Data-driven analysis of ACC car-following time gaps for different driving contexts

Master's thesis in Electrical Engineering

RUOHAN LI

YUSHENG YANG

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Data-driven analysis of ACC car following time gaps for different driving contexts

RUOHAN LI
YUSHENG YANG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Systems and Control
Chalmers University of Technology
Gothenburg, Sweden 2025

Data-driven analysis of ACC car following time gaps for different driving contexts
RUOHAN LI
YUSHENG YANG

Supervisor: Jie Zhu, Volvo Cars; Fangting Zhou, Department of Architecture and Civil Engineering
Examiner: Balázs Kulcsár, Department of Electrical Engineering

Master's Thesis 2025
Department of Electrical Engineering
Systems and Control
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Ruohan Li
Yusheng Yang
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Adaptive Cruise Control (ACC) is a key component of the Advanced Driver Assistance System which automatically adjusts the vehicle speed to maintain a safe distance from vehicles ahead. While current ACC strategies mainly take vehicle speed as the dominant factor, human drivers adapt their time headway (THW) according to a wider range of contextual factors, including infrastructure, traffic density, lead vehicle type, and environmental conditions. This thesis adopts a data-driven approach to analyze how such factors influence car-following time gaps. Using public trajectory datasets and the Volvo Cars sample dataset, we first extract steady car-following states and apply regression models with permutation importance to rank influential factors. Detailed analyses are then conducted on lead vehicle type, lane selection, and lighting conditions, supported by statistical comparisons across traffic groups. In addition, scenario-based studies examine interactions of multiple factors, such as cut-in intentions and on-ramp infrastructure, using clustering methods (Hidden Markov Models and K-means) and sensor signal filtering. The results show that contextual conditions affect human time gap preferences in predictable ways. These insights provide recommendations for more adaptive and human-like ACC time gap strategies, with potential benefits for driving comfort and user experience.

Adaptive Cruise Control, Car Following, Time Gap Control, Human Driving Behavior, Data-driven Analysis

Acknowledgments

We would like to express our sincere gratitude to our supervisors, Dr. Jie Zhu from Volvo Cars and Dr. Fangting Zhou from Chalmers, and our examiner, Professor Balázs Kulcsár, for their invaluable guidance, continuous support, and constructive feedback throughtout the project.
We are also deeply grateful to Viktor Wänerlöv for providing essential support with the dataset from Volvo Cars, to Dr. Pengcheng Wu for sharing valuable insights on research methods, and Sten Elling Tingstad Jacobsen for offering his insights on scenarios that greatly enhanced our work.
We would also like to thank our colleagues at Volvo Cars who provide us with valuable research insights.
The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of Volvo Cars.

Ruohan Li and Yusheng Yang, Göteborg, August 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ACC	Adaptive Cruise Control
ADAS	Advanced Driver Assistance Systems
CC	Cruise Control
DHW	Distance Headway
HDI	Highest-Density Intervals
HMM	Hidden Markov Model
ID	Identifier
IIR	Infinite Impulse Response
KDE	Kernel Density Estimation
K-S	Kolmogorov-Smirnov
MSE	Mean Squared Error
PCA	Principal Component Analysis
RQ	Research Question
R^2	Coefficient of Determination
STD	Standard Deviation
THW	Time Headway
XGBoost	Extreme Gradient Boosting

Contents

List of Figures

List of Tables

1
Introduction

Adaptive Cruise Control (ACC) is an advanced driver assistance system that automatically adjusts a vehicle's speed to maintain either a driver-selected cruising speed or a safe following distance from the vehicle ahead by automatically controlling the throttle and the brake [1]–[3]. It is believed to enhance driving safety, traffic efficiency, and comfort by reducing driver workload and mitigating human limitations in the driving process [4]–[6]. One of the most critical parameters affecting ACC performance is the car-following time gap between vehicles. A short time gap may increase the risk of rear-end collisions and cause more frequent, abrupt speed oscillations, while a long time gap can lead to driver dissatisfaction and reduce road capacity. Therefore, determining an appropriate following time gap is crucial for the development of advanced ACC systems.

The speed of the vehicle is the primary factor considered in the current ACC function. This may cover the major scenarios, while other factors can affect the car-following time gap. Human drivers continuously adjust car following distance not only based on velocity, but also according to a variety of contextual factors such as visibility, traffic density, and the behavior of surrounding vehicles. For example, human drivers may maintain a longer time gap at night than during the day due to reduced visibility, allowing more time to react to potential threats [7]. Similarly, the presence of a lead truck may influence human drivers' perception of risk and choice of car-following distance [8]. Including these contextual factors can help ACC functions operate in a more natural, intelligent, and context-aware manner, making them better aligned with human driving behavior.

In this study, contextual factors such as visibility and infrastructure are systematically incorporated into the car-following time gap analysis. The findings are cross-validated with both a public dataset and a Volvo dataset, and critical scenarios are identified. These results offer implications not only for the development of ACC functions and autonomous driving functions but also for the advancement of traffic simulation and analysis.

1.1 Background

ACC is one of the earliest and most widely deployed Advanced Driver Assistance Systems (ADAS). It has been 30 years for the first generation of ACC equipped vehicles to be on the market, and it has been 22 years since the first ISO standard of ACC systems was produced [2]. ACC systems are an extension of conventional Cruise Control (CC) systems that adjust vehicle velocity and provide a specified

distance to the preceding vehicle by automatically controlling the throttle and the brake. However, the application scenarios of CC are limited because the vehicle speed cannot change according to local traffic. In order to solve this problem, ACC system has been developed. A typical ACC system has two function modes [9]. The first mode is traditional CC mode which is activated when there is no preceding vehicle or the preceding vehicle is far away. This mode is also known as speed control mode. The second mode is the distance control mode. It is activated when the ego vehicle needs to keep the inter-distance with the preceding vehicle to avoid a collision. In this mode the ego vehicle can automatically adapt its speed when a preceding vehicle appears, so the application scenario of CC is largely expanded [10].
Time gap (or time headway) is a pivotal control parameter in ACC systems, as it governs both traffic flow efficiency and safety. Excessively short time gaps may compromise driver safety and string stability, whereas overly long gaps can lead to reduced traffic capacity and efficiency [11].
Current ACC systems mainly rely on vehicle speed and relative distance to determine the time gap. While this approach captures the fundamental aspects of car-following, human drivers adapt their time gaps to different contextual conditions. In practice, drivers often adjust their following distance depending on factors such as road type, visibility, and the presence of other vehicles. This study, therefore, investigates how human drivers adapt time gaps under varying scenarios and contextual factors, and aims to provide recommendations for more adaptive and human-like ACC functions.

1.2 Related work

Research on car-following has a long history, and numerous models have been proposed to describe how vehicles follow one another on a road. Early approaches were predominantly theory-based, relying on mathematical formulations from traffic flow theory. Among them, kinematic models emphasized the physical aspects of vehicle motion by considering parameters such as velocity, headway, acceleration, and relative distance. Two well-known examples are Gipps' model [12] and the Intelligent Driver Model [13]. They remain widely used because of their ability to reproduce macroscopic traffic phenomena.
Psycho-physical models were introduced to account for the driver's perception and reaction processes. These models incorporate thresholds of perception and response times, thereby capturing more realistic driver behavior in response to surrounding vehicles. In parallel, models explicitly tailored for ACC emerged, focusing on the mechanistic aspects of longitudinal control, such as throttle and braking, to ensure safety and comfort [14].
With the increasing availability of trajectory data, data-driven car-following models have gained prominence. Leveraging machine learning and statistical methods, these models can capture heterogeneous traffic behavior and complex interactions. For example, Khodayari et al. [15] proposed a neural-network-based model incorporating driver reaction delay, while Kamjoo et al. [16] developed models for winter maintenance scenarios involving snowplows, showing that accounting for driver heterogeneity improves predictive performance. Recent reviews highlight that such

approaches outperform traditional models under complex traffic and mixed environments, including lane-free and pedestrian interactions [17], [18].
Despite these advances, most car-following models (theory-based, psycho-physical, or data-driven) primarily emphasize speed, distance, and headway as the key determinants of following behavior. Contextual factors, such as road type, visibility, traffic density, and other road users, are rarely incorporated in a systematic way. This limitation motivates the present study, which investigates how human drivers adapt their time gaps under different contextual conditions and provides insights for the development of more adaptive and human-like ACC systems.

1.3 Purpose

To address the challenges above, this study seeks to answer the following research questions (RQs):

RQ1: What contextual factors influence the car-following time gap maintained by human drivers?

RQ2: How do these contextual factors influence the car-following time gap maintained by human drivers?

RQ3: In which typical driving scenarios do multiple contextual factors combine to significantly affect the car-following time gap, prompting drivers to adapt their following distance?

RQ4: How may the ACC time gap strategy adapt to these factors?

In summary, existing car-following models have greatly advanced our understanding of vehicle interactions. The way human drivers adapt their time gaps to different scenarios remains insufficiently understood. Addressing this research gap is crucial for developing ACC systems that are adaptive and human-like.
In this research, we identify contextual factors that influence human drivers' choice of car-following time gaps, quantify how time gaps vary across different conditions, and highlight key scenarios where multiple factors interact. By systematically analyzing these effects with data-driven methods, the study provides new insights into driver behavior and offers recommendations for developing more adaptive and human-centric ACC functions.

1.4 Scope

In this project, we focus on the analysis of car-following time gaps within ACC application scenarios. Specifically, the study considers steady highway car-following situations where (1) a lead vehicle is present, (2) both vehicles travel on a highway, (3) the ego vehicle maintains a consistent following pattern, and (4) no lane-change is performed. Within this scope, we identify contextual factors that may influence drivers' choice of time gaps, quantify their effects, and highlight scenarios where multiple factors interact.
This study specifically focuses on the car-following time gap within ACC application scenarios under steady car-following conditions. This implies that the primary focus is on the time gap maintained when a vehicle is steadily following a preceding vehicle,

without abrupt speed or gap adjustments, and in the absence of significant traffic disruptions. Scenarios involving urban driving, lane-free, extreme weather, abrupt braking in response to the preceding vehicle, or reactions to a vehicle already having cut into the lane in front of the ego vehicle, do not fall within the scope of this study.

1.5 Thesis outline

This thesis is organized into four main chapters. Chapter 2 (Theory) introduces the theoretical background relevant to this research, including regression model, clustering model, random forest classifier and filters. Chapter 3 (Implementation) describes the methodology, detailing the dataset, preprocessing steps, factor selection, and the data-driven modeling approach. Chapter 4 (Results) presents the main findings of the study, with a focus on how different contextual factors influence car-following time gaps and the research on critical scenarios. Specifically, this chapter investigates:

- **Ranking and Individual Factor Influence:** A comprehensive importance ranking of individual contextual factors influencing the car-following time gap.
- **Detailed Individual Factor Analysis:** In-depth analysis of specific individual contextual factors, including: (1) Lead vehicle type; (2) Number of lanes and lane selection; (3) Lighting conditions.
- **Scenario-Based Studies:** Investigation of car-following time gaps in specific scenarios, focusing on the combined influence of contextual factors: (Scenario 1): Cut-in intention of other vehicles; (Scenario 2): Passing on-ramp area without merging vehicles; (Scenario 3): Passing on-ramp area with merging vehicles

Finally, Chapter 5 (Conclusion) summarizes the key results, discusses their implications for ACC development, and outlines the limitations of this work along with directions for future research.

1.6 Delimitation

This study is subject to several delimitation. The distinct characteristics of the two datasets used in this study (the public dataset and the proprietary Volvo Cars sample dataset) may each influence the results and should be taken into account.

- **Public Dataset :** A key challenge with the public dataset is the relatively short duration of continuous car-following events. This limits the analysis of long-term behavioral patterns and potential driver adaptation over extended periods. Furthermore, there is no information on whether a vehicle's ACC function was active or not. This is a significant point of discussion, as our study's fundamental assumption is that all observed car-following behavior is human-driven. The increasing ACC penetration rate in the vehicle fleet means that a portion of the data might reflect automated rather than human behavior, potentially influencing the statistical distributions we observed. The dataset is also restricted to specific German motorways, which, while providing

a rich source of high-quality data, may not fully reflect the comprehensive driving scenarios and cultures found across the world.

