

Homography information from each camera that transforms coordinates from the camera coordinate system to the global coordinate system is provided. Additionally, the transformation from the global coordinate system to the roadway mile marker system is also provided.

7.7.5 Data

The dataset includes long-distance vehicle trajectories with position, speed, and lane information, along with vehicle attributes and roadway geometry. It is publicly available, with code for reproduction and further development at this [link](#), and a video tutorial at this [link](#).

7.7.6 Data processing

The main purpose of this dataset is to improve the cross-camera tracking algorithm. The raw data for detection and the tracking is provided as it is and the postprocessing code in [103] is provided at this [github repo](#) for reproducible usage.

7.7.7 Challenges in the methodology employed/ future work / improvements;

In our paper we demonstrate the challenges of tracking vehicles for 234 cameras using a Kalman Filter-based method. Also the detection algorithm is not up-to-date. We invite the community to use the dataset and work on this challenging problem to improve the state-of-the-art. We sincerely encourage the community to share any issues and provide feedback at this [link](#).

7.8 Zen Traffic Data Dataset

7.8.1 What was the research question targeted by the team? What can be considered as a potential use?

Zen Traffic Data, collected by Hanshin Expressway Co., Ltd. using image-sensing technology, is designed to support research on transportation safety and efficiency.

This high-frequency, detailed dataset enables in-depth investigations into traffic flow characteristics, vehicle interactions, and driving behaviors, ultimately contributing to safer and more efficient roadways. Researchers have leveraged Zen Traffic Data for various applications, including traffic flow modeling (car-following and lane-changing behavior), traffic operations, traffic phenomena analysis, safety assessments,

traffic management strategies, predictive models for transportation planning and infrastructure improvements, automated and connected driving systems, etc. It has also been utilized in. A (incomplete) list of publications that used Zen Traffic Data can be found on its official website. The first paper that introduced Zen Traffic Data seems to be Kodama et al. (2018) [43].

7.8.2 Data collection perimeter definition

As of February 2023, traffic data are available for approximately 2 km sections of the Hanshin Expressway, including:

- Route 11 (Ikeda Line, Osaka-bound) near Tsukamoto Junction, which is a key corridor connecting Osaka International (Itami) Airport (ITM), the Meishin and Chugoku Expressways, and other major routes to Osaka City. Located on the right bank of the Yodo River, just outside the city center, this junction experiences heavy traffic on its two-lane roads in each direction. The complex road environment, featuring S-shaped curves, sags, and the Tsukamoto Interchange, frequently generates shock waves at convergence points, curves, and the Yodo River bridge.
- Route 4 (Wangan Route, Osaka-bound) at the Ohama-Sambo section, which is a major route connecting southern Osaka to the city center. The section has two lanes in each direction and experiences increased traffic volume in the morning. Due to its complex road environment, including Ohama Junction, left-hand curves, and sags, shock waves are occasionally observed.
- Route 13 (Higashiosaka Route, Nara-bound) near Morinomiya Junction, which is a key corridor linking Osaka and Hyogo to Nara, carrying the highest daily traffic volume on the Hanshin Expressways. Frequent congestion occurs due to multiple route mergers and gradient slopes causing speed reductions.

7.8.3 Experimental design

Following a structured plan to ensure thorough coverage of the designated road section and timeframe, Zen Traffic Data is obtained through cameras installed on light poles along Japan's Hanshin Expressway.

7.8.4 Meta data

This dataset consists of five one-hour traffic recordings with traffic conditions ranging from free-flow to congestion. Vehicles in the target sections are recorded at 0.1-second intervals, with location data measured at the rear center of each vehicle. However, motorcycles are not included. In addition to trajectory data, the dataset provides vehicle attributes (e.g., car length),

precise location details, and road surface characteristics.

Two types of traffic-related data are included: section datasets and trajectory datasets.

