

Large Car-following Data Based on Lyft level-5 Open Dataset: Following Autonomous Vehicles vs. Human-driven Vehicles

Guopeng Li^{1,*}, Yiru Jiao¹, Victor L. Knoop¹, Simeon C. Calvert¹, and J.W.C. van Lint¹

Abstract—Car-Following (CF), as a fundamental driving behaviour, has significant influences on the safety and efficiency of traffic flow. Investigating how human drivers react differently when following autonomous vs. human-driven vehicles (HV) is thus critical for mixed traffic flow. Research in this field can be expedited with trajectory datasets collected by Autonomous Vehicles (AVs). However, trajectories collected by AVs are noisy and not readily applicable for studying CF behaviour. This paper extracts and enhances two categories of CF data, HV-following-AV (H-A) and HV-following-HV (H-H), from the open Lyft level-5 dataset. First, CF pairs are selected based on specific rules. Next, the quality of raw data is assessed by anomaly analysis. Then, the raw CF data is corrected and enhanced via motion planning, Kalman filtering, and wavelet denoising. As a result, 29k+ H-A and 42k+ H-H car-following segments are obtained, with a total driving distance of 150k+ km. A diversity assessment shows that the processed data cover complete CF regimes for calibrating CF models. This open and ready-to-use dataset provides the opportunity to investigate the CF behaviours of following AVs vs. HVs from real-world data. It can further facilitate studies on exploring the impact of AVs on mixed urban traffic.

Index Terms—Car-following, trajectory dataset, autonomous vehicle, driving behaviour

I. INTRODUCTION

Autonomous vehicles (AVs) have been rapidly developing in recent years, bringing potential benefits such as enhancing traffic safety [1], reducing congestion [2], and increasing mobility accessibility [3]. However, the extent of these improvements remains unclear, which depends not only on the performance of AVs but also on human drivers' reactions to AVs. Clarifying the impact of AVs on human driving behaviour is thus crucial for safe and efficient integration of AVs into transportation systems [4,5].

Car-following (CF), which refers to one vehicle following another, is the most common driving behaviour. CF plays a critical role in maintaining smooth traffic flow and reducing congestion [6,7]. The presence of AVs may reshape CF behaviours and thus mixed traffic flow. How an AV follows its leading vehicle is determined by the specific driving algorithm, which is continuously being improved. Therefore, it is more important to examine how human-driven vehicles (HVs) react differently when following an AV vs. an HV.

In the literature, the influence of AVs on the CF behaviours of human drivers has mainly been studied through field experiments, driving simulators, and real-world AV datasets. In field experiments, participants are asked to follow a real

or seemingly-real AV in different scenarios [8–10]. Similar experiments can also be carried out in a virtual environment by using driving simulators [11]. Field tests and simulations are controllable so researchers can focus on specific points of interest. However, due to cost limitations, these two approaches cannot provide comprehensive and large data covering diverse scenarios.

Recently, the release of autonomous driving datasets, such as Waymo [12], nuScenes [13], and Lyft5 [14], has enabled researchers to study AVs' impacts on traffic with real-world data. Hu et al. [15] offer the first attempt to process a CF dataset from the Waymo dataset. However, because AVs are not marked in the entire dataset of Waymo, only 274 HV-following-AV (H-A) pairs and 1032 HV-following-HV (H-H) pairs are extracted. The limited amount of samples leads to contradictory findings. For example, Wen et al. [16] conclude that, compared with H-H, H-A has lower driving volatility, smaller time headways, and higher Time-to-Collision (TTC); while Hu et al. [17] found no significant difference between H-H and H-A, except for smaller spacing during congestion. To reduce the biases when using small datasets, a larger and comparative CF dataset balanced comprising both scenarios is indispensable.

In this paper, the Lyft level-5 open dataset is processed. We select, assess, and enhance 29k+ HV-following-AV pairs and 42k+ HV-following-HV pairs in similar environments. The dataset covers diverse CF regimes and the enhanced dataset provides smooth, ready-to-use motion information for CF model calibration/training. The contribution of this paper is double-fold. First, we propose a processing procedure and demonstrate its validity. Second, the processed data is openly shared as the first large CF dataset that allows comparing the behaviours of following AV vs. HV. It is expected to help better evaluate the influence of AVs in mixed traffic. The dataset is available at <https://github.com/RomainLITUD/Car-Following-Dataset-HV-vs-AV>, including detailed instructions on reading and filtering desired CF pairs.