- **Volvo Cars sample dataset**: The Volvo Cars sample dataset offers a different perspective, capturing longer car-following durations and a wider range of infrastructure. However, unlike the fully curated public dataset, the raw data in Volvo Cars sample dataset requires extensive preprocessing and filtering. These steps may introduce subtle biases into the analyzed results. In addition, the dataset is collected with test drivers rather than in fully naturalistic settings, this recorded behavior may not perfectly represent everyday driving patterns.
- **Common Limitations**: A significant limitation of both datasets is the absence of information on individual driver characteristics or driving style, as noted by Karrouchi et al. [19]. Therefore, our analysis focuses on the collective driving patterns. As this study focuses on cars, the findings may not be directly applicable to other vehicle classes, such as trucks or vans. For the lead vehicle type study, we only investigate the influence of whether the lead vehicle is a car or a truck due to the lack of data for other vehicle type, such as motorcycles and vans, other than trucks.

Beyond the dataset-specific characteristics, our study was also subject to certain methodological limitations: Our in-depth scenario study, focusing on cut-in intention and on-ramp interactions, was primarily conducted and verified on the public dataset. A full cross-validation of these scenario-specific findings on the Volvo Cars sample dataset was not performed. This means that while our conclusions are robust for the public data, the human preference car-following time gap adaption in the Volvo Cars sample dataset remains a subject for future investigation.

The selection of the contextual factors for our analysis was necessarily restricted to the factors available in the dataset. There may be other significant contextual factors, such as weather conditions, road geometry, or traffic events that are not captured in the datasets, which could also play a crucial role in influencing a driver's choice of time headway.

2
Theory

Car-following behavior influenced by contextual factors is inherently complex, which requires a combination of complementary methods to analyze. To address the research questions introduced in Chapter 1, this section employs several analytical approaches. Regression models are used to estimate the impact of contextual factors on time headway, while clustering techniques help uncover latent patterns in driving behavior. Filtering methods are applied to improve the quality of time-series data, and a quantitative analysis of the shortest interval is conducted to characterize the predominant driving performance. Together, these methods provide a robust framework for understanding and modeling car-following gaps under different driving contexts.

2.1 Regression model

In this study, regression models are used to analyze how contextual factors influence the car-following time gap. As widely used and powerful tools in machine learning, regression models are suitable for discovering the relationships between a dependent variable and multiple independent variables. They make predictions and reveal interpretable insights into feature contributions.
In particular, Gradient Boosting and Extreme Gradient Boosting (XGBoost) are used in handling nonlinear relationships and interactions between variables.
[20] Gradient Boosting Regressor builds an additive model in a forward stage-wise manner. It iteratively fits new models to the residual errors made by prior models, effectively reducing prediction error by combining weak learners:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where $h_m(x)$ is the base learner at stage m, and γ_m is the learning rate.
[21] XGBoost, or Extreme Gradient Boosting, is an efficient and scalable implementation of gradient boosting. It includes advanced regularization (L1 and L2), which helps prevent overfitting, and supports parallel computation, making it highly suitable for large-scale regression tasks.
In our study, Gradient Boosting Regression model and XGBoost Regression model are used to estimate the car-following time headway based on multiple factors. By modeling THW as a function of these features, we aim to assess the importance of individual features.

- **Gradient Boosting** and **XGBoost** can capture nonlinear dependencies and feature interactions that are typical in human driving behavior.

- They are robust to multicollinearity and can handle heterogeneous feature scales without preprocessing.
- These models provide feature importance estimates that are helpful for interpretability and policy recommendation in driver behavior analysis.
- Their regularization techniques help prevent overfitting, making them ideal for real-world noisy traffic data.

By combining high prediction accuracy with interpretability, these regression models are well-suited for estimating the relative influence of driving factors on car-following time gaps.

2.2 Permutation Importance

Permutation importance is a model-agnostic technique designed to quantify the contribution of predictor variables to a model's predictive performance. The key idea is to randomly permute the values of a feature, thereby breaking its relationship with the response variable, and then measuring the resulting increase in prediction error. A greater increase in error indicates that the feature plays a more important role in the model. Permutation importance methods are used to correct biases in feature importance assessments and are particularly useful for tree models and other complex algorithms [22].

Traditional feature importance measures often suffer from bias: variables measured on larger scales or those with many categories tend to appear artificially more important [23].

Permutation importance addresses this issue by using a performance-based criterion that is not affected by the scale or number of categories. This makes it particularly suitable for complex machine learning models, such as random forests, gradient boosting, or neural networks.

In applied research, permutation importance has proven useful in domains where interpretability and robust variable selection are crucial. For example, Su et al. [24] applied this method to traffic event detection, showing that identifying influential flow variables through permutation-based assessment improved detection accuracy. Similarly, in bioinformatics and medical research, Altmann et al. demonstrated that permutation-based correction yields more stable and unbiased feature importance rankings compared to standard random forest measures [22].

By applying permutation importance, we are able to rank contextual factors according to their influence, which offers clear evidence for analyzing the influence of individual factors in car-following behavior.

2.3 Clustering Method

Understanding and categorizing driving behavior patterns is essential for traffic behavior analysis. Clustering methods are widely used to identify typical behavioral modes. In this study, we focus on two representative clustering approaches: the Hidden Markov Model (HMM), which captures temporal and latent-state dynamics for time series data, and K-Means, a classical clustering algorithm suitable for

grouping multivariate observations based on factor similarity.

2.3.1 Hidden Markov Model

HMM is a probabilistic framework commonly used for modeling the dynamics of multivariate time series, where the observed data are considered as stochastic functions of an underlying, unobservable (hidden) finite-state Markov process. In this context, the hidden states represent distinct behavioral modes or regimes, and the model captures both the temporal dependencies between states and the distribution of observations conditioned on those states. Zhang et al. fuse traffic contextual information into the driving behavior estimation of target vehicles by using continuous HMMs [25]. Yue Zhanga and Yajie Zou use the Nonhomogeneous Hidden Markov Model to decompose the merging processes into primitives containing semantic information [26].
An HMM is defined by:

- A set of hidden states $X = \{x_1, x_2, ..., x_N\}$ that evolve according to a Markov chain.
- An observation sequence $O = \{o_1, o_2, ..., o_T\}$ where each observation o_t is generated from a probability distribution conditioned on the current hidden state.
- A transition probability matrix $A = [a_{ij}]$ defining the probabilities of switching from one hidden state to another.
- An emission probability distribution B for generating observations from each state.
- An initial state distribution π.

In this study, HMM is applied for: Car-following behavior clustering and pattern recognition, and for Cut-in behavior modeling.
HMM is particularly suitable for these use cases since driver behavior involves transitions between discrete but unobservable modes, the latent-state formulation of HMM allows for the detection and interpretation of early-stage transitions, which is crucial for pattern recognition and prediction tasks in traffic scenarios.

2.3.2 K-means Model

K-Means is a widely used unsupervised machine learning algorithm for partitioning data into a predefined number of clusters. It is particularly effective for identifying groupings in multivariate space where similar observations are located close to one another based on a distance metric, typically the Euclidean distance. The algorithm iteratively refines cluster assignments by minimizing the within-cluster variance, which makes it computationally efficient and easy to implement. Higgs and Abbas use k-means to segments and clusters car-following behaviors based on eight state–action variables and get state–action clusters that define the driving pattern of drivers [27].
Formally, given a set of observations $X = \{x_1, x_2, \ldots, x_n\}$, where each $x_i \in \mathbb{R}^d$, K-Means aims to partition the data into K clusters $C = \{C_1, C_2, \ldots, C_K\}$ such that the total within-cluster sum of squares is minimized:

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where μ_k is the centroid of cluster C_k.
In this study, K-Means is applied to analyze and categorize different types of driver deceleration behavior using features such as maximum deceleration, average deceleration, and deceleration duration.
K-Means is an appropriate choice for this task since it is fast and simple, allowing for efficient experimentation with different clustering configurations and enough for this case due to the data distribution is easy to separate in different cluster by the distinct features. It provides interpretable cluster centers that can serve as representative behavioral profiles.
These clustering methods enable the identification and interpretation of driving patterns, which is essential for analyzing how contextual factors shape car-following behavior.

2.4 Random Forest Classifier

The Random Forest Classifier is a powerful and widely used ensemble machine learning algorithm, particularly effective for classification tasks. A Random Forest constructs a "forest" of multiple decision trees during its training phase. The process involves two key randomization mechanisms:

- Bagging (Bootstrap Aggregating): For each individual decision tree in the forest, a separate training dataset is created by bootstrapping from the original training data. Data points are randomly sampled from the original dataset. As a result, each bootstrap sample will likely contain some duplicate data points and omit others, ensuring diversity among the training sets for different trees.
- Random Feature Subspace Selection: When building each decision tree, at every node split, a random subset of features is selected from the total available features. The algorithm then searches for the best split. This additional layer of randomness further decorrelates the trees, preventing any single highly predictive feature from dominating all trees and thereby reducing the variance of the overall model.

Each tree is grown to its maximum possible depth without pruning (or until a minimum number of samples per leaf node is reached). For a new, unseen observation, each decision tree in the forest makes an independent prediction. In the case of classification, the final output of the Random Forest is determined by a majority vote among all the individual tree predictions.
In this study, Random Forest Classifier is used for selecting the key features for the scenarios analysis. The Random Forest Classifier was particularly well-suited for the task due to its ability to handle non-linear relationships inherent in complex driving behaviors, its robustness to noise, and its capacity to provide intrinsic feature importance scores. This latter capability is good for identifying which feature that have the most significant influence on the distribution model [28].

2.5 Filters

2.5.1 Butterwort Low pass filter

The Butterworth filter is a type of signal processing filter designed to have a frequency response that is as flat as possible in the passband. This characteristic is crucial as it minimizes distortion to the desired signal components. Unlike other filters (e.g., Chebyshev filters) that might exhibit ripples in the passband, the Butterworth filter maintains a smooth, monotonic frequency response.
For a low-pass Butterworth filter of order N, its squared magnitude response $|H(j\omega)|^2$ is given by:

$$|H(j\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2N}}$$

where:

- ω represents the angular frequency of the input signal.
- ω_c is the cutoff angular frequency, often defined as the frequency at which the filter's power output is half of the input power (i.e., -3 dB point). Signals with frequencies significantly below ω_c are passed with minimal attenuation, while those significantly above ω_c are attenuated.
- N is the filter order, which determines the steepness of the filter's roll-off in the stopband. A higher order N results in a sharper transition from the passband to the stopband, meaning a more aggressive attenuation of high-frequency components.

The key advantages of the Butterworth filter for this application are its maximally flat passband, which preserves the integrity of the slow-varying lighting signal, and its smooth, monotonic roll-off in the stopband, which avoids introducing unwanted oscillations or ringing artifacts into the filtered data.

2.5.2 Gaussian smoothing filter

A Gaussian smoothing filter is a linear filter that uses a Gaussian function to calculate the weights of the neighboring data points. When applied to a time-series signal, it replaces each data point with a weighted average of its neighbors, where the weights are determined by a Gaussian curve. Data points closer to the center of the window receive higher weights, while those further away receive lower weights.
The one-dimensional Gaussian function is given by:

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where:
x is the distance from the center of the distribution (or the current data point).
σ is the standard deviation of the Gaussian function. This parameter is crucial as it controls the "width" of the smoothing kernel and, consequently, the degree of smoothing. A larger σ results in a wider kernel, incorporating more distant data points into the average and thus leading to more aggressive smoothing.

This effectively blurs the signal, reducing high-frequency noise and smoothing out sudden fluctuations. A key characteristic of the Gaussian filter is that it prioritizes nearby points more heavily, providing a very natural and gentle smoothing effect that minimizes the introduction of artificial sharp edges or oscillations that can sometimes be associated with other types of filters.
Both filtering methods are employed to preprocess the time-series data from the Volvo Cars sample dataset, ensuring the removal of noise for further analysis.

2.6 Shortest Interval with majority of the data

In this study, we derive a quantitative analysis method that we use consistently throughout this study to extract the shortest interval with the majority of data from the data distribution. This approach is inspired by the principles of statistical hypothesis testing, Standard statistical practices use quantiles (e.g., interquartile range for 50% or 5th-95th percentile ranges) to describe the spread of the bulk of the data. For example, Akhilesh et al. use 85 quantile as one of the Statistical parameters to model time headway distribution [29]. We optimized it in a way that more suitable for our situation, where the distribution is multimodal or skewed and percentile cuts from the ends might not capture the true "majority" peak.
The idea of the algorithm is for a density distribution, we calculated the cumulative density function of the distribution, then calculate the shortest interval for the distribution axis that encompasses majority of the probability.
This shortest interval method is applied in the analysis of single contextual factor influence and scenario-based studies to consistently capture the predominant driving behavior.
In summary, the methods presented in this chapter provide a complementary toolkit for analyzing car-following time gaps under different driving contexts. Regression models and permutation importance quantify the influence of individual factors, while clustering techniques reveal latent behavioral patterns. Filtering methods ensure data quality. In addition, the shortest-interval approach offers a consistent measure of predominant behavior in multimodal distributions. These methods form a framework that directly supports our analysis of contextual factors and scenario studies.