- Section datasets provide macroscopic traffic flow information at 500-meter intervals for each lane, including:
 - An overview of the target sections, detailing location, road surface conditions, road markers, and signs.
 - An overview of traffic, featuring cross-sectional traffic volume, congestion occurrences, speed contour graphs, and more.
- Trajectory datasets offer microscopic driving information for individual vehicles, including vehicle ID, time, position (latitude, longitude, lane, kilopost), speed, vehicle dimensions and class, and road surface characteristics (slope, cross slope, curvature).

While the dataset does not explicitly include weather conditions, researchers can integrate external sources for this information. It provides lane markings, GPS coordinates, and kilopost distances, ensuring a detailed spatial representation of the recorded roadway.

7.8.5 Data processing

The Zen Traffic Data team implemented some basic methods to handle missing data. Due to occlusion and other image-sensing limitations, some vehicle trajectories may be discontinuous. However, each vehicle retains its ID throughout trips within the target section. When trajectory data is missing, it is reconstructed using interpolations based on measured points before and after the gap. Each record specifies whether the data was directly captured by cameras or derived through calculation.

Recently, Li et Al. (2025) [104] evaluated six outlier removal methods using Zen Traffic Data, finding that their adaptive multi-layer vehicle trajectory reconstruction method outperformed all others across multiple evaluation criteria. The dataset processed using their method is available upon request. Additionally, Li et al. (2024) [105] applied a car-following (CF) extraction algorithm to Zen Traffic Data, identifying 246 CF pairs and generating three datasets: kilopost-based, geography-based, and velocity-based. The extracted CF data captured a wider range of driving conditions than NGSIM, highlighting Zen Traffic Data's richness in driving behavior insights.

7.9 Third Generation Simulation (TGSIM) Data

7.9.1 What was the research question targeted by the team? What can be considered as a potential use?

The Third Generation Simulation (TGSIM) dataset aims to provide real-world data from human-ADAS in-

teractions at various levels of automation in both highway and urban environments. The highway interactions focus on car-following (collected using SAE Level 1 and Level 2 vehicles) and automated lane-changing maneuvers (collected using SAE Level 2 vehicles), covering behavior along different highway geometries (including near on/off ramps as well as a major diverge section). The urban environment focuses on data from SAE Level 3 vehicle interactions with vehicles, pedestrians, bicyclists, and scooterists in an urban setting, including stopped-controlled and signal-controlled intersections. The highway environment data covers two locations in Chicago, IL (i.e., I-294 and I-90/I-94) and one location in Washington D.C. (i.e., I-395). It is important to note that in addition to data from human-ADAS interactions, a considerable amount of data is available from human-human interactions in both highway and urban environments. The TGSIM dataset can be utilized to investigate the impacts of ADAS on human behavior in both highway and urban environments, analyze the string stability of ACC systems in real-world settings, investigate interactions with heavy vehicles and their impact on traffic flow, analyze mandatory and discretionary lane-changing maneuvers, and analyze the car-following behavior of human drivers and ADAS-equipped vehicles. Some of the data in the TGSIM dataset can be also utilized for macroscopic traffic flow analysis.

7.9.2 Data collection perimeter definition

The TGSIM dataset was collected on I-90/I-94 (two datasets) near Chicago, IL, I-294 near Hinsdale, IL, I-395 in Washington D.C., and George Washington University Campus (Foggy Bottom) in Washington D.C. Based on the data collection objectives and location-related limitations, three different data collection methods were utilized: (1) Fixed Location Aerial Videography where a hovering helicopter covered a fixed length of a roadway. This approach was utilized on I-90/I-94 to cover dynamics at a major diverge section; (2) Moving Aerial Videography where a helicopter followed one or more ADAS-equipped vehicles as they move through the target segment. This approach was utilized on I-90/I-94 as well as I-294 and offers several minutes long trajectories; (3) Infrastructure Based Videography where data was collected using cameras attached to infrastructure. This approach was utilized on both I-395 (cameras located on overpasses) and in George Washington University Campus (cameras mounted on multiple buildings).

7.9.3 Metadata

Trajectory data from vehicles, pedestrians, bicyclists, and scooterists were recorded at 0.1-second intervals.