II. LYFT LEVEL-5 DATASET

A. Dataset description

The Lyft level-5 dataset [14] is a large-scale dataset of high-resolution sensor data collected by a fleet of 20 self-driving cars. The dataset includes 1000+ hours of perception and motion data collected over a 4-month period from urban and suburban environments along a fixed route in Palo Alto, California. The route is shown in Fig. 1.

The motion prediction dataset comprises about 170,000 scenes, with each scene spanning approximately 25 s.

¹ Department of Transport & Planning, Delft University of Technology, the Netherlands

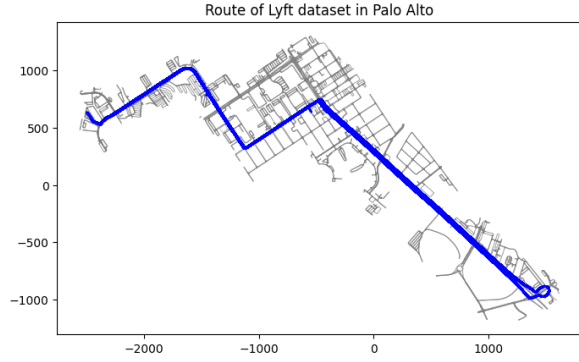


Fig. 1. The road maps of Palo Alto, California (in meters). The blue lines mark the fixed route in the Lyft level-5 dataset.

These scenes may be collected continuously or intermittently. For each scene, the track ids of agents are re-numbered (from 1). Each scene includes the movement states (e.g. position, yaw angle, size, speed) of perceived vehicles, cyclists, and pedestrians, as well as the position and orientation of the AV. Information about the driving environment, including high-definite maps and traffic light status, is also provided. The dataset is available from the website: <https://woven.toyota/en/prediction-dataset>. The material used in this paper includes the full training and validation datasets, the semantic map file, and the python toolkit l5kit (<https://woven-planet.github.io/l5kit/>) provided by Lyft developers.

B. Data Processing Framework

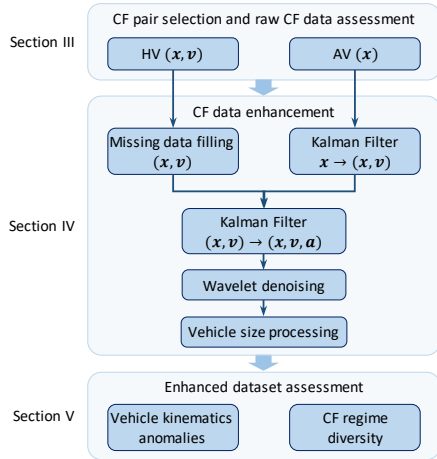


Fig. 2. Flowchart of the data processing.

The flowchart in Fig.2 presents the procedure of CF data selection, assessment, and enhancement. In section III, CF pairs and their raw trajectories are first selected from the unlabelled dataset based on certain rules. Next, the raw data quality is assessed from the perspective of anomaly analysis. In section IV, the raw data is enhanced. For AVs (only position x is given) and HVs (speed v and position x are

given), we use two different methods to fill in the missing segments and estimate smooth x and v . The acceleration a is estimated by the combination of the Kalman Filter and wavelet denoising methods. After the processing procedure, in section V, the enhanced data is assessed again, including both anomaly analysis and the diversity of CF regimes. In the following sections, we will introduce each step in detail.

III. CF PAIR SELECTION AND ASSESSMENT

A. CF pair selection

The first step is identifying CF pairs from unlabelled data. Previous studies about the rules of extracting CF pairs have been reviewed by Wen et al. [16]. We refer the readers to this recent paper for more details.