3
Implementations

With the theoretical methods introduced in Chapter 2, this chapter explains how they are implemented with the available datasets. We begin by describing the datasets used in this study and the data preprocessing steps. A steady car-following filter is then constructed and compared with an experienced-based filter, providing the foundation for selecting valid car-following segments. Following the overall research outline, the chapter proceeds with a detailed explanation of the analytical workflow, including multi-factor ranking, selected single-factor analysis, and scenario-based studies.

3.1 Dataset

This project utilizes two primary sources of naturalist driving data for analysis: publicly available datasets highD and exiD, and a Volvo Cars sample dataset. While both sources contain a large quantity of data on vehicle trajectories and driving contents, including fundamental parameters like vehicle speed, acceleration, and distance headway, they also possess distinct information that complements the overall study.

1. highD and exiD, large-scale vehicle trajectory datasets collected from German highways.The two datasets are collected on different types of road infrastructure. The highD dataset provides naturalistic driving trajectories on straight highway segments, whereas the exiD dataset covers highway sections with on- and off-ramps. This distinction enables a more comprehensive analysis of car-following behavior in relation to straight roads and ramp areas.
2. Volvo Cars sample dataset, a dataset collected by test vehicles from multiple countries and locations. The real-world driving data complement the public datasets by providing longer car-following durations, diverse infrastructures and lighting condition.

3.1.1 Public Datasets: highD and exiD

For the analysis of public data, we utilize two well-known datasets: highD and exiD which provide extensive vehicle trajectory data collected on German highways by drone.

- The highD dataset provides vehicle trajectory data collected on German highways, focusing primarily on straight lanes without ramps. Making it suitable for analyzing car-following behavior without the complexities of ramps [30].

- The exiD dataset includes highway infrastructure with both on-ramps and off-ramps, capturing more complex merging and diverging behaviors.This allows for the capture and analysis of more complex merging and diverging behaviors, which are crucial for understanding car-following in dynamic traffic environments [31].

Both highD and exiD datasets are naturalistic driving data collected from an aerial perspective on German Highways using a drone at 6 different locations by the team from the Institute for Automotive Engineering, RWTH Aachen University [30], [31]. These datasets include information about the vehicle classes, widths, sizes, driving dynamics such as velocity, accelerations, time gap, information about surrounding vehicles, and the infrastructure information of all the passing vehicles within the record road section, which are essential for our project which make it suitable for constructing realistic car-following scenarios due to their detailed information on all road users and infrastructure.
The main difference between the two datasets is the ramp infrastructure. This enables us to investigate and compare the car-following behavior with and without the influence of ramps and merging vehicle. Furthermore, since the dataset records comprehensive information and assigns a unique identifier (ID) to all vehicles captured in the recording, we are able to extract the trajectories of surrounding vehicles relative to the ego vehicle, as well as other information such as vehicle type. This allows us to thoroughly investigate the influence of other road users on car-following behavior. In our study, we assume that human drivers drive all the vehicles recorded and analyzed, and that ACC is not activated.

Based on these public datasets, our analysis investigates the influence of the following factors and scenarios on car-following time gaps:

- Cut-in intention of other vehicles
- On-ramp areas (with and without merging vehicles)
- Number of lanes and lane selection
- Lead vehicle type

3.1.2 Volvo Cars sample dataset

Volvo Cars sample dataset is a high-quality naturalistic driving dataset collected by the data center at Volvo Cars. In this study, we only utilize manual driven data collect on highway/motorway. Compared to public datasets, it offers longer car-following driving sequences and additional information, including lighting condition and road curvature. These features enable analysis of conditions that are not covered by public datasets. The dataset also assigns a unique ID to other road users, allowing for the study of the influence of surrounding vehicles. Since it's recorded in multiple countries and locations in different road infrastructure and conditions, the Volvo Cars sample dataset also provide more data variability. These additional inputs allow for a more comprehensive understanding of driving behavior under real-world conditions.
Since the data are collected by professional test drivers on test purpose, their driving behavior tends to be more by the book, more consistent, and steadier compared to

that of naturalistic everyday driving.
Figure 3.1 illustrates the longitudinal velocities of a leading vehicle and its following ego vehicle from a random drive segment within the Volvo Cars sample dataset. It can be observed that the ego vehicle closely follows the lead vehicle, and their speeds are highly correlated for most of the duration.

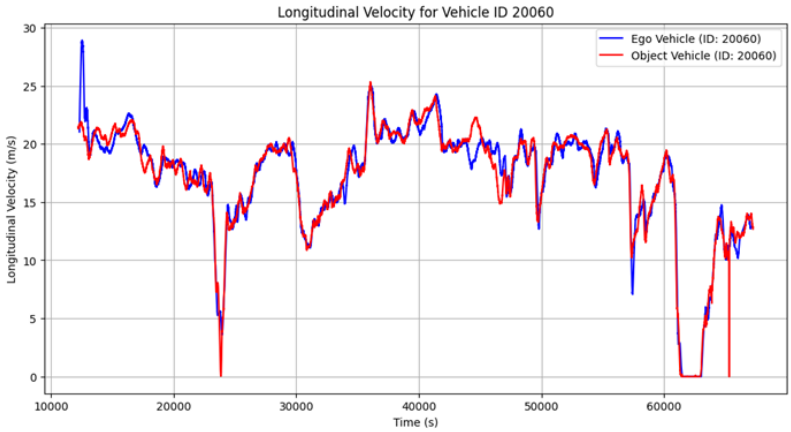


Figure 3.1: Longitudinal velocity for leading and ego vehicles (ID: 20060) during a car-following event from the Volvo Cars sample dataset.

The additional inputs available in the Volvo Cars sample dataset allow for a more comprehensive understanding of driving behavior under real-world conditions, particularly concerning environmental and contextual factors not typically available in public datasets. Specifically, the Volvo Cars sample dataset enables us to explore the impact of:

- Lighting conditions (e.g., day vs. night driving)
- Number of lanes and lane selection
- Lead vehicle types

Incorporating these factors allows for a more adaptive ACC system that better imitates human driving preferences.

3.2 Data Pre-processing

We performed a consistent analysis across heterogeneous datasets (highD, exiD, and Volvo Cars sample dataset). Core features such as surrounding vehicles, lane information, and environmental conditions are processed in a similar and comparable way, while dataset-specific adjustments are applied due to the existing differences in datasets.
The public datasets are already calibrated by the data owner using state-of-the-art computer vision algorithms, exhibit typical positioning errors of less than ten centimeters [30]. This high level of pre-calibration allows for the direct application of the steady-state car-following filter. In contrast, the Volvo Cars sample dataset requires specialized pre-processing, including removing outliers and noise, before similar filters can be applied. Differences also exist in the processing of infrastructure data between the two sources.

Gaps accurately reflect scenarios relevant to typical ACC function application; we made specific data selection choices. We limited the dataset to vehicle-following instances that occur on a highway. Since ACC systems are mainly designed for and engaged in these types of roads, characterized by multi-lane traffic, higher speed limits, additionally, the data used were restricted to highway/motorway driving and vehicle-following instances where the ego vehicle's speed was within the range of 20 – 40 m/s. Since where traffic is typically uncongested and flowing relatively freely so human drivers prefer to use the ACC function, and the majority of the public data falls into this speed range.
The main steps are as follows:

1. **Surrounding Vehicles:** For each ego vehicle under study, we identify the surrounding vehicles. Using the unique vehicle ID and the frame index, we extract the dynamics (velocity and acceleration) of both ego and surrounding vehicles.
   - In the highD dataset, this is achieved directly using the provided unique IDs and frame numbers.
   - In the exiD dataset, a similar procedure is applied; however, the lane has been separated into segments for further discussion of different roads close to the on-ramp and off-ramp infrastructure. Therefore, we need to collect the fragment information and store the information to a json file so that we know whether the ego vehicle is passing by on/off ramp section, which is one of the scenarios we are interested in.
2. **Lane Identification** Lane information is essential for determining the driving context of the ego vehicle.
   - In the highD dataset, the lane number detected by the drone provides the current lane of the ego vehicle.
   - In the exiD dataset, lanes are further divided into segments, particularly to capture behavior near on- and off-ramp sections. Segment information is stored in a JSON file to preserve this context.
   - In the Volvo Cars sample dataset, lane information is obtained from HD map and cross-verified with the detection of vehicle perception system.
3. **Environmental and Sensor Signals.** These factors are provided only in Volvo Cars sample dataset, so the preprocessing is only based on this dataset.
   Lighting Condition
   Volvo Cars sample dataset provides a numerical indicator for the ambient lighting conditions during each drive. This data ranges from 0 to 10000 (arbitrary units). As illustrated in Figure 3.2 and Figure 3.3, a significant amount of the collected data is distributed at the extreme ends of this range, specifically around 0 (representing very dark conditions) and 10000 (representing very bright conditions).
   This data proved to be sensitive to road conditions. We observed significant noise influence and sudden drops in brightness values, presumably caused by environmental factors such as the vehicle driving under trees or bridges, which would lead to abrupt reductions in light readings. This required further scrutiny and potential mitigation strategies to ensure the reliability of lighting condition analysis.

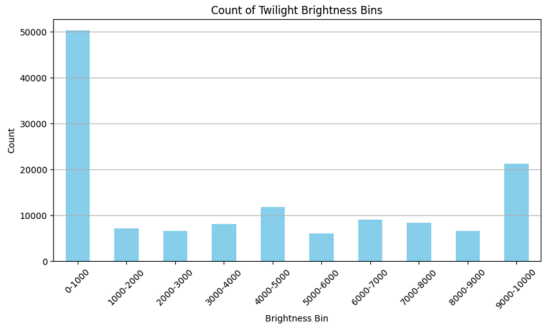


Figure 3.2: Count of lighting condition data points across different bins in the Volvo Cars sample dataset.

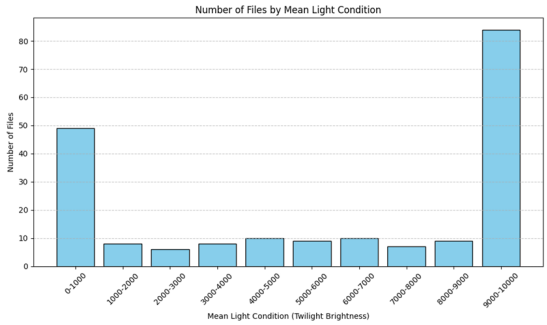


Figure 3.3: Number of files (car-following drives) by mean light condition bin, showing data distribution by drive.

The two pre-processing approaches for lighting condition data are as follows:

- **Method 1: Custom Thresholding and Gaussian Smoothing.** This approach was designed to specifically target sudden data drops and general noise.
  (a) *Outlier Removal:* To eliminate sudden drops in brightness values, a rolling mean window was applied to the lighting condition data. A dynamic threshold, determined by this rolling mean, was then set. If a data point deviated excessively from the rolling mean and surpassed this threshold, it was identified as an outlier and replaced by the corresponding rolling mean value. This method effectively mitigates abrupt, presumably erroneous, light reductions.
  (b) *Noise Reduction:* Following outlier removal, a Gaussian smoothing filter was applied to the data. Gaussian filters are particularly suitable for smoothing out general noise and high-frequency fluctuations while preserving the overall trend of the signal, given their bell-shaped weighting function that prioritizes values closer to the center of the window.
- **Method 2: Time-Frequency Conversion with Butterworth Filter.** This method leveraged frequency domain analysis for noise and artifact suppression.
  Low-Pass Filtering: To isolate noise and sudden drops, which typically manifest as high-frequency components. A low-pass Butterworth filter was applied to the lighting condition data. Butterworth filters are known for their maximally flat frequency response in the passband and monotonic roll-off in the stopband, making them ideal for effectively filtering out high-frequency noise and sudden signal artifacts without introducing significant ripples or phase distortion in the desired low-frequency signal, thus preserving the underlying trend of the brightness.