A total of six datasets are included with durations ranging from 2 to 3 hours. The vehicle trajectory data for highway scenarios include Vehicle ID, time stamp, x-location, y-location (location along the highway), lane number, vehicle speed, vehicle acceleration, vehicle length, vehicle width, vehicle type (ADAS-equipped or otherwise). The urban scenario data include object ID, time stamp, x location, y location, lane number (if applicable), speed in x direction, speed in y direction, acceleration in x direction, acceleration in y direction, object length, object width, and object type.

7.9.4 Data processing

While the fundamentals of trajectory extraction remains the same across different TGSIM datasets, the details differ based on the method of data collection and location requirements. [106] introduces the details of the process for each dataset. The overall process, however, consists of (1) data processing which focused on extracting raw images and generating a reference image (using a combination of a satellite image and collected images); (2) object detection which focused on generating bounding boxes associated with each object; (3) object tracking which focused on tracking each detected object across all observations; (4) image stabilization which focused on converting local coordinates to global coordinates based on the reference image; (5) trajectory construction which focused on extracting the trajectories based on the tracking and image stabilization output; and (6) data cleaning which focused on removing outliers and fixing potential issues in the trajectory data.

7.9.5 Challenges and opportunities for improvement

The TGSIM dataset captures interactions among ADAS-equipped vehicles and other roadway users in both highway and urban environments. The focus of the dataset, however, is on lower levels of autonomy. With the increase in availability of SAE Level 4 data, it is critical to collect a similar set of data from higher levels of autonomy.

Abbreviations

CF: car-following; LC: lane change; HDV: human-driven vehicle; ACC: automated cruise control