Considering the large size of the dataset (1000+ hours), we make two groups of rules and propose a two-step procedure. The first step is quick screening. The entire dataset is scanned second by second (every 1-5 seconds) based on the first group of rules to roughly identify possible CF events. Next, the second group of rules rigorously check each CF event frame by frame. The rules are listed in Table.I. The extracted CF pairs are categorized based on the type of the leading vehicle (AV or HV). Statistics of the CF dataset are presented in Table.II. The total duration of this CF dataset spans over 460+ hours, covering a total distance of 15,000+ km.

TABLE I
CF EVENT SELECTION RULES

No.	Description
Group-1: second-to-second	
1.1	The probability of both vehicles being passenger cars > 0.95 .
1.2	The longitudinal distance between the vehicles < 85 m.
1.3	The lateral distance between the vehicles < 1.75 m.
1.4	No obstacles or other road agents interrupt the CF pairs.
1.5	Both vehicles are on the same side of the traffic lights.
1.6	Both vehicles are on a straight road.
1.7	Rules 1.1-1.6 must hold with a duration > 16 s.
Group-2: frame-to-frame	
2.1	The deviation of yaw for both vehicles < 0.035 rad.
2.2	The maximum yaw to the driving direction for both vehicles < 0.087 rad.
2.3	The interval between all adjacent timestamps < 0.42 s.
2.4	The distance between all adjacent positions < 5 m.
2.5	The average speed of both vehicles > 1 m s ⁻¹ .

TABLE II
EXTRACTED CF PAIRS FROM LYFT DATASET

Dataset	Nb pairs	Distance(km)	Duration(h)
H-A	29449	6996	220
H-H	42892	8125	240

For each CF pair, the initial position of the leading vehicle is set as the origin. The straight road lane defines the driving direction. Because the angle between the yaw of the vehicles and the lanes is small (as stated in Rules 2.1 and 2.2), it is reasonable to assume that the lateral movement is

independent of the car-following behaviours. Therefore, this paper will only focus on the longitudinal movement going forward.

B. Raw CF data assessment

Calibrating CF models requires highly accurate and consistent position, speed, and acceleration data. It is necessary to first assess the quality of the raw CF data. For HV (either in H-A or H-H dataset), we assess (1) anomalies that violate constraints of vehicle kinematics, and (2) missing data. For AV, because only x is given, the anomalies of vehicle kinematics are assessed only. According to Punzo et al. [18], 3 constraints on acceleration a and jerk j must be satisfied:

- $a \in [-8, 5] \text{ m/s}^2$
- $j \in [-15, 15] \text{ m/s}^3$
- The jerk's sign cannot be inverse more than once in 1 s (denoted as *Jerk Sign Inversion*, JSI).

For HV, a and j are directly computed by the 1st and the 2nd-order differentiation of v (v -based). While for AV, a and j are derived from the 2nd and the 3rd-order differentiation of x (x -based). Violating one of the 3 constraints is identified as an anomaly. The result is shown in Table III, which is also compared with the CF data in the Waymo dataset (evaluated in [15]). The v -based jerk and JSI anomaly proportion for HVs in Lyft are on-par with the Waymo dataset. For x -based AVs, Lyft shows higher quality than Waymo (lower anomaly proportion).

TABLE III
ANOMALY ASSESSMENT OF RAW CF DATA

	Lyft		Waymo	
	AV (x -based)	HV (v -based)	x -based	v -based
Statistics				
Anomaly acc (%)	0.004	0.226	—	—
Anomaly jerk (%)	1.007	0.656	5.3	0.439
Anomaly JSI (%)	34.2	25.4	86.1	37.2

Next, abnormal segments are further investigated. Fig. 3 shows an example that compares the given speed and x -based speed of an HV. The result demonstrates that the speed provided by Lyft is smoothed from the measurements of position. However, there also exist some errors. At the beginning and the end of each scene (25 s duration), there is a 0-speed value, which seems to have been added artificially and was not excluded in smoothing. We guess that perhaps these 0-values are used to assist in segmenting scenes (for deep-learning purposes) but for some reason not removed. Therefore, the first 0.5 to 1.5 s data in these scenes are unreliable.