Here's examples of the raw and processed lighting signals using the first method, which combines a custom threshold with a Gaussian filter. Figures 3.4 and 3.5 display the lighting condition evolution for random drives, both in their original form and after the application of this smoothing technique.
As can be observed from these examples, the applied filtering effectively mit-

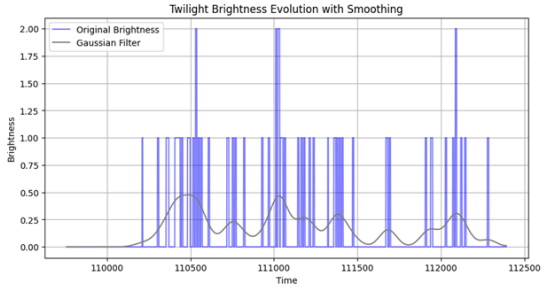


Figure 3.4: Raw and Processed Lighting Condition Evolution with Smoothing for Random Drives Example 1

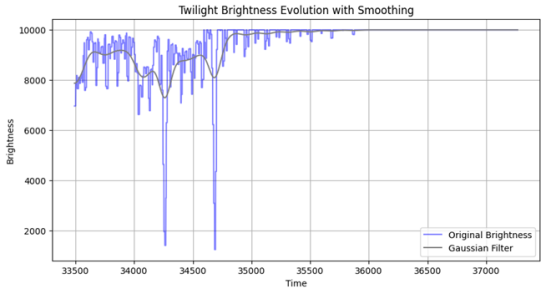


Figure 3.5: Raw and Processed Lighting Condition Evolution with Smoothing for Random Drives Example 2

igates or removes sudden drops in the data. Furthermore, the Gaussian filter effectively removes high-frequency noise, while the custom threshold helps to eliminate glitches or spurious readings in the raw brightness signal. This preprocessing ensures that the lighting condition data used for subsequent analysis is more stable and accurately reflects the ambient brightness.
Then we evaluate the processed lighting signal using the second method, which employs a Butterworth low-pass filter in the frequency domain. Figures 3.6 and 3.7 present examples of the raw and Butterworth-filtered lighting condition signals for random drives.

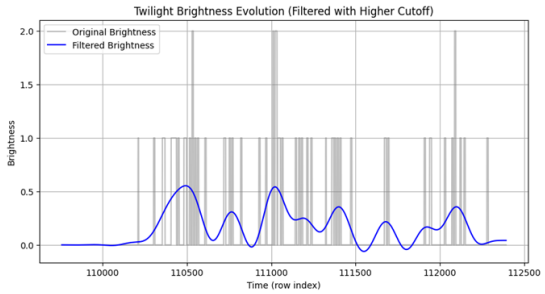


Figure 3.6: Raw and Butterworth-Filtered Lighting Condition Evolution for Random Drives Example 1

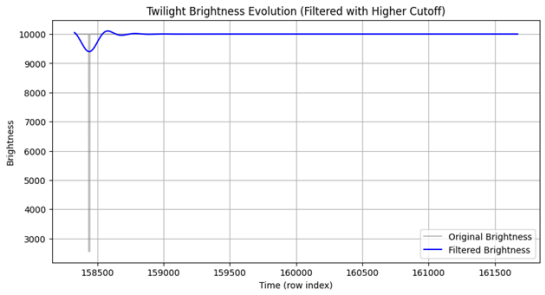


Figure 3.7: Raw and Butterworth-Filtered Lighting Condition Evolution for Random Drives Example 2

As observed in Figures 3.6 and 3.7, the Butterworth filter demonstrates a similar effectiveness in smoothing the signal compared to the custom threshold and Gaussian filter method. It successfully mitigates sudden drops and effectively removes high-frequency noise from the signal.
However, certain characteristics inherent to Butterworth filters are also visible. For instance, in Figure 3.7, a small ripple is noticeable at the onset of the filtered signal. Additionally, in Figure 3.6, the filtered brightness values occasionally drop slightly below zero. These phenomena are attributable to the Butterworth filter's design. While it offers a maximally flat passband and a monotonic roll-off (a key advantage for preserving signal integrity within the desired frequency range and avoiding sharp discontinuities), its infinite impulse response (IIR) nature can sometimes introduce minor ringing arti-

facts or, when applied to non-negative signals, produce small negative values. This occurs because the filter's impulse response can have negative components, which, when convolved with a signal near zero, can result in temporary negative outputs.
Despite these minor trade-offs, the results obtained from the Butterworth filtering are acceptable and yield conclusions similar to those derived from the custom threshold and Gaussian filtering method. Furthermore, the Butterworth filtering approach is often more straightforward to implement and less complex in terms of parameter tuning compared to developing a custom thresholding and smoothing pipeline, making it a robust and efficient alternative for signal preprocessing.
This dual-approach was adopted to compare and verify the effectiveness of different signal processing techniques from varied perspectives, ensuring a robust and reliable processing method for lighting condition data. This dual approach served as a verification step, allowing us to assess if both methods yielded consistent underlying brightness trends, thereby increasing confidence in our processed data.

3.3 Steady Car-following Filter

Since this study specifically focuses on the car-following time gap under steady-state conditions, it is imperative to derive a comprehensive and robust criterion to accurately filter steady car-following data from the raw datasets. A meticulously defined steady car-following criterion is crucial to avoid introducing bias into the dataset, which is fundamental to the reliability of our subsequent analysis.
From the literature review, a rule-based steady car-following criterion has been proposed by Loulizi [32]. This criterion includes the following conditions:

1. There is a lead vehicle, and the lead vehicle does not change during the car following state.
2. Ego vehicle does not change its lane during the car following state.
3. Absolute value of the speed difference between the lead vehicle and ego vehicle $< 5\%$
4. Acceleration and deceleration during the event < 0.2 m/s^2
5. Time headway < 4 s
6. Distance headway < 100 m

To determine the suitability and validity of this filter criterion for our specific datasets and to ensure it does not introduce unintended bias, we conducted an analysis of this criteria and identified typical steady car-following patterns from the public data. This internal validation process was essential to confirm the fundamental assumptions of our study.
To effectively model and describe the dynamic relationship between the lead vehicle and the ego vehicle, the following steps are taken to select the feature to ensure the robustness and interpretability of for the car following pattern recognition model:

1. Select all contextual features from the public datasets, which may influence drivers' choice on car-following time gap. These features capture the essential aspects of car-following behavior.

2. A Standard Scaler is applied to the selected feature data to ensure that each feature contributes equally to the analysis, preventing features with larger scales from dominating.
3. Principal Component Analysis (PCA) is applied to identify the most significant underlying patterns in high-dimensional data, thereby simplifying the model and mitigating multicollinearity. Working with many features directly increases the computational complexity of training and interpreting the HMM. If highly correlated features are fed directly into a clustering model like HMM, it can lead to less reliable clustering results since multicollinearity can make the model's parameters less stable and more sensitive to small changes in the input data. PCA is an appropriate choice here since it effectively reduces the dimensionality of the dataset while preserving most of the variance by transforming correlated features into a set of linearly uncorrelated variables.
4. The results of the PCA are then visualized. We also thoroughly examined the explained variance ratio and its cumulative sum to determine the optimal number of principal components that capture a substantial portion of the dataset's variance. Furthermore, feature importance was assessed through the analysis of component loadings.

Through these methods, we discovered that ego vehicle acceleration, distance headway, and the speed difference between the preceding vehicle and the ego vehicle emerged as the contextual and important features for modeling the dynamic relationship in car-following scenarios, which is also the factor used in the rule based filter criteria, suggesting that the rule base filter feature selection is reasonable.

To gain a deeper understanding of car-following behavior and identify distinct steady car-following patterns, an unsupervised clustering approach was employed. Initially, data was filtered to include only instances where a vehicle was following another, under the filter criteria: distance headway of less than 120 meters and a time headway of less than 5 s.

Subsequently, the set of contextual features capturing the vehicle's dynamic state was selected: ego vehicle acceleration, distance headway, speed difference between the preceding and ego vehicles, mean acceleration, and mean speed difference. These selected features were then standardized using the same Standard Scalar, consistent with the previous method to ensure uniform scaling.

An HMM model was then applied to cluster the car-following behavior into distinct states. HMMs are particularly suitable for this task because they can model sequential data, capturing both the underlying hidden states (car-following patterns) and the probabilistic transitions between these states, which is essential for understanding dynamic behaviors over time. They allow for the identification of behavioral patterns that might not be immediately apparent from individual data points.

The optimal number of hidden states for the HMM was determined using the elbow method, which indicated that four states provided the most appropriate balance between model complexity and explanatory power. Ultimately, a robust HMM was trained, effectively clustering the car-following behaviors into these four distinct groups, each representing a unique car-following pattern.

3.4 Multi-factor Importance Ranking

To investigate the impact of driving dynamics, surrounding road users, road infrastructure, and other environmental factors, we first aggregated and preprocessed trajectory data from files in different maps. We selected Map 2 in the exiD dataset as the primary infrastructure for this study due to its clearly defined lane structure and well-separated regions, including the main road and the acceleration lane. The dataset was filtered to include only three vehicle types: car, van, and truck (denoted as trackClass 0, 1, and 2, respectively). For each vehicle, a unique identifier was created by combining its file ID and track ID.

3.4.1 Factor Selection and Grouping

The ego vehicle in this study is restricted to cars, as the focus of this research is on car driving behavior.
The selection of factors in this study follows the driver–vehicle–environment framework discussed by Han et al. , which emphasizes that realistic car-following modeling should incorporate driver–vehicle dynamics, surrounding road users, and broader traffic environment conditions [33]. Accordingly, we include variables representing vehicle motion (velocity, acceleration, relative speed), interactions with surrounding vehicles (e.g., presence of lead, rear, and alongside vehicles in adjacent lanes), and traffic measures (traffic flow and density).
For infrastructure-related variables, previous research has shown that the design of the acceleration lane significantly affects vehicle behavior near on-ramps. Kita found that longer merging lanes can alter how mainline drivers adjust their spacing [34]. This supports our inclusion of distance-to-merge-point and acceleration-lane-end distance as contextual features, since they capture the spatial constraints influencing driver anticipation and gap adjustment in merging areas.
In the Volvo Cars sample dataset, we also include additional environmental factors such as lighting conditions and wiper status. However, after filtering for stable car-following data, the processed curvature values were consistently within the range of –0.001 to 0.001, indicating that all selected segments were essentially straight road sections[35]. Therefore, curvature was not included in the subsequent analysis.
We focus only on vehicles located in the target lane. Therefore, we use the lane IDs provided in the dataset's Lanelet2 map to generate a map information file (data.json) for lane matching.
Since speed is the dominant factor in keeping a car-following time gap, we computed the mean velocity for each vehicle instance and categorized vehicles into three speed groups:

1. 20 – 25 m/s
2. 25 – 30 m/s
3. 30 – 35 m/s
4. 35 – 40 m/s

However, in the research in the public dataset, the data ranging 35 – 40 m/s are really limited, so we only compare the other three speed groups. Additionally, jerk was derived as the time rate of change of acceleration using a fixed frame interval

of 0.04 s.
The factors selected from the public dataset are:

1. **Vehicle dynamics:**
   (a) Velocity
   (b) Acceleration
   (c) Relative speed to the lead vehicle
2. **Lead vehicle type:** car / truck
3. **Traffic:**
   (a) Traffic flow: Number of unique vehicles passing through the exit lane within a one-second window centered on the current frame.
   (b) Traffic density: Number of vehicles across all same-direction lanes at the current frame. Since vehicles are counted within the same frame, the roadway segment length considered is identical for all measurements.
4. **Other road users:**
   (a) Other road users include vehicles on the left and right side, including the presence of vehicles in the lead, rear, and alongside positions. They are derived from ID fields (leftLeadId, leftRearId, leftAlongsideId, rightLeadId, rightRearId, rightAlongsideId).
   (b) Lateral distance to the right-side vehicle was computed based on vehicle offsets and widths.
5. **Infrastructure:**
   (a) Distance to the end of the acceleration lane, calculated based on map-specific geometric rules.
   (b) Distance to the merging point (distJoin), which can be positive or negative depending on vehicle location relative to the merge.
   (c) Lateral offset from the lane center was used to estimate lateral distances to neighboring vehicles.

Features were preprocessed to ensure comparability across variables. Continuous features (e.g., velocity, acceleration, lateral offset) were standardized to have zero mean and unit variance, facilitating model convergence and avoiding scale dominance. Traffic flow and traffic density values, which have bounded but skewed distributions, were normalized using min–max scaling to the [0,1] range. Binary and categorical variables, such as the presence of surrounding vehicles and lead vehicle type, were retained in their original form to preserve interpretability.[36] There are no missing or invalid entries because the dataset has undergone initial preprocessing before being published.

3.4.2 Regression model and permutation importance

A Gradient Boosting Regressor model is trained for each velocity group. An 80/20 train-test split was used within each group. After fitting the model, permutation importance was computed to evaluate the contribution of each feature to the prediction of time headway.
The feature importance results were visualized using heatmap plots. Additionally, the plots are shown without velocity-related features, such as velocity, acceleration, and jerk, to better understand the influence of other contextual factors.

Although we also experimented with the XGBRegressor, it did not significantly affect the results, and the Gradient Boosting Regressor was retained for final analysis due to comparable performance and interpretability.