Author details

References

- Liu, C. (2025). Vehicle trajectory data processing, analytics, and applications. *ACM Computing Surveys*.
- Du, J., Meng, C., and Liu, X. (2024). Analysis of urban residents' travelling characteristics and hotspots based on taxi trajectory data. *Applied Sciences*, 14(3), p. 1279.
- Zhou, H., Ma, K., Liang, S., Li, X., and Qu, X. (2024). A unified longitudinal trajectory dataset for automated vehicle (ultra-av). *Scientific Data*, 11, p. 1123.
- Chandler, R. E., Herman, R., and Montroll, E. W. (1958). Traffic dynamics: Studies in car following. *Operations Research*, 6(2), p. 165–184.
- Gerlough, D. L. and Huber, M. J. (1975). *Traffic flow theory: a monograph*. Special report / Transportation Research Board, National Research Council. National Research Council, Washington, DC.
- Gazis, D. C., Herman, R., and Rothery, R. W. (1961). Nonlinear follow-the-leader models of traffic flow. *Operations Research*.
- Herman, R. and Potts, R. B. (1961). *Single lane traffic theory and experiment*. International symposium on transportation and traffic theory. Elsevier Ltd, Warren, Michigan, r. herman edition.
- Li, L., Jiang, R., He, Z., Chen, X., and Zhou, X. (2020). Trajectory data-based traffic flow studies: A revisit. *Transportation Research Part C: Emerging Technologies*, 114, pp. 225–240.
- Treiterer, J. and Myers, J. (1974). The hysteresis phenomenon in traffic flow. In D. J. Buckley (editor), *Proc., Sixth International Symposium on Transportation and Traffic Theory*, pp. 13–38.
- Lee, C. E., Rioux, T. W., Copeland, C. R., and of Texas at Austin. Center for Highway Research, U. (1977). *The TEXAS model for intersection traffic development*. FHWA/TX78-184-1.
- Wicks, D. A., Andrews, B. J., Goldblatt, R. B., and KLD Associates, I. (1977). *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model, Volume 4: Program Documentation*. FHWA-RD-76-79.
- (1996). *CORSIM user manual version 1.01*.
- Fellendorf, M. and Vortisch, P. (2010). *Microscopic Traffic Flow Simulator VISSIM*, volume 145 of *International Series in Operations Research & Management Science*, p. 63–93. Springer New York, New York, NY.
- Krajzewicz, D., Erdmann, J., Behrisch, M., and Bieker, L. (2012). Recent development and applications of sumo - simulation of urban mobility. *International Journal On Advances in Systems and Measurements*, 5(3 & 43 & 4), p. 128–138. Container-title: International Journal On Advances in Systems and Measurements.
- Ferrer, J. L. and Barceló, J. (1993). Aimsun2: advanced interactive microscopic simulator for urban and non-urban networks. *Research Report*.
- Smith, S. A. and Associates, J. . (1985). *Freeway data collection for studying vehicle interactions: technical report*. FHWA-RD-85-108.
- U.S. Department of Transportation, Federal Highway Administration (2024). Next generation simulation (ngsim) vehicle trajectories and supporting data. <https://catalog.data.gov/dataset/next-generation-simulation-ngsim-vehicle-trajectories-and-supporting-data>. Accessed: 2025-08-15.
- Ossen, S. J. L. (2008). Longitudinal driving behavior: theory and empirics.
- Ossen, S. and Hoogendoorn, S. P. (2011). Heterogeneity in car-following behavior: Theory and empirics. *Transportation research part C: emerging technologies*, 19(2), pp. 182–195.
- Ossen, S., Hoogendoorn, S. P., and Gorte, B. G. (2006). Interdriver differences in car-following: A vehicle trajectory-based study. *Transportation Research Record*, 1965(1), pp. 121–129.
- Buisson, C., Villegas, D., and Rivoirard, L. (2016). Using polar coordinates to filter trajectories data without adding extra physical constraints. Event-title: Transportation Research Board 95th Annual Meeting.