The assessment of the raw CF data is summarized as follows:

- The given position data of AVs is of high quality.
- Data are missing at the beginning and end of each scene.
- For both AVs and HVs, acceleration and jerk derived by differentiation is not smooth enough for calibrating CF models.

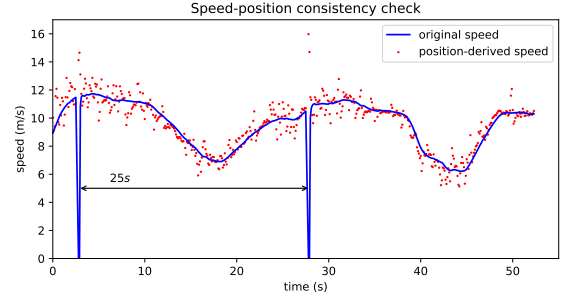


Fig. 3. Comparison between the speed data and the position-derived speed.

In the next section, the raw data will be enhanced to address these problems.

IV. CF DATA ENHANCEMENT

A. Missing data filling

Before further processing raw HV data, the missing part around the 0-speed points must be filled in first. In this study, we remove the speed and position data within 1.5 s of the last 0-value timestamp. This segment of motion is estimated by the polynomial of degree seven (7-DOP) jerk-minimization method. The position of the vehicle is assumed to have the form of a polynomial of degree seven with 8 unknown coefficients:

$$x(t) = \sum_{i=0}^7 p_i t^i \quad (1)$$

The initial and final positions, speeds, and accelerations (derived from speed) pose 6 boundary conditions (constraints). The objective function to minimize is the jerk in the duration T :

$$J = \int_0^T [x'''(t)]^2 dt \quad (2)$$

which is typical quadratic programming that can be easily solved. Notice that only the position and speed are estimated. The acceleration data will be derived later.

One example of the estimation results is shown in Fig. 4. The time interval is not always uniform. Clearly, the 0-value influences the given speed at around 3.0 s (the red line). The removed segment is filled in by the black-star curves (the timestamp series are not changed), which is a 6-degree polynomial. We see that the estimated speed profile is smoother due to the vehicle's kinematic constraints.

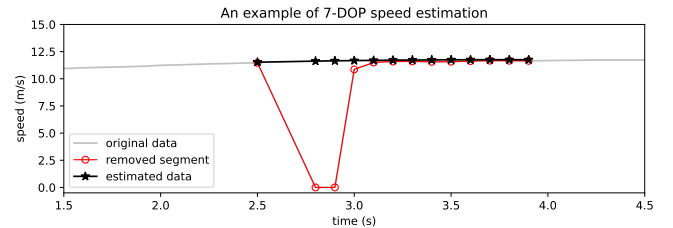


Fig. 4. Use the 7-DOP method to fill in the incorrect data segment.

B. Kalman filtering for speed estimation

Because the speed of AVs is not provided, to facilitate the following acceleration estimation and smoothing steps, we need to derive the speed of AVs from position data first. The well-known Kalman Filter (KF) [19] is used. In this Kalman Filter, we employ the constant-speed model. For each time interval (mostly $\Delta t = 0.1$ s), the state transition equation is:

$$\begin{bmatrix} x(t_{i+1}) \\ v(t_{i+1}) \end{bmatrix} = \begin{bmatrix} 1 & \Delta t_i \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x(t_i) \\ v(t_i) \end{bmatrix} \quad (3)$$

The given position and its differentiation are regarded as the measurement of x and v . The process covariance matrix \mathbf{Q}_1 and the measurement covariance matrix \mathbf{R}_1 control the trade-off between accuracy and smoothness. Considering the range of acceleration in Eq.III-B, their values are set as follows by trial and error:

$$\mathbf{Q}_1 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.8 \end{bmatrix}^2, \quad \mathbf{R}_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1.1 \end{bmatrix}^2 \quad (4)$$