3.5 Baseline time gap for traffic group

In our study, to evaluate if a particular variable significantly influences the car-following time gap, we proceed as follows: we construct two distinct traffic groups by applying identical and consistent filtering criteria to the data, with the sole exception of the single variable under investigation. For each of these variable-effected traffic groups, we then calculate key statistical measures including the mean, standard deviation (std), 25th percentile, 75th percentile, and the majority data selection which mean the shortest interval that contain most of the data. This systematic quantitative analysis of differences across groups allows us to rigorously assess the influence of the single factor on car-following time gaps. This method is inspire by the common use of Kolmogorov-Smirnov (K-S) test to study whether distribution are significantly difference [29].

The time headway (THW) distribution across different traffic groups typically exhibits a trapezoidal shape as we can observe from Fig 3.8. This characteristic distribution is marked by a high concentration of data points in the central portion, representing the preferred or common time gaps, while fewer data points (often considered outliers) are distributed at the lower and upper tails of the range. Our primary interest lies in analyzing this central, or majority, time gap chosen by human drivers, as these time headway values are generally perceived as safe and comfortable by most. This focus allows us to specifically filter out scenarios where the time headway might be excessively small (indicating following too closely, with a high risk of collision) or unnecessarily large (suggesting an excessively far following distance that potentially reduces road capacity). These driving behavior (too conservative or too aggressive) are not representative of a driver's preferred or stable car-following behavior under normal conditions, nor are they the intended operational scenarios for ACC systems,. Including them would skew the analysis towards reactive, high-risk situations rather than typical following patterns.
To systematically identify where the majority of the data points are distributed within a given traffic group, we derived the algorithm to identify the shortest interval containing the majority (specifically, 90%) of the time headway data points within a given traffic group 2.6. The choice of 90% as the threshold for this majority selection was made after empirical observation of the time headway distributions across various traffic groups and lane types within the dataset, it robustly captures the central, "flat" or most dense region of these trapezoidal distributions as we seem in Fig 3.8. This central region represents the predominant and most preferred time headway choices made by the majority of drivers under typical car-following conditions. This threshold effectively excludes the "triangle areas" or tails at both ends of the distributions. The process unfolds in several steps:

1. **Density Distribution Estimation:** First, the density distribution of the time headway for the single-variable-effected traffic group is calculated us-

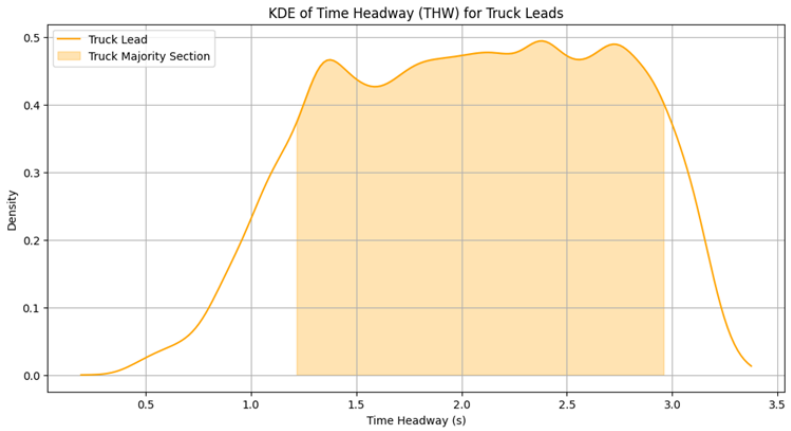


Figure 3.8: Time headway distribution for Truck leads data, highlighting 80% of the data in shortest interval(20 – 30 m/s, Volvo Cars sample dataset)

ing Kernel Density Estimation (KDE). This provides a continuous probability density function ($f(x)$) for the time headway values (x).

2. **Normalization:** To ensure that the total area under the density curve is unity, thus avoiding any bias from varying sample sizes or scales, the estimated density distribution is normalized. This effectively transforms it into a true probability distribution.
3. **Cumulative Density Computation:** The cumulative density function is then computed from the normalized density distribution. This function represents the probability that a time headway value is less than or equal to a given point, providing a continuous sum of the probabilities.
4. **Shortest Interval Identification:** From the cumulative density function, we systematically search for the shortest interval on the time headway axis that encompasses 90% of the data's total probability. This involves iterating through all possible start and end points on the time headway axis and identifying the interval for which the difference in cumulative density is at least 0.9, while simultaneously minimizing the length of that interval.

This robust method allows us to quantitatively characterize the central tendency and spread of preferred time gaps, enabling a focused analysis on the most typical human driving behaviors rather than being skewed by outliers.
The statistical baseline is made on recording locations that have 2 lanes. The data is calculated on the inner lane for velocity over 20 m/s using highD dataset.

Time headway (s)	**mean**	**25%**	**75%**	**std**
All speed highway	1.19	0.75	1.50	0.59
20 – 25 m/s highway	1.18	0.73	1.50	0.61
25 – 30 m/s highway	1.17	0.72	1.47	0.59
30 – 35 m/s highway	1.21	0.76	1.53	0.60
>35 m/s highway	1.21	0.75	1.55	0.60

Table 3.1: Baseline calculated with straight lane data

3.6 Selected individual factors analysis

In this section, we delve into the individual influence of specific factors on the car-following time gap, which are of particular interest to this study. We closely investigate the impact of:Lead vehicle types, Number of lanes and lane selection, Lighting conditions. These feature are chosen given the following reason: Firstly, they play a notable role in shaping time headway preferences from factor important ranking 4.2. especially for lighting condition from the Volvo Cars sample dataset result. Secondly, For highly dynamic and real-time influencing factors such as the immediate presence and behavior of surrounding vehicles (e.g., tail vehicles, left-lead vehicles), their interactions are complex to analyze due to their transient nature and strong interdependencies. In contrast, lead vehicle type, the number of available lanes, and lighting conditions represent more stable, persistent, and relatively independent contextual signals that drivers continuously process throughout a drive. Their influence is not momentary but sustained, making them amenable to a more systematic and robust analysis of driver adaptation. They provide a foundational understanding of driver preferences that is less prone to the rapid fluctuations of direct vehicle-to-vehicle interactions. Thirdly, these factor can be easily interpreted and incorporated into future ACC algorithms design.
Given that vehicle speed has been identified as the primary and most influential factor governing car-following time gaps, and recognizing that time headway distributions and driver behavior vary significantly across different speed ranges, we have categorized the traffic data into four distinct speed categories for a more granular analysis. These speed categories are defined as: 20 – 25 m/s, 25 – 30 m/s, 30 – 35 m/s, and 35 – 40 m/s. This categorization allows for a detailed examination of how the single factors influence time gap preferences within different speed ranges for human drivers.

3.6.1 Lead vehicle Types

Research finds that the desired headway of passenger car drivers are lower when following a truck than when following a passenger car [8]. To investigate and quantify the influence of lead vehicle type on car-following time gaps, we use both the public highD dataset and the Volvo Cars sample dataset.
In the public highD dataset, collected on German highways, only two primary vehicle types are present: cars and trucks. For this analysis, we filtered the data to include vehicles within the same predefined speed categories (20 – 25 m/s, 25 – 30 m/s, 30 – 35 m/s, and 35 – 40 m/s) and focused on car-following scenario occurring on all lanes of straight road segments.
Volvo Cars sample dataset offers a broader range of lead vehicle types, including cars, trucks, and motorcycles. However, it is important to note that the data volume for motorcycles on motorways in the dataset is relatively smaller and more limited compared to the data for cars and trucks. For the Volvo data analysis, we applied filters out with different lead vehicles in the different speed categories and focus on car-following events occurring on the outermost lane of the road. The lane filter criteria is base on the fact that there's a significantly larger volume of lead truck

data in outermost lane and little in inner lane since the heavy vehicle prefer to drive in outer lane. Also it's easier to filter data base on outermost lane from Volvo Cars sample dataset. Subsequently, we compared and analyzed the time headway distributions for each lead vehicle type: car, truck, and motorcycle.

3.6.2 Number of Lanes and Lane selection

Similarly, to investigate the influence of the number of lanes and lane selection on car-following time gaps, we utilized both the public highD dataset and the Volvo Cars sample dataset, applying similar filtering strategies.
For the public highD dataset, which was recorded on German highways featuring either two or three lanes, we examined the time headway distribution across different speed categories , varying numbers of lanes on the road, and the ego vehicle's lane selection (e.g., inner, middle, outer lane). This allowed us to assess how the lane configuration and a driver's choice of lane might affect car-following behavior.
For the Volvo Cars sample dataset, to maintain consistency and allow for comparative analysis, we similarly filtered for drives occurring on roads with two, or three lanes. We then analyzed the time headway distribution under different speed categories, the specific number of lanes on the road, and the ego vehicle's lane choice. This comparative approach across both datasets provides a comprehensive understanding of the impact of lane characteristics on car-following time gaps in various driving contexts.

3.6.3 Lighting Condition

To investigate the influence of lighting conditions on car-following time gaps, we focused exclusively on the Volvo Cars sample dataset due to it capture the ambient light. The lighting condition data, ranges from 0 to 10,000 lux. A significant majority of these values are distributed around the extremes, i.e., close to 0 and close to 10,000, with values near 0 typically indicating night conditions. Given this distribution, our investigation proceeds from two distinct perspectives:

1. **Average Light Condition vs. Average THW:** This approach seeks to understand the relationship between general light levels and preferred time gaps over extended driving periods.
2. **Categorical Analysis (Dark vs. Bright Conditions):** This method classifies driving segments into distinct "dark" and "bright" categories to analyze their respective THW distributions.

Average Light Condition vs. Average Time Headway
For this method, our interest lies in the driver's preferred time headway over the long term, rather than transient THW changes caused by sudden, momentary shifts in light conditions (e.g., passing under a bridge). We assume that, for a given drive segment, the overall lighting condition remains relatively consistent, and we are interested in the driver's general THW preference under this prevailing light. To achieve this, we calculated the mean lighting condition and the mean THW for every 500 data points, the data is collect by time series with fix time interval. A block size of 500 data points was chosen as a good balance: it is sufficiently large to reflect

a driver's long-term preference under a stable lighting condition, yet small enough to capture meaningful shifts in light over the course of a drive. We then fitted a linear regression model to describe the time gap evolution as a function of average light condition changes. The distribution trend and relationship between THW and lighting condition are subsequently visualized using a scatter plot.

Categorical Analysis (Dark vs. Bright Conditions)
For the second method, we employ the traffic group baseline and majority selection methodology again. Based on the observed distribution of lighting condition values, we classify the data into two distinct lighting condition groups:
Dark Condition Traffic Group: Data points where the lighting condition value falls within the range of [0 – 10000] (arbitrary units). These segments primarily represent night-time or very low-light driving scenarios. The selection of this range is supported by the bimodal distribution of the lighting data shown in Fig 3.2 and Fig 3.3 that the values concentrated is heavily concentrated in two distinct clusters: one centered around the [0 – 1000] range and another around the [9000 – 10000] range. The intermediate range [1000 – 9000] represents transitional periods. The range of [0 – 1000] value indicate environments with effectively no ambient light, representing complete darkness To distinctly investigate and compare driver behavior specifically under no light conditions versus light conditions (which encompasses both bright and transitional states), the threshold of 1000 arbitrary units was chosen to because it represents a natural inflection point where the frequency of data points for genuinely low-light conditions drops significantly.This allows us to create a clean separation, isolate and study driving dynamics specifically in genuinely dark environments, separate from twilight or daylight influences.
Bright Condition Traffic Group Data points where the lighting condition value falls within the range of [1000 – 10000] (arbitrary units). These segments primarily correspond to day-time or non-dark driving scenarios. It includes both the high-light cluster concentrated around the [9000 – 10000] range (representing full daylight) and the intermediate range (approximately [1000 – 9000]) that corresponds to transitional periods like dawn, dusk, or driving through a tunnel. This categorization allows for a direct and clear comparison between driving behavior in genuinely dark environments versus all other conditions where ambient light is a factor.
We then applied our established methods to compare and analyze the time headway distributions of these two distinct traffic groups, focusing on their majority sections to identify significant differences in driver preferences under dark versus bright conditions.

3.7 Scenarios where the time gap is influenced by infrastructure and road users

In our analysis of public driving datasets, we identified merging behavior as a frequent and influential scenario. Comparing the time headway baseline between highD and Exid suggests that drivers begin to adjust their time gaps in anticipation of a potential merge, even before the actual merging occurs. Cut in behavior is also of

interest for ACC system design. We aimed to identify and quantify the changes in car-following time gaps induced by drivers' reactions to potential events. Following this, we investigated how these scenario influences the time gap modifications preference by human drivers.