- Punzo, V. and Simonelli, F. (2005). Analysis and comparison of microscopic traffic flow models with real traffic microscopic data. *Transportation Research Record: Journal of the Transportation Research Board*, 1934(1), p. 53–63.
- Punzo, V., Formisano, D. J., and Torrieri, V. (2005). Nonstationary kalman filter for estimation of accurate and consistent car-following data. *Transportation Research Record: Journal of the Transportation Research Board*, 1934(1), p. 2–12.
- Kang, S., Hu, Z., Liu, L., Zhang, K., and Cao, Z. (2025). Object detection yolo algorithms and their industrial applications: Overview and comparative analysis. *Electronics*, 14(66), p. 1104.
- Ali, Y., Sharma, A., and Zheng, Z. (2025). Empirical research on car-following and lane-changing: Recent developments, emerging vehicle technologies' impact, and future research needs. *Transportation Research Interdisciplinary Perspectives*, 31, p. 101368.
- (FHWA), F. H. A. (2006). Next generation simulation (ngsim) data. Accessed: 2025-03-16.
- Li, W. and Sun, J. (2020). Understanding ngsim trajectory data: Challenges and opportunities. *IEEE Transactions on Intelligent Transportation Systems*, 21(3), pp. 1234–1245.
- Zheng, Z. and Buisson, C. (2018). Lessons learned from using ngsim data for traffic modeling and calibration. *Transportation Research Record*, 2672, pp. 45–57.
- Makridis, M. and Ciuffo, B. (2019). Guidelines for data collection inspired by the ngsim experiment. *Journal of Transportation Engineering, Part A: Systems*, 145(8), p. 04019041.
- Arruda, H., Silva, E. R., Lessa, M., Proença Jr, D., and Bartholo, R. (2022). Vosviewer and bibliometrix. *Journal of the Medical Library Association: JMLA*, 110(3), p. 392.
- Sun, P., Kretschmar, H., Dotiwala, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2446–2454.
- Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Charles, Q., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, J., Alexander Shlens, and Anguelov, D. (2021). Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. *arXiv preprint arXiv:2104.10133*.
- Coifman, B., Wu, M., Redmill, K., and Thornton, D. (2016). Collecting ambient vehicle trajectories from an instrumented probe vehicle – high quality data for microscopic traffic flow studies. *Transportation Research Part C*, 72, pp. 254–271.
- Wu, M. and Coifman, B. (2019). Quantifying what goes unseen in instrumented and autonomous vehicle perception sensor data – a case study. *Transportation Research Part C*, 107, pp. 105–119.
- Makridis, M., Mattas, K., Anesiadou, A., and Ciuffo, B. (2021). OpenACC: An open database of car-following experiments to study the properties of commercial ACC systems. *Transportation Research Part C: Emerging Technologies*, 125, p. 103047.
- Zheng, S., Jiang, R., Tian, J., Li, X., Treiber, M., Li, Z., Gao, L., and Jia, B. (2022). Empirical and experimental study on the growth pattern of traffic oscillations upstream of fixed bottleneck and model test. *Transportation Research Part C: Emerging Technologies*, 140, p. 103729.
- Amft, M. A. and Tampère, C. M. J. (2023). Empirical study of lane-changing maneuvers in a weaving area based on reconstructed trajectories of floating car data. *Transportation Research Record*.
- Traffic Open Data (2025). Traffic open data platform. Accessed: March 13, 2025.
- Wu, M. and Coifman, B. (2019). Quantifying what goes unseen in instrumented and autonomous vehicle perception sensor data—a case study. *Transportation research part C: emerging technologies*, 107, pp. 105–119.
- Maji, A., Preetham, K., and Ghosh, I. (2024). JATAYU: A large-scale indian UAV dataset for vehicle detection and tracking. In *Proc. IEEE Int. Conf. on Electronics, Computing and Communication Technologies (CONECT)*.
- Gludemans, D., Wang, Y., Ji, J., Zachar, G., Barbour, W., Hall, E., Cebela, M., Smith, L., and Work, D. B. (2023). I-24 motion: An