The next step is estimating the acceleration. In principle, this can also be accomplished by a Kalman filter that gives both speed and acceleration. However, tuning the process and measurement covariance is difficult and usually the acceleration covariance changes with time for a CF event. Therefore, we propose a 2-step method. First, continuing in this subsection, a second KF is applied for both AVs and HVs to derive an over-smoothed acceleration. Next in subsection IV-C, wavelet denoising will be used to smooth the v -based (differentiation) acceleration. For the second KF, the state transition equation is:

$$\begin{bmatrix} x(t_{i+1}) \\ v(t_{i+1}) \\ a(t_{i+1}) \end{bmatrix} = \begin{bmatrix} 1 & \Delta t_i & \frac{1}{2}a(t_i)(\Delta t_i)^2 \\ 0 & 1 & a(t_i)\Delta t_i \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x(t_i) \\ v(t_i) \\ a(t_i) \end{bmatrix} \quad (5)$$

The \mathbf{Q}_2 and \mathbf{R}_2 are set as follows. We set a high value of measurement error for acceleration to over-smooth it.

$$\mathbf{Q}_2 = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 1.5 \end{bmatrix}^2, \quad \mathbf{R}_2 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10 \end{bmatrix}^2 \quad (6)$$

Denote the v -based acceleration as a_v and the over-smoothed acceleration given by KF as a_k , their RMSE, σ_a , is regarded as an approximation of the total noise. This estimated hyper-parameter will be used in wavelet denoising to further smooth acceleration.

C. Wavelet denoising for acceleration smoothing

Wavelet denoising [20] is a robust technique to remove noise from signals by transforming the time series into wavelet space and then thresholding the high-frequency coefficients while preserving the important information in the low-frequency coefficients. Compared with KF, wavelet denoising is non-parametric. It is effective for removing different types of noise, including Gaussian noise, impulsive noise, and mixed noise. Especially, wavelet denoising can adapt to non-stationary noise characteristics, which meets the requirements of acceleration estimation and smoothing.

In this study, we use the wavelet denoising tool in `skimage` python package. The noise standard deviation is set as σ_a from the result of KF. The type of wavelet is the Daubechies family with 6 vanishing moments ('db6'), which can effectively capture both short-term and long-term features. A soft threshold method and up to 4 wavelet decomposition levels are used.

Fig.5 presents a 1 min H-A car-following example. The silver lines are raw position, speed, v -based acceleration and jerk of the following HV. This CF pair approaches an intersection, decelerates, stops and waits, and then accelerates to a desired speed. The missing segments at 24s and at 49s are filled in. Processed acceleration and jerk profiles are significantly smoother.

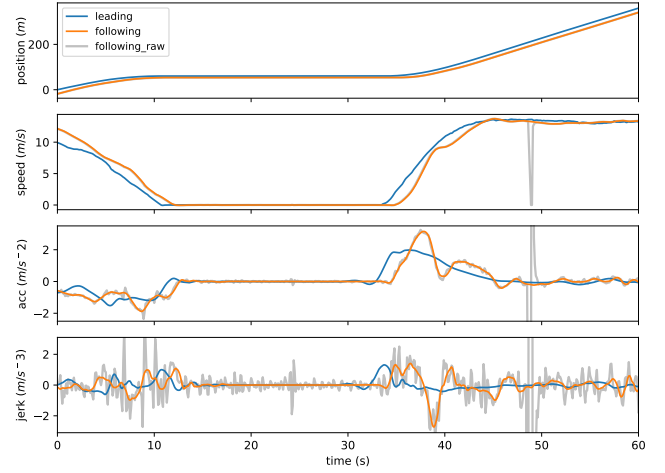


Fig. 5. Comparison between the raw data and the processed data in a 60s CF event.

D. Vehicle size processing

Besides motion information, the length of the vehicle is also important, especially for calculating some safety and efficiency metrics, such as bump-to-rear gap and TTC. The size of the AVs is a given, fixed value (4.87 m length, 1.85 m width). However, the size of perceived HVs is not always stable. It varies with time due to perception errors (such as shading). Many datasets, including the Waymo CF data [15], ignore this step.

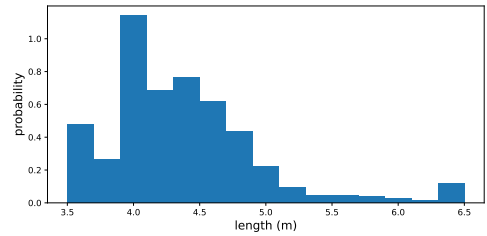


Fig. 6. Distribution of HV lengths.