3.7.1 Scenario 1: Cut-in intention of other vehicles

To investigate the influence of other vehicles' cut-in intentions on the ego vehicle's time headway, we use the highD dataset. This dataset's focus on straight road segments is crucial as it allows us to minimize the confounding influence of factors like road curvature and other infrastructure elements.

Scenario Definition
The scenario of interest, as presented in Figure 3.9, is defined as follows: The ego vehicle is driving on a straight road while a vehicle in an adjacent lane initiates a cut-in maneuver, but the cut-in maneuver has not yet been fully completed (the intruding vehicle has not yet crossed the lane marking and established itself as the new lead vehicle directly in front of the ego vehicle). This is identified by the adjacent vehicle performing a noticeable lateral movement towards the ego vehicle's lane. Our primary terest lies in understanding the ego vehicle's pre-cautionary reactions, such as braking or increasing its following distance/time gap, in response to this potential cut-in intention.

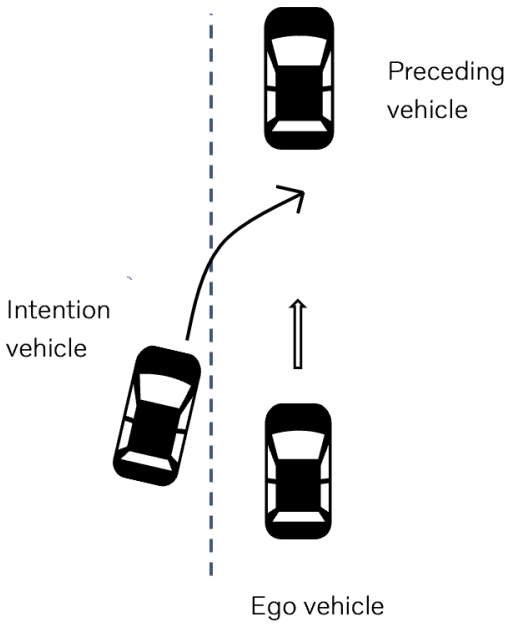


Figure 3.9: Scenario demonstration for Scenario 1

Data processing
To effectively analyze this, the first step is to accurately identify the intention period of the cutting-in vehicle, enabling the extraction of relevant data pertaining to both

the cut-in intention event and the ego vehicle's subsequent reaction. To accurately estimate lane-changing intention, we draw inspiration from Schlechtriemen et al. [37]. We first filtered the dataset to identify vehicles that successfully performed a cut-in maneuver (a lane change of surround vehicles resulting in a new lead vehicle for the ego vehicle).

Subsequently, we applied a methodology using feature analysis and selection then apply HMM model to cluster similarly to previous analyses, selecting the most relevant features to model the vehicle's movement and behavior throughout the cut-in process, which is inspired by Qiu et al. [38]. The chosen features include longitudinal velocity, longitudinal acceleration, lateral velocity and lateral acceleration. These features comprehensively describe the dynamic behavior of a vehicle during a lane change.

Cluster Cut in Behavior

An HMM model was then trained on this feature set to cluster the driver behaviors during these cut-in intention scenarios. HMMs are particularly suitable for this application because they excel at modeling sequential data, enabling the identification of distinct, underlying behavioral states (such as different phases of a cut-in intention or reaction) within a continuous time series. They can effectively capture the temporal dependencies and probabilistic transitions inherent in dynamic driving maneuvers. The optimal number of clusters was determined to be 5 using the elbow method, resulting in a robust HMM that efficiently categorizes driver behavior into these four distinct groups, with 3 of cluster can be used to identify cut in intention period and the rest of the cluster as normal driving.

Following the successful clustering of cut-in behaviors using the HMM, the data points and drive identified as corresponding to a "cut-in state" are then retrieved. This enables the specific analysis of the ego driver's reaction to vehicles exhibiting a cut-in intention or being in their intention period.

Cluster Driver Reaction

The second step of this investigation focuses on characterizing the ego driver's behavioral responses to these cut-in events, specifically examining deceleration behavior. Deceleration is considered a robust indicator of a driver's precautionary reaction to an intending cut-in vehicle. For each identified cut-in event, we calculated three key metrics for the ego vehicle: maximum deceleration, duration of deceleration (deceleration time), and mean deceleration. These metrics are computed for the interval from the moment the cut-in vehicle first shows its intention until the moment it fully cuts in front of the ego vehicle.

Then K-means clustering was applied to these three features (maximum deceleration, deceleration time, and mean deceleration) to categorize the ego vehicle's precautionary reactions. The elbow method was again employed to determine the optimal number of clusters, indicating that three distinct types of precautionary reactions provided the most meaningful segmentation of driver behavior in response to cut-in intentions.

Data visualization and Identify feature

Following the clustering of ego driver reactions into three distinct groups, the final step of this analysis aims to understand the underlying reasons for these varied reactions and to identify the specific scenarios that trigger each type of response. To achieve this, we investigated feature and scenario differences across the identified driver reaction clusters.

Specifically, for each reaction cluster, we calculated the average speed difference and average longitudinal distance headway between the cut-in intention vehicle and the ego vehicle during the entire cut-in intention period. Furthermore, we computed the average lateral acceleration and average lateral velocity of the cut-in intention vehicle during this same period.The feature is selected using the feature importance from random forest classifier that trained on the scenario specific features, random forest is chosen referring to Xue et al. [39] By examining the differences in these calculated metrics across the distinct reaction groups, we sought to uncover the contextual cues that lead to different precautionary behaviors. The variations across these groups were further visualized using KDE plots inspired by Björn Filzek et al. [40], providing a clear graphical representation of the distribution trends for each feature within each reaction cluster.

3.7.2 Scenario 2: Passing on-ramp area without merging vehicles

Lane-changing and merging behavior was identified by tracking changes in vehicle "lane IDs" over time, which allowed us to determine directly from the map whether, and at which location, a vehicle changed lanes within the segment. The analysis revealed that the most common merging event occurs when a vehicle in the acceleration lane merges into lane 0, the lane adjacent to on-ramp area as shown in Figure 3.10. The colors indicate different functional areas of the road: the red region corresponds to the acceleration lane and on-ramp, the blue region marks the analyzed mainline segment (lane 0, adjacent to the on-ramp). The division into numbered blocks (e.g., 1499, 1500) follows the segmentation provided by the exiD dataset, which splits the road into discrete segments for analysis. The small blue rectangle highlights an example ego vehicle, included to illustrate the reference position of the analyzed vehicle within the segment.

Based on these observations, Scenarios 2 and 3 were defined to examine both the effect of the on-ramp's physical presence on car-following behavior and the influence of vehicles present in the on-ramp lane on those in the adjacent lane.

Scenario definition This scenario investigates whether the presence of an on-ramp infrastructure influences car-following behavior when no merging vehicles are present in the adjacent acceleration lane. As shown in Figure 3.10, the ego vehicle is located in Lane 0, while the red road segment represents the ramp lane, which remains empty. The purpose of this setup is to isolate the effect of the ramp infrastructure itself, distinguishing it from the influence of merging traffic, and to compare it against baseline driving on a standard straight highway segment.

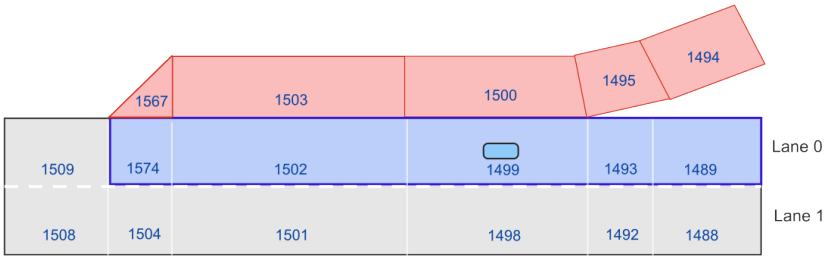


Figure 3.10: Example infrastructure for Scenario 2

Data selection The exiD dataset provides us with 6 infrastructures with different ramp infrastructures. We selected Map 2 due to its clearly defined lane structure and well-separated regions, including the main road and the acceleration lane. This clear spatial organization makes it particularly suitable for analyzing merging behaviors and on-ramp infrastructure according to Li et al. [41]. Moreover, the layout of a two-lane main road with a single on-ramp lane is a common highway configuration, enhancing the generalizability of our analysis.
The analysis focuses on vehicles driving in the blue-marked segment, corresponding to the lane adjacent to the acceleration lane (lane 0), as shown in Figure 3.10.

Data filtering Both conditions are filtered to include only steady car-following states. For the exiD dataset, additional filtering ensures that the acceleration lane contains zero vehicles during the selected frames.

Approach For both scenarios, we analyzed three key longitudinal variables: speed, acceleration, and THW. To statistically evaluate group differences, we applied a linear mixed-effects model, with `trackId` included as a random intercept to account for repeated measurements from the same vehicle. The model was estimated using maximum likelihood, and the general form of the model was:

$$y_{ij} = \beta_0 + \beta_1 \cdot \text{Condition}_{ij} + u_i + \epsilon_{ij},$$

where y_{ij} denotes the outcome variable (speed, acceleration, or THW) for vehicle i at frame j; Condition_{ij} indicates whether the vehicle was in the on-ramp-adjacent lane or the straight lane; u_i is the random intercept for vehicle $i \sim \mathcal{N}(0, \sigma_u^2)$; and ϵ_{ij} is the residual error term $\sim \mathcal{N}(0, \sigma^2)$. Hypothesis testing for fixed effects was conducted using likelihood ratio tests, with significance thresholds set at $p < 0.05$.
To characterize the empirical distributions of the three variables, we employed KDE with a Gaussian kernel. From the estimated densities, we extracted the shortest 80% highest-density intervals (HDI) to summarize the most probable range of values, providing a robust comparison of distributional characteristics across conditions.

Statistical analyses were performed in Python using the `statsmodels` package, and density estimation was carried out using `scipy` and `seaborn`.

3.7.3 Scenario 3: Passing on-ramp area with merging vehicles

Figure 3.11 follows the same definitions as Figure 3.10. In addition, the small blue rectangle in the red acceleration lane highlights an example ego vehicle located on the on-ramp.

Scenario definition This scenario investigates whether the presence of a vehicle in the adjacent acceleration lane, regardless of whether it merges, affects car-following behavior in the lane next to the on-ramp. As shown in Figure 3.11, the ego vehicle is located in Lane 0, while the red road segment represents the ramp lane containing another vehicle that may or may not merge into the ego lane to become a lead vehicle. The purpose of this setup is to isolate the influence of a neighboring vehicle in the ramp lane on the ego vehicle's car-following behavior and to compare it against both baseline highway driving and the empty-ramp condition.

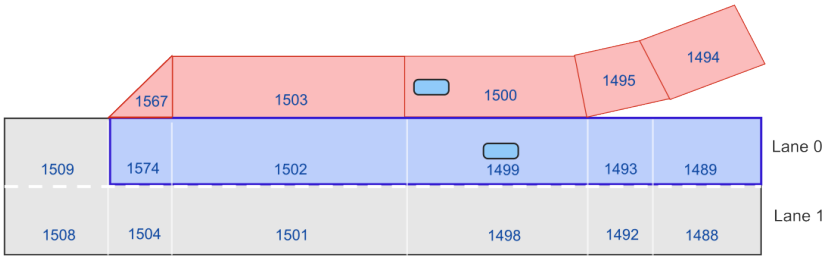


Figure 3.11: Example infrastructure for Scenario 3

Data selection Data are drawn from Map 2 of the exiD dataset, focusing on vehicles in lane 0, which is similar to scenario 2. The "with right vehicle" condition includes time frames where at least one vehicle is continuously present in the acceleration lane. The "without right vehicle" condition includes only time frames where the acceleration lane is empty.

Data filtering For both conditions, only steady car-following states within the target lane are retained. In the "with right vehicle" condition, data segments are truncated at the point where the adjacent vehicle leaves the acceleration lane.

Approach The analysis procedure follows the same steps as in Scenario 2. Speed, acceleration, and THW were extracted, and group differences were evaluated using a linear mixed-effects model with `trackId` as a random intercept. KDE was applied to estimate the empirical distributions, and the shortest 80% HDI was computed to summarize the range of most probable values. In this scenario, the two comparison groups were defined by the presence or absence of vehicles in the adjacent acceleration lane. To capture the effect of vehicle appearance or disappearance, segments were restricted to periods of continuous presence of the right-lane vehicle.

In summary, the implementation chapter addressed our work on dataset-related challenges, and provided the technical solution for analyzing car-following behavior. We emphasized the importance of data preprocessing for comparable factors provided to different theoritical models, and filtering when working with naturalistic driving datasets. Noisy or inconsistent signals, such as lighting conditions or lane assignment, required systematic filtering strategies to ensure that subsequent analyses were based on reliable input. Roads are separated into segments with unique IDS in public datasets, which helps us to clearly research the influence of specific infrastructures, but introduces dependencies on dataset-defined segmentation and additional data-processing effort.