- instrument for freeway traffic science. *Transportation Research Part C: Emerging Technologies*, 155, p. 104311.
42. Seo, T. et al. (2020). Evaluation of a large-scale complete vehicle trajectory dataset on a two-kilometer highway segment: Zen Traffic Data. In *Proc. International Symposium on Transportation Data & Modelling (ISTDM)*.
 43. Kodama, T., Masumoto, H., Suzuki, K., Tanabe, J., Nakanishi, M., Tago, Y., Nakajo, S., and Koibuchi, M. (2018). Utility study on traffic flow data generated by image sensing. In *Proceedings of Infrastructure Planning*, 57. Tokyo. (in Japanese).
 44. Talebpour, A., Mahmassani, H. S., and Hamdar, S. H. (2024). Third generation simulation data (tgsim): A closer look at the impacts of automated driving systems on human behavior. Technical report, U.S. Dept. of Transportation (FHWA).
 45. Mahajan, V. et al. (2023). Treating noise and anomalies in vehicle trajectories from an experiment with a swarm of drones. *IEEE Transactions on Intelligent Transportation Systems*.
 46. Krajewski, R., Bock, J., Kloecker, L., and Eckstein, L. (2018). The highD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *21st IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2118–2125.
 47. Coifman, B., Li, L., and Xiao, W. (2018). Resurrecting the lost vehicle trajectories of treiterer and myers with new insights into a controversial hysteresis. *Transportation Research Record*, p. 14p.
 48. Treiterer, J. and research staff (1969). Investigation of traffic dynamics by aerial photogrammetry techniques. Technical Report EEB 278-2, Experiment Station, Ohio State University, Columbus.
 49. Coifman, B. (2024). Coifman trajectory data sets. <https://u.osu.edu/coifman.1/data-sets/>.
 50. Thiemann, C., Treiber, M., and Kesting, A. (2008). Estimating Acceleration and Lane-Changing Dynamics from Next Generation Simulation Trajectory Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2088(1), pp. 90–101.
 51. Punzo, V., Borzacchiello, M. T., and Ciuffo, B. (2011). On the assessment of vehicle trajectory data accuracy and application to the next generation simulation (ngsim) program data. *Transportation Research Part C: Emerging Technologies*, 19(6), pp. 1243–1262.
 52. Montanino, M. and Punzo, V. (2015). Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. *Transportation Research Part B: Methodological*, 80, pp. 82–106.
 53. Coifman, B. and Li, L. (2022). A new method for validating and generating vehicle trajectories from stationary video cameras. *IEEE Transactions on Intelligent Transportation Systems*, pp. 16227–16236.
 54. Chen, X., Zhu, M., Chen, K., Wang, P., Lu, H., Zhong, H., Han, X., Wang, X., and Wang, Y. (2023). Follownet: A comprehensive benchmark for car-following behavior modeling. *Scientific Data*, 10(1), p. 828.
 55. Zhou, X., Wang, Y., Li, J., and et al. (2024). Ultra-av: A unified longitudinal trajectory dataset for autonomous vehicle research. *Scientific Data*.
 56. Commission, E. (2024). FAIR and open Data Management Plan template. <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/horizon-results-platform/75669>. Template DMP conforme aux principes FAIR.
 57. Parliament, E. and Council (2023). Regulation (EU) 2023/2854 – Data Act. <https://data.europa.eu/eli/reg/2023/2854/oj>. Entré en vigueur en janvier 2024, applicable depuis septembre 2025.
 58. Riehl, K., Kouvelas, A., and Makridis, M. A. (2025). Revisiting reproducibility in transportation simulation studies. *European Transport Research Review*, 17(1), p. 22.
 59. Riehl, K., El-Baklisch, S. K., Kouvelas, A., and Makridis, M. A. (2025). Consistent vehicle trajectory extraction from aerial recordings using oriented object detection. *Scientific Reports*, 15, p. 27522.
 60. Chaudhari, A. A., Treiber, M., and Okhrin, O. (2025). Mitra: A drone-based trajectory data for an all-traffic-state inclusive freeway with ramps. *Scientific Data*, 12(1), p. 1174.
 61. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee.