Because we excluded vehicles that are not passenger cars when selecting CF pairs, we set a maximum length of 6.5 m

(pick-up truck) and a minimum length of 4 m (small vehicle). The following rules are used to estimate the size. (1) For each HV, if the variance of the length time series < 0.3 m, then we choose the mean value. (2) Otherwise, perceived length series are clamped between 3.5 m and 6.5 m, then we choose the percentile at 0.95 (95%). The distribution of estimated HV lengths is shown in Fig. 6. The mean value is 4.38 m, which is close to the average length of passenger cars in the US [21].

By now, the entire data enhancement procedure is finished. Next, we will assess the quality of enhanced data.

V. ENHANCED DATASET EVALUATION

The enhanced dataset is split into two groups, following AV (denoted as H-A) vs. following HV (denoted as H-H). They will be assessed based on anomaly analysis and diversity evaluation. We will show that the enhanced trajectories have higher quantity and the dataset covers diverse regimes for calibrating CF models.

A. Vehicle kinematics anomalies

For anomaly detection, we use the same rules as stated in section III-B. The result is shown in Table IV. After processing, the anomaly percentage is significantly reduced, especially abnormal jerk sign inversion. The enhanced data is smoother and better conforms to the vehicle's kinematic constraints.

TABLE IV
ANOMALY ASSESSMENT OF THE ENHANCED DATASET

Statistics	H-A pairs	H-H pairs
Anomaly acc (%)	0.0082	0.0234
Anomaly jerk (%)	0.0039	0.0186
Anomaly jerk sign inversion (%)	0.455	0.454

B. CF regime diversity

Next, we will evaluate the regime diversity in both H-H and H-A subsets. Here a regime refers to a driving situation experienced by the following vehicle (usually restricted by its leader). A higher diversity of regimes means that the dataset is not just 'big' but also informative. For some CF models, e.g. Intelligent Driver Models (IDM), certain parameters cannot be calibrated if some specific regimes are missing in the data [22]. Insufficient regime diversity in CF data may also result in unrepresentative or over-fitted models [23], and thus contradictory conclusions about driving behaviours.

To evaluate the regime diversity in the enhanced dataset, we adopt the identification algorithm proposed in [24]. The algorithm consists of three steps: 1) segmenting the follower's speed profile into various sections, 2) categorizing the sections into car-following (CF) and free-following (FF), and 3) determining regimes based on the acceleration within these sections. For the second step, a threshold is selected based on the mean and variance of the distribution of time gaps that are calibrated from **Newell's car-following model** [25] (we refer the readers to [24] for more details). Given

that the followers in H-H and H-A may exhibit different time gap distributions, we perform two separate threshold selection. Fig. 7 compares the distributions of time gaps for H-H and H-A, where a clear difference is shown between following an HV vs. an AV.

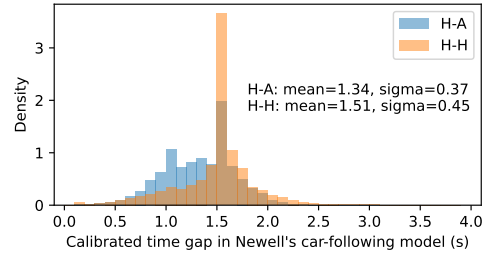


Fig. 7. Distribution comparison between H-H and H-A time gaps calibrated from Newell's car-following model.

Sharma et al. [24] classify regimes based on the previous studies by Treiber et al. [22], where standstill and deceleration are excluded from free-flowing. This consideration is reasonable for highways because the traffic flow on highways is highly continuous. However, vehicles face more interruptions in urban environments, such as vulnerable road agents (cyclists, pedestrians) and traffic signals. Therefore, this paper considers the following 7 regimes: free acceleration (**Fa**), free deceleration not caused by the leading vehicle (**Fd**), cruising at a desired speed (**C**), acceleration following a leading vehicle (**A**), deceleration following a leading vehicle (**D**), constant speed following (**F**), and standstill (**S**).