A further limitation of the current implementation is the focus on on-ramp scenarios. While on-ramps are a critical source of interactions and merging behavior, they represent only one subset of highway driving contexts. Other factors, such as off-ramps, or cut-out behaviors, were not explicitly addressed here but may serve as key scenarios.

4
Results

In this section, we present a comprehensive overview of all findings and results derived from our data-driven analysis on car following time gap. Our study aimed to investigate as many factors and scenarios relevant to ACC function design as possible that can be extract from the dataset. To ensure the robustness and validity of our conclusions, and to effectively isolate the influence of individual variables of interest on the car-following time gap, we employed rigorous control strategies. For example, recognizing speed as the primary factor influencing time gap, we constructed and extracted traffic groups base on speed category and derive our findings on each group, ensuring that other confounding variables were minimized when analyzing the impact of a specific factor. This systematic approach results in a lot of detailed findings and conclusions, which will be elaborated upon in the subsequent subsections.

4.1 Car-following pattern recognition

To identify distinct car-following patterns, we trained an HMM using the features: ego vehicle acceleration, distance headway (DHW), and speed difference between the lead and ego vehicles. This analysis was performed on different speed groups within the highD dataset. Here, we present the results for the 20 – 25 m/s speed group.
Figure 4.1 displays the distributions of key features — DHW, THW, ego velocity, preceding vehicle velocity, and speed difference — across the four identified HMM clusters. These distributions serve as indicators to verify and interpret the characteristics of each cluster.
Based on the analysis of these feature distributions, we can interpret the meaning of each cluster:

- **Cluster 1 & Cluster 2 : Steady Car-Following Behavior.** These two clusters most likely reflect the behavior of steady car-following. This is primarily indicated by the "Distribution of Speed_Diff Across HMM States" plot in Figure 4.1, where the speed difference for both Cluster 1 and Cluster 2 is highly concentrated around 0. This signifies that the ego vehicle is maintaining a consistent speed relative to its lead vehicle.
  - **Cluster 1 (Light Blue): Steady Following with close distance:** This cluster represents a closely steady following behavior. As shown in the plots, the time headway and distance headway distributions for Cluster 1 are relatively small, with a small standard deviation. This

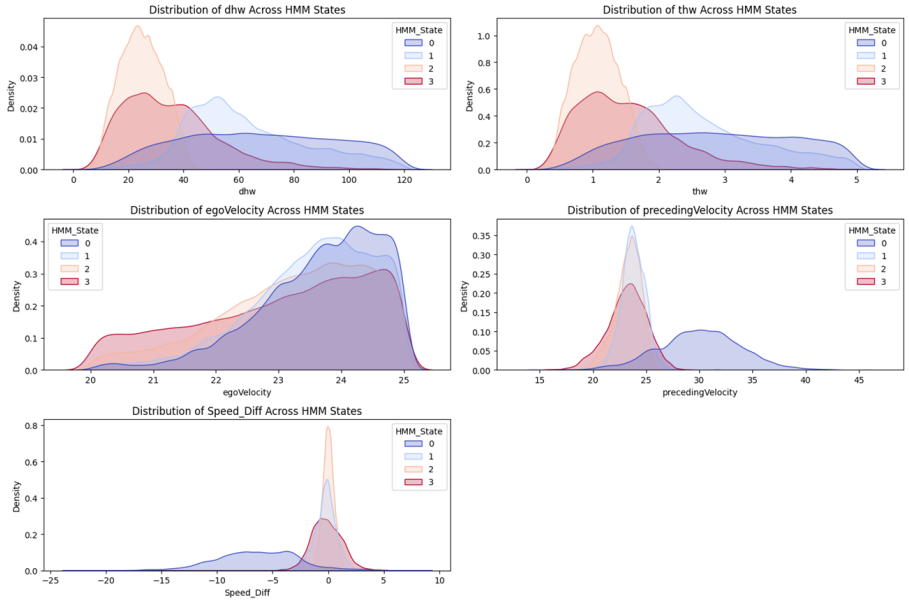


Figure 4.1: Distribution of Distance Headway, Time Headway, Ego Velocity, Preceding Velocity, and Speed Difference Across HMM States for 20 – 25 m/s Speed Group by highD dataset

indicates that most vehicles in this cluster are closely following the lead vehicle, typically maintaining a distance under 20 – 40 meters.

  - **Cluster 2 (Orange): Steady Following with Longer Distance.** Cluster 2 also represents steady following but with a preference for a longer time and distance gap. Suggesting drivers maintain a greater buffer while still closely matching the lead vehicle's speed.

- **Cluster 0 (Purple): Lead Vehicle Driving Away Behavior.** Cluster 0 primarily represents a scenario where the lead vehicle is driving away. This is evident from the "Distribution of Speed_Diff Across HMM States" plot, where the speed difference for Cluster 0 is predominantly negative and the preceding vehicle distribute around 20 – 40 m/s. This indicates that the preceding vehicle's speed is generally larger than the ego vehicle's speed.
- **Cluster 3 (Red): Unsteady car following Behavior** While not explicitly display any feature , typically represents scenarios where the ego vehicle is adjusting speed since the speed difference distribution is larger, which indicate a non steady following behavior.

Then, we compared the aggregated data from HMM Clusters 1 and 2 (representing steady car-following) with data classified as steady car-following by a pre-defined rule-based model. Figure 4.2 illustrates the distribution differences across various features between these two classification methods.

From the figure 4.2 we can see a significant overlap in the distributions of distance headway, time headway, ego vehicle Velocity, Lead vehicle Velocity, Speed Difference between ego and lead vehicle, and ego vehicle Acceleration.

Quantitatively, we calculate that:

- Percentage of data points that cluster as steady following by HMM model that

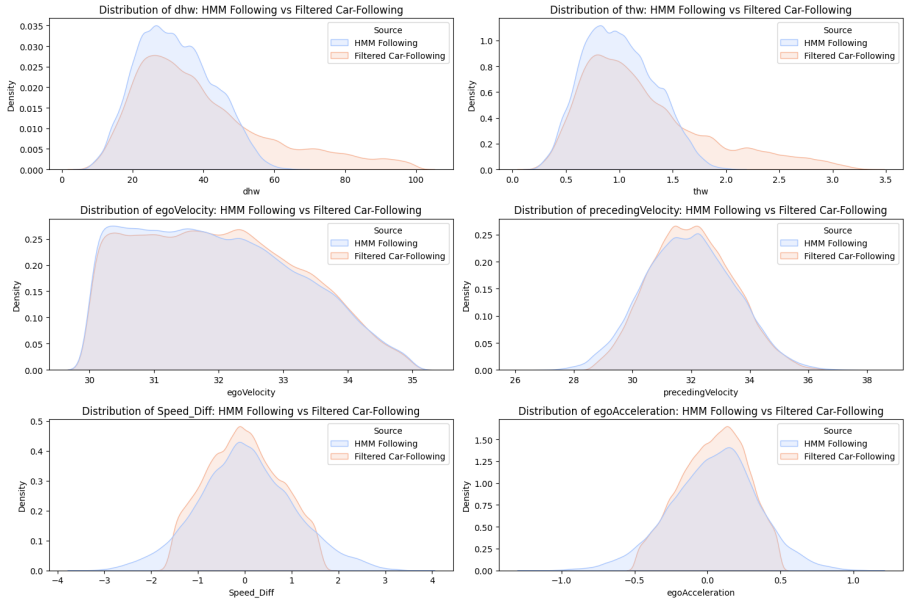


Figure 4.2: Distribution Comparison: HMM-identified Car-Following vs. Rule-Based Filtered Car-Following for 20 – 25 m/s group

also meet the rule based model car-following criteria: 52.50%.

- Percentage of data point that cluster as steady car-following by rule based model that are also cluster as steady car following by HMM model: 70.36%.

The high degree of overlap between these two groups, particularly the substantial portion of rule-based car-following points captured by the HMM, verifies that the rule-based model is effective and its classification of steady car-following is reasonable. This demonstrates the consistency between a data-driven HMM approach and a more traditional rule-based classification for identifying fundamental driving patterns.

Similarly we examined the feature distributions across identified HMM clusters for the 30 – 35 m/s speed group. The result can be found in Appendix A

After examine the result for all the traffic group, we verifies that the rule-based model is effective and reasonable in filtering out steady car-following data. One interesting finding is that the ego vehicle acceleration criteria need to be loosen to 0.5 m/s^2 for speed higher than 30 m/s for in high speed, the vehicle maneuvers is more responsive. This process not only validates the rule-based approach but also provides valuable insight into car-following behavior within the highD dataset, particularly how the parameters for "steady" following may adapt with increasing speed.

4.2 Multi-factor Importance Ranking

4.2.1 Feature importance on public dataset

From the exiD dataset, we selected steady car-following data collected on the lane adjacent to the on-ramp area. Both Gradient Boosting and XGBoost regression

Speed Range	R² Train	R² Test	MSE Train	MSE Test
20 – 25 m/s	0.70	0.68	0.21	0.22
25 – 30 m/s	0.96	0.94	0.043	0.057
30 – 35 m/s	0.98	0.97	0.016	0.022

Table 4.1: Gradient Boosting Regression – exiD (lane adjacent to on-ramp) Evaluation Metrics

Speed Range	R² Train	R² Test	MSE Train	MSE Test
20 – 25 m/s	0.99	0.99	0.0012	0.0074
25 – 30 m/s	0.99	0.99	0.001	0.004
30 – 35 m/s	0.99	0.99	0.001	0.004

Table 4.2: XGBoosting Regression – exiD (lane adjacent to on-ramp) Evaluation Metrics

models were applied to evaluate feature importance. Their evaluation metrics are summarized in Tables 4.1 and 4.2, while the corresponding feature importance results are shown in Figure 4.3.

Table 4.1 shows the Gradient Boosting results. The test R^2 ranges between 0.68 – 0.97 across speed intervals, with a corresponding MSE between 0.22 and 0.022. By contrast, Table 4.2 shows that XGBoost attains higher coefficients of determination ($R^2 \approx 0.99$) and substantially lower mean squared errors (MSE < 0.01). R^2 measures the proportion of variance explained by the model, while MSE quantifies the average squared difference between predicted and observed values. This indicates that XGBoost provides a more stable and accurate fit to the data. Although XGBoost models often risk overfitting due to their high flexibility, in our case this risk is not high. The training and test R^2 are close to 0.99 and the consistently low MSE across speed ranges indicate that the model generalizes well in this dataset. The relatively weaker performance of Gradient Boosting, especially in the 20 – 25 m/s range, suggests that the model struggles with the larger variability in contextual factors at lower speeds.

In this analysis, the choice of regression model does not substantially affect the permutation importance results. Since both regressors yield similar outcomes, we present the XGBoost regression as a representative example.

After filtering, the dataset contained 355 unique tracks in the 20 – 25 m/s range, 27 in the 25 – 30 m/s range, 28 in the 30 – 35 m/s range, and only 5 in the 35 – 40 m/s range. Results for the highest speed range were excluded due to the very small sample size, which would make the metrics unreliable. The differences in sample sizes also help explain why Gradient Boosting performance fluctuates more strongly than XGBoost because small datasets amplify instability in models that lack strong regularization.