 62. Dua, D. and Graff, C. (2019). Uci machine learning repository [<http://archive.ics.uci.edu/ml>]. irvine, ca: University of california, school of information and computer science. *IEEE transactions on pattern analysis and machine intelligence*, 1(1), pp. 1–29.
 63. Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., and Liu, H. X. (2023). Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953), pp. 620–627.
 64. Arun, A., Haque, M. M., Bhaskar, A., Washington, S., and Sayed, T. (2021). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. *Accident Analysis & Prevention*, 153, p. 106016.
 65. Haghani, M. (2024). E-scooters are linked with injuries and hospital visits – but we can't say they are riskier than bikes yet. UNSW Newsroom.
 66. Fernandez, A. N., Li, K. D., Patel, H. V., Allen, I. E., Ghaffar, U., Hakam, N., and Breyer, B. N. (2024). Injuries with electric vs conventional scooters and bicycles. *JAMA network open*, 7(7), pp. e2424131–e2424131.
 67. Deutsche Welle (2024). Germany: E-scooter accidents and fatalities on the rise. Deutsche Welle.
 68. Yan, Z., Li, J., Hang, P., and Sun, J. (2025). Onsitevru: A high-resolution trajectory dataset for high-density vulnerable road users. *arXiv preprint arXiv:2503.23365*.
 69. Heinonen, S. e. a. (2024). The e-biking in switzerland (ebis) study: methods and dataset.
 70. International Energy Agency (2023). Transport – tracking report. International Energy Agency.
 71. Wu, F., Stern, R. E., Cui, S., Delle Monache, M. L., Bhadani, R., Bunting, M., Churchill, M., Hamilton, N., Haulcy, R., Piccoli, B., et al. (2019). Tracking vehicle trajectories and fuel rates in phantom traffic jams: Methodology and data. *Transportation Research Part C: Emerging Technologies*, 99, pp. 82–109.
 72. Yang, Z., Zheng, Z., Kim, J., and Rakha, H. (2025). Eco-cooperative adaptive cruise control for platoons in mixed traffic using single-agent and multi-agent reinforcement learning. *Transportation Research Part D: Transport and Environment*, 142, p. 104658.
 73. Hu, X., Zheng, Z., Chen, D., and Sun, J. (2023). Autonomous vehicle's impact on traffic: Empirical evidence from waymo open dataset and implications from modelling. *IEEE Transactions on Intelligent Transportation Systems*, 24(6), pp. 6711–6724.
 74. Ali Y, C. D., Sharma A (2024). Investigating autonomous vehicle discretionary lane-changing execution behaviour: Similarities, differences, and insights from waymo dataset. *Analytic Methods in Accident Research*, 42, p. 100332.
 75. Liu, J., Hang, P., Hu, X., and Sun, J. (2024). Evaluation of drivers' interaction ability at social scenarios: a process-based framework. *Transportmetrica B: Transport Dynamics*, 12(1), p. 2380913.
 76. Hu, X., Zheng, Z., Chen, D., Zhang, X., and Sun, J. (2022). Processing, assessing, and enhancing the waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research Part C: Emerging Technologies*, 134, p. 103490.
 77. Jin, X., Li, X., and Wang, J. (2022). Car-following model development using naturalistic driving data. *IEEE Transactions on Intelligent Transportation Systems*, 23(4), pp. 3456–3471.
 78. Wang, M., Li, S., and Zhang, R. (2021). Autonomous vs human driving: A comparative analysis of naturalistic trajectories. *Transportation Research Part C: Emerging Technologies*, 129, p. 103305.
 79. Zhang, W., Li, T., and Sun, Y. (2022). Estimating reaction times in human and automated driving systems. *IEEE Transactions on Intelligent Transportation Systems*, 23(6), pp. 4168–4180.
 80. Fakhra, A., Jin, X., and Zhao, L. (2020). Time gap estimation in naturalistic driving: A data-driven approach. *Transportation Research Record*, 2674(2), pp. 145–157.
 81. Sun, Y., He, Q., and Zhou, M. (2022). Development of robust vehicle controllers for connected automated vehicles. *Transportation Research Part C: Emerging Technologies*, 135, p. 103801.
 82. Li, H., Zhang, Y., and Zhao, Q. (2023). Evaluating energy consumption and emissions in mixed traffic flow with connected and automated vehicles. *IEEE Transactions on Intelligent Transportation*