Fig. 8 presents the proportion of accumulated duration of the 7 regimes in H-A and H-H subsets. The regimes F, D, and A constitute around 74% and 70% of the total duration in H-A and H-H, respectively. This suggests that the followers' behaviours depend on their leaders most of the time. This dataset is therefore suitable for studying how followers react differently to the leading AV vs. HV.

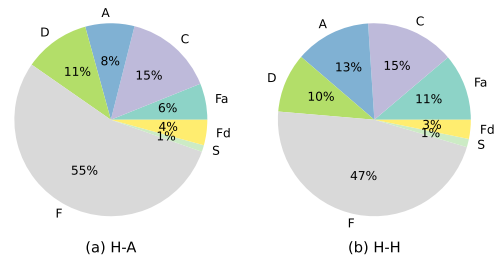


Fig. 8. Time proportion of car-following regimes.

According to Sharma et al. [23], to calibrate all parameters in IDM, at least 3 regimes, A, D, and F, must be included in a CF event. Based on this principle, we categorize all CF pairs into two groups. One is 'ADF+n', which means all ADF regimes are included and there are n extra regimes. The other is 'others', which means at least one of ADF regimes is missing (no matter how many regimes they have). The results in Fig. 9 show that 64% and 56% CF pairs fall in the

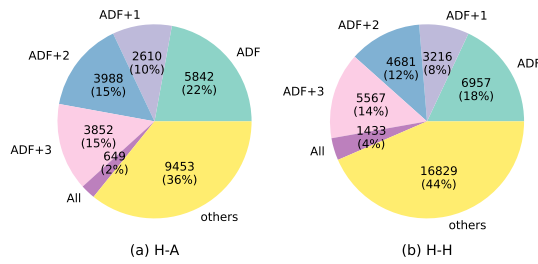


Fig. 9. Regime diversity of car-following pairs.

ADF+ n group in H-A and H-H subset, respectively. These pairs support calibrating IDMs. Meanwhile, we would like to emphasize that ‘others’ do not mean that these CF pairs are useless. They are not suitable for calibrating IDM but they are still informative for calibrating other more complex car-following models, e.g. deep-neural-networks-based methods.

In summary, this section shows that the enhanced dataset has fewer anomalies and high car-following regime diversity.

VI. CONCLUSION

This paper proposes a car-following trajectory data processing procedure. This procedure has been applied to an openly available dataset and validated by anomaly analysis and regime assessment. The Lyft level-5 dataset, which contains information about both autonomous vehicles and human-driven vehicles, has been processed with this technique and the enhanced car-following trajectories are publicly available. The initial dataset now is processed into a high-quality, ready-to-use dataset. It contains human drivers following autonomous as well as human-driven vehicles in diverse scenarios. The processing procedure is essential for doing further analysis. The published enhanced car-following dataset is expected to help researchers better understand the impact of AVs on traffic flow and to develop safer and more effective AV systems in the future.

ACKNOWLEDGMENT

This research is sponsored by the NWO/TTW project MiRRORS with grant agreement number 16270.

REFERENCES

- [1] Z. Yao, R. Hu, Y. Jiang, and T. Xu, “Stability and safety evaluation of mixed traffic flow with connected automated vehicles on expressways,” *Journal of Safety Research*, vol. 75, pp. 262–274, 2020.
- [2] F. Duarte and C. Ratti, “The impact of autonomous vehicles on cities: A review,” *Journal of Urban Technology*, vol. 25, no. 4, pp. 3–18, 2018.
- [3] J. Meyer, H. Becker, P. M. Bösch, and K. W. Axhausen, “Autonomous vehicles: The next jump in accessibilities?,” *Research in transportation economics*, vol. 62, pp. 80–91, 2017.
- [4] R. Hoogendoorn, B. van Arem, and S. Hoogendoorn, “Automated driving, traffic flow efficiency, and human factors,” *Transportation Research Record*, vol. 2422, no. 1, pp. 113–120, 2014.
- [5] S. Calvert, W. Schakel, and J. Van Lint, “Will automated vehicles negatively impact traffic flow?,” *Journal of advanced transportation*, vol. 2017, 2017.
- [6] J. Sun, Z. Zheng, and J. Sun, “Stability analysis methods and their applicability to car-following models in conventional and connected environments,” *Transportation research part B: methodological*, vol. 109, pp. 212–237, 2018.