Figure 4.3 shows the feature importance for car-following behavior in exiD dataset, grouped by different speed intervals. The results indicate that traffic density is the most important factor among the two velocity intervals. The number of surrounding vehicles is also influential, particularly in the 25–30 m/s group. Traffic-related

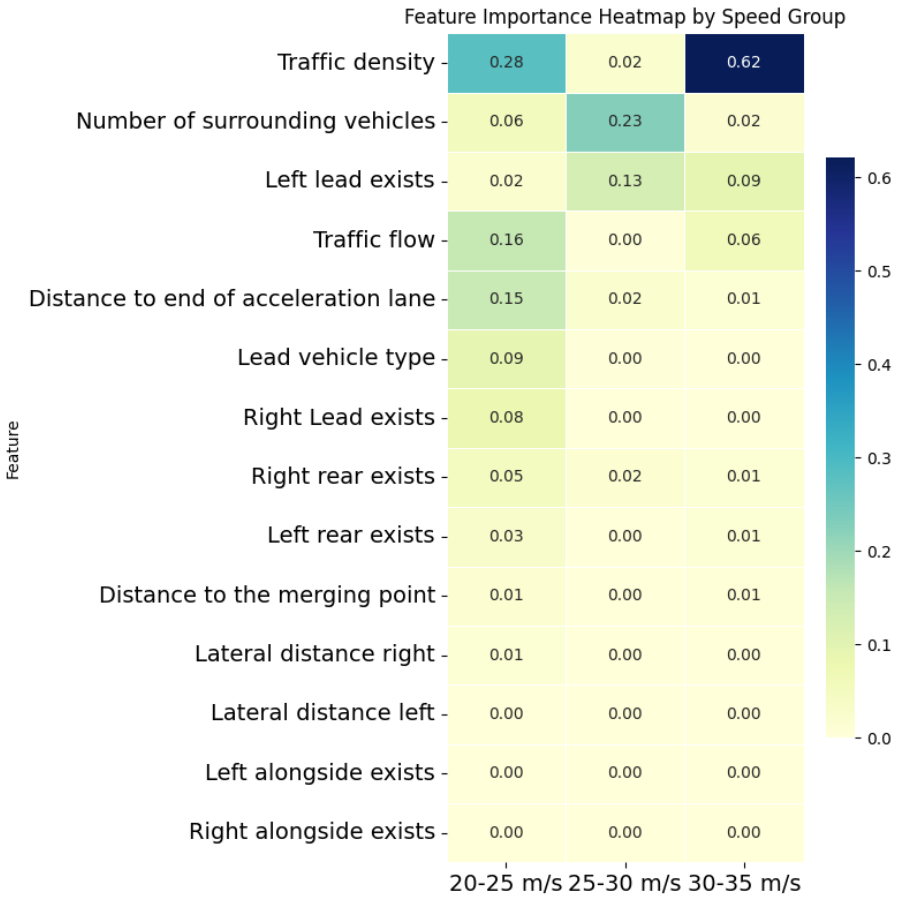


Figure 4.3: Feature Importance Heatmap – Volvo Cars sample dataset, German Highway (3 Lanes, XGBoosting)

Speed Range	R² Train	R² Test	MSE Train	MSE Test
20 – 25 m/s	0.67	0.67	0.34	0.33
25 – 30 m/s	0.52	0.52	0.31	0.31
30 – 35 m/s	0.37	0.37	0.28	0.29
35 – 40 m/s	0.75	0.73	0.098	0.10

Table 4.3: Gradient Boosting Regression – German Highway Lane 3 Evaluation Metrics

Speed Range	R² Train	R² Test	MSE Train	MSE Test
20 – 25 m/s	0.98	0.96	0.016	0.022
25 – 30 m/s	0.96	0.95	0.020	0.028
30 – 35 m/s	0.97	0.97	0.018	0.022
35 – 40 m/s	0.99	0.98	0.0048	0.0072

Table 4.4: XGBoosting Regression – German Highway Lane 3 Evaluation Metrics

factors, Traffic density, traffic flow and number of surrounding vehicles are the core contextual factors.
Figure 4.3 (based on XGBoost) highlights the relative contribution of contextual features, with velocity and relative velocity intentionally excluded to focus on external factors. Across speed intervals, traffic density emerges as the most influential feature, followed by the number of surrounding vehicles, particularly in the 25 – 30 m/s group. This finding is consistent with the intuitive role of traffic flow conditions in shaping car-following dynamics. Distance to the end of the acceleration lane has moderate influence at lower speeds, reflecting the localized effect of on-ramp infrastructure. Vehicle type and the presence of right-lane vehicles show limited importance, though their effects appear more noticeable at lower speeds, which motivates future research on explicit merging scenarios. Rear and alongside vehicle features generally rank very low, likely due to limited representation in the dataset.
In this part, XGBoost regression provides more reliable performance across heterogeneous conditions, making it the preferred choice for subsequent modeling tasks. The dataset-specific map definitions support infrastructure-oriented analysis. Therefore, the results of the ranking motivate further single-factor analyses and investigation of on-ramp effects.

4.2.2 Feature importance on Volvo Cars sample dataset

From the Volvo Cars sample dataset, we selected steady car-following data collected on a German highway with three lanes. This setup is comparable to the public exiD dataset, which was also recorded on a three-lane German highway (include one acceleration lane), ensuring that results can be compared without confounding differences in lane number or country-specific traffic regulations. Both Gradient Boosting and XGBoost regression models were applied to evaluate feature importance, and their evaluation metrics are summarized in Tables 4.3 and 4.4.
Table 4.3 shows that Gradient Boosting regression exhibits modest performance, with test R^2 values ranging from 0.37 to 0.73 depending on speed interval. The rela-

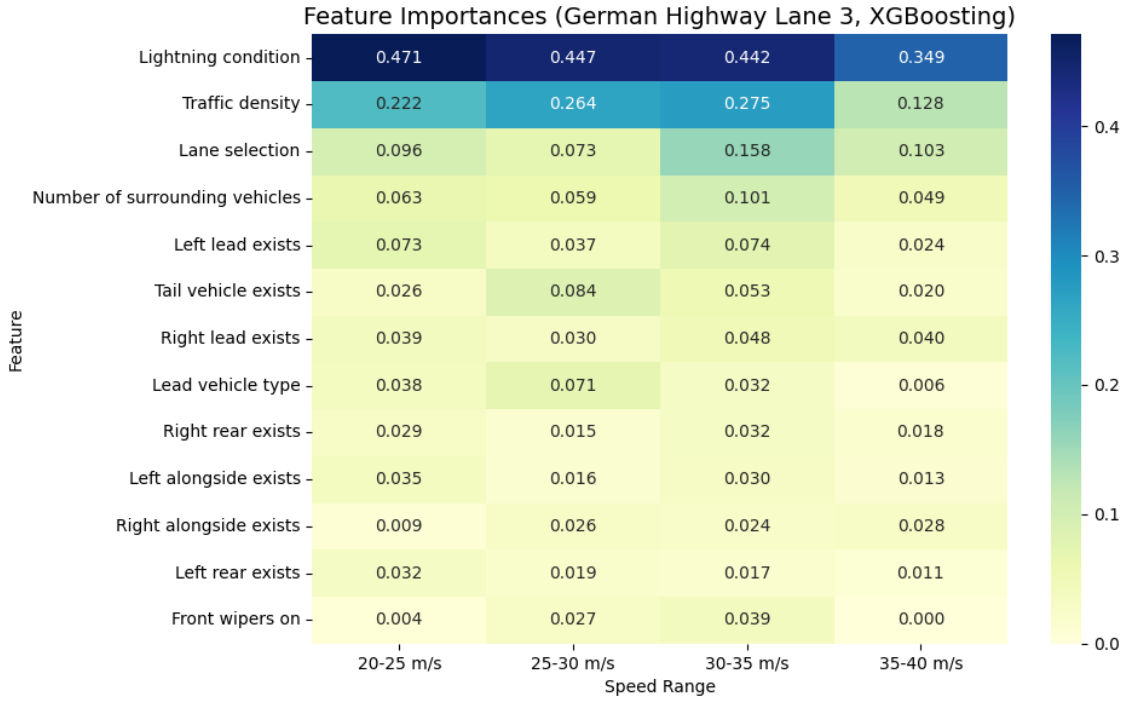


Figure 4.4: Feature Importance Heatmap – Volvo Cars sample dataset, German Highway (3 Lanes, XGBoosting)

tively low R^2 at intermediate speeds (30–35 m/s) and higher MSE values (0.28–0.29) suggest that the model struggles to generalize in these ranges. In contrast, Table 4.4 shows that XGBoost achieves higher predictive accuracy, with R^2 values above 0.95 across all speed intervals and very low MSE values (all below 0.03). The close agreement between training and test R^2 in the XGBoost results indicates that overfitting is not a concern here.

The flickering pattern of MSE in the Gradient Boosting results (e.g., slightly lower error at $20 - 25$ m/s and $35 - 40$ m/s, but much higher in the mid-speed intervals) may be explained by data variability: intermediate speeds are associated with diverse traffic conditions and possible lane-changing events, which are harder for regression models to capture reliably.

Although all data belong to the free-flow regime, driving behavior is not equally homogeneous across speed intervals. As Treiber and Kesting show in their stability analysis, free-flow stability increases with larger headways, since small perturbations are more likely to decay [42]. This is consistent with our findings: at higher speeds ($35 - 40$ m/s), vehicles maintain longer gaps and more uniform behavior, leading to higher R^2 and lower MSE compared to intermediate speed intervals.

Figure 4.4 display the heatmap of feature importance ranking of contextual factors with Volvo Cars sample dataset. It shows that lighting condition is the most influential factor for car-following behavior across all speed ranges. Traffic density also shows high importance, which is in line with expectations. The number of surrounding vehicles, although related to local traffic density, appears less important than the overall traffic density. One possible reason is that the range of surrounding vehicles is zero to six, which may not fully capture differences in heavy traffic conditions.

Lane selection is another important factor, indicating that the chosen lane has a significant effect on time gap behavior. This finding is also supported by our further analysis on lane-specific patterns.
In contrast, the type of lead vehicle has lower importance than expected. This may be due to data imbalance, as there are few instances of trucks in the lead vehicle position. The feature indicating whether the front wipers are on also has little influence, likely because of the limited number of such samples.
With feature importance scores for surrounding vehicles, we observe that the lead vehicle in adjacent lanes generally has a greater influence on time gap behavior than vehicles located behind or alongside the ego vehicle. Rear vehicles also show a more important influence in the speed range of 25 – 30 m/s. Vehicles alongside the ego vehicle show much lower importance across all speed ranges, suggesting that drivers primarily focus on the behavior of vehicles in front and the rear vehicle in the same lane when adjusting their time gap, rather than those behind or next to them. This pattern is consistent across different speed intervals.

4.2.3 Overall results from the factor importance analysis

Across both datasets, traffic density consistently emerges as the most influential contextual factor affecting car-following time gaps, regardless of speed interval. The number of surrounding vehicles also plays a secondary role, although its influence is weaker than that of overall traffic density. Lane-related variables are important in Volvo Cars sample dataset, reflecting the impact of lane choice on driver behavior. In the exiD dataset, the research is done on a certain lane adjacent to the on-ramp, so the presence of right-side vehicles are more relevant, highlighting merging effects in that environment. Other factors, such as lead vehicle type or weather-related indicators (e.g., wiper status), show limited influence, likely due to data imbalance or small sample sizes.
These results suggest that car-following behavior is shaped primarily by traffic conditions and road context, with dataset-specific effects reflecting the on-ramp infrastructure. This provides a basis for selecting key factors in further scenario-based modeling and for interpreting differences in driver behavior between settings.
However, the dataset is imbalanced in two aspects. First, vehicles following a truck are much fewer than those following a car, which may lead to an underestimation of the influence of lead vehicle type. Second, some vehicles contribute disproportionately long car-following durations, giving their behavior more weight in the training process. Both issues suggest that the relative importance of certain factors should be interpreted with caution, as they may be either underestimated or overrepresented due to the data distribution.

4.3 Lead Vehicle Type (Car or Truck)

Here we investigate the influence of lead vehicle type on car-following time gaps. For this analysis, we specifically filtered traffic segments on motorways where the lead vehicle was either a car or a truck, within the predefined speed category. Furthermore, to simplify the analysis and ensure consistency, we restricted the data to

car-following instances occurring in the rightmost/outermost lane.

4.3.1 Analysis of Volvo Cars sample dataset

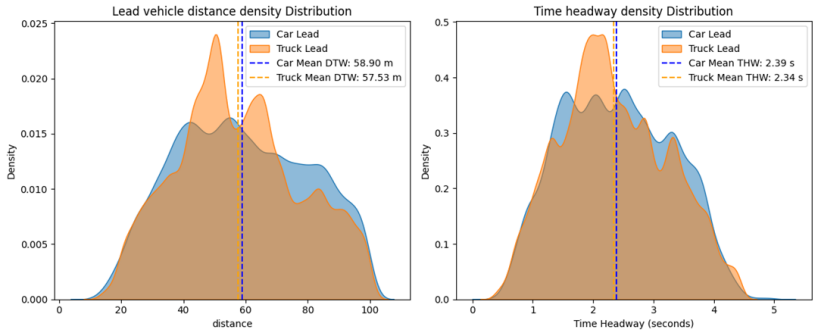
Figure 4.5 compare lead Vehicle Distance and Time Headway Distributions by Lead Vehicle Type for Different Speed Ranges for Volvo Cars sample dataset. We can see the difference between speed range varies across different speed group. In general, The gap is not significant at the low speed group, then we can observe Time headway different at median speed group, then the difference again disappear on the high speed group.

(1) 20 – 25 m/s (Fig. 4.5a): It is evident that the DHW and THW distributions for car-lead and truck-lead vehicles are highly overlapped, with no statistically significant difference can be observed. In this speed traffic is relatively low-speed and potentially slightly congested, human drivers do not significantly distinguish between car and truck leads and they do not exhibit