- Systems*, 24(5), pp. 5289–5301.
83. Yan, H., Wang, P., and Zhao, L. (2020). Assessing traffic safety with naturalistic driving trajectory data. *Accident Analysis and Prevention*, 146, p. 105794.
 84. Zhou, C., He, Y., and Liu, K. (2023). Investigating traffic flow dynamics using naturalistic vehicle trajectories. *Transportation Research Part B: Methodological*, 174, p. 102785.
 85. Chen, W., Zhang, R., and Li, T. (2024). String stability in connected automated vehicle systems: Insights from real-world trajectory data. *Transportation Research Part C: Emerging Technologies*, 140, p. 104525.
 86. ZalaZone Test Facility (2025). Zalazone testing and track. Accessed: March 13, 2025.
 87. AstaZero Test Facility (2025). Astazero testing and research facility. Accessed: March 13, 2025.
 88. Jiang, R., Hu, M.-B., Zhang, H., Gao, Z., Jia, B., Wu, Q., Wang, B.-H., and Yang, M. (2014). Traffic experiment reveals the nature of car-following. *PLOS ONE*, 9(4), p. e94351.
 89. Jiang, R., Hu, M., Zhang, H., Gao, Z., Jia, B., and Wu, Q. (2015). On some experimental features of car-following behavior and how to model them. *Transportation Research Part B: Methodological*, 80, pp. 338–354.
 90. Jiang, R., Jin, C., Zhang, H., Huang, Y., Tian, J., Wang, W., Hu, M., Wang, H., and Jia, B. (2018). Experimental and empirical investigations of traffic flow instability. *Transportation Research Part C: Emerging Technologies*, 94, pp. 83–98.
 91. Huang, Y., Jiang, R., Zhang, H., Hu, M., Tian, J., Jia, B., and Gao, Z. (2018). Experimental study and modeling of car-following behavior under high speed situation. *Transportation Research Part C: Emerging Technologies*, 97, pp. 194–215.
 92. Zheng, S., Jiang, R., Tian, J., Zhang, H., Li, Z., Gao, L., and Jia, B. (2021). Experimental study on properties of lightly congested flow. *Transportation Research Part B: Methodological*, 149, pp. 1–19.
 93. Zheng, S., Jiang, R., Tian, J., Li, X., Jia, B., Gao, Z., and Yu, S. (2022). A comparison study on the growth pattern of traffic oscillations in car-following experiments. *Transportmetrica B: Transport Dynamics*, 1(11), pp. 706–724.
 94. Arman, M. A. and Tampère, C. M. J. (2025). Traffic flow crystallization method for trajectory approximation and lane change inference. *IEEE Transactions on Intelligent Transportation Systems*, 26(7), pp. 9305–9325.
 95. Vlaamse Overheid (2025). Rijstrooksignalisatie vlaanderen: dynamische snelheidsbeperkingen. <https://www.vlaanderen.be/datavindplaats/catalogus/rijstrooksignalisatie-vlaanderen-dynamische-snelheidsbeperkingen>. Accessed: 2025-07-23.
 96. Vlaams Verkeerscentrum (2025). Verkeerssituatie — vlaams verkeerscentrum. <https://www.verkeerscentrum.be/>. Accessed: 2025-07-23.
 97. Arman, M. A. and Tampère, C. M. J. (2024). Choice-based macroscopic lane-change prediction model for weaving areas. *Transportation Research Part C: Emerging Technologies*, 169.
 98. Arman, M. A. and Tampère, C. M. J. (2022). Lane-level trajectory reconstruction based on data-fusion. *Transportation Research Part C: Emerging Technologies*, 145.
 99. Peng, J., Shangguan, W., Arman, M. A., and Tampère, C. M. J. (2025). Empirical study of lane-changing patterns with dynamic discharge flow in a weaving area. In *Proceedings of the Transportation Research Board 104th Annual Meeting (TRB)*. Washington, D.C. Presented paper.
 100. Vlaamse Overheid (2025). Wegenregister segment. <https://www.vlaanderen.be/datavindplaats/catalogus/wegenregister-segment>. Accessed: 2025-07-23.
 101. Vlaamse Overheid (2025). Meten-in-vlaanderen: minuutwaarden verkeersmetingen. <https://www.vlaanderen.be/datavindplaats/catalogus/meten-in-vlaanderen-minuutwaarden-verkeersmetingen>. Accessed: 2025-07-23.
 102. Gloudemans, D., Zachár, G., Wang, Y., Ji, J., Nice, M., Bunting, M., Barbour, W. W., Sprinkle, J., Piccoli, B., Monache, M. L. D., et al. (2024). So you think you can track? In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4528–4538.
 103. Wang, Y., Gloudemans, D., Ji, J., Teoh, Z. N., Liu, L., Zachár, G., Barbour, W., and Work, D. (2024). Automatic vehicle trajectory data reconstruction at scale. *Transportation research part C: emerging technologies*, 160, p. 104520.
 104. Li, R., Zheng, Z., Ngoduy, D., and Li, L. (2025). Developing and validating an adaptive multi-layer vehicle trajectory reconstruction method for outlier removal. *Transportation Research Part C: Emerging Technologies*, 171, p. 104946.
 105. Li, R., Zheng, Z., Ni, D., and Li, L. (2024). A method for long car-following pair extraction and comprehensive data quality assessment: a case study using zen traffic data. *Transportation Letters*, pp. 1–20.
 106. Ammourah, R., Beigi, P., Fan, B., Hamdar, S. H., Hourdos, J., Hsiao, C.-C., James, R., Khajeh-Hosseini, M., Mahmassani, H. S., Monzer, D., et al. (2024). Introduction to the third generation simulation dataset: Data collection and trajectory extraction. *Transportation Research Record*, p. 03611981241257257.