- [7] Y. Shiomi, G. Li, and V. L. Knoop, “Analysis of stochasticity and heterogeneity of car-following behavior based on data-driven modeling,” *Transportation Research Record*, p. 036119812311692, May 2023.
- [8] Y. Rahmati, M. K. Hosseini, A. Talebpour, B. Swain, and C. Nelson, “Influence of autonomous vehicles on car-following behavior of human drivers,” *Transportation Research Record*, vol. 2673, no. 12, pp. 367–379, 2019.
- [9] I. Mahdinia, A. Mohammadnazar, R. Arvin, and A. J. Khattak, “Integration of automated vehicles in mixed traffic: Evaluating changes in performance of following human-driven vehicles,” *Accident Analysis & Prevention*, vol. 152, p. 106006, 2021.
- [10] S. Soni, N. Reddy, A. Tsapi, B. van Arem, and H. Farah, “Behavioral adaptations of human drivers interacting with automated vehicles,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 86, pp. 48–64, 2022.
- [11] N. Reddy, S. P. Hoogendoorn, and H. Farah, “How do the recognizability and driving styles of automated vehicles affect human drivers’ gap acceptance at t-intersections?,” *Transportation research part F: traffic psychology and behaviour*, vol. 90, pp. 451–465, 2022.
- [12] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [13] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [14] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, “One thousand and one hours: Self-driving motion prediction dataset,” in *Proceedings of the 2020 Conference on Robot Learning* (J. Kober, F. Ramos, and C. Tomlin, eds.), vol. 155, pp. 409–418, Nov 2021.
- [15] X. Hu, Z. Zheng, D. Chen, X. Zhang, and J. Sun, “Processing, assessing, and enhancing the waymo autonomous vehicle open dataset for driving behavior research,” *Transportation Research Part C: Emerging Technologies*, vol. 134, p. 103490, 2022.
- [16] X. Wen, Z. Cui, and S. Jian, “Characterizing car-following behaviors of human drivers when following automated vehicles using the real-world dataset,” *Accident Analysis & Prevention*, vol. 172, p. 106689, 2022.
- [17] X. Hu, Z. Zheng, D. Chen, and J. Sun, “Autonomous vehicle’s impact on traffic: Empirical evidence from waymo open dataset and implications from modelling,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2023.
- [18] V. Punzo, M. T. Borzacchiello, and B. Ciuffo, “On the assessment of vehicle trajectory data accuracy and application to the next generation simulation (ngsim) program data,” *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1243–1262, 2011.
- [19] G. Welch, G. Bishop, *et al.*, “An introduction to the kalman filter,” 1995.
- [20] Q. Pan, L. Zhang, G. Dai, and H. Zhang, “Two denoising methods by wavelet transform,” *IEEE transactions on signal processing*, vol. 47, no. 12, pp. 3401–3406, 1999.
- [21] G. Tiwari, J. Fazio, and S. Pavitras, “Passenger car units for heterogeneous traffic using a modified density method,” in *Proceedings of fourth international symposium on highway capacity*, pp. 246–257, 2000.
- [22] M. Treiber and A. Kesting, “Microscopic calibration and validation of car-following models – a systematic approach,” *Procedia - Social and Behavioral Sciences*, vol. 80, pp. 922–939, 2013.
- [23] A. Sharma, Z. Zheng, and A. Bhaskar, “Is more always better? the impact of vehicular trajectory completeness on car-following model calibration and validation,” *Transportation Research Part B: Methodological*, vol. 120, pp. 49–75, 2019.
- [24] A. Sharma, Z. Zheng, and A. Bhaskar, “A pattern recognition algorithm for assessing trajectory completeness,” *Transportation Research Part C: Emerging Technologies*, vol. 96, pp. 432–457, 2018.
- [25] G. Newell, “A simplified car-following theory: a lower order model,” *Transportation Research Part B: Methodological*, vol. 36, no. 3, pp. 195–205, 2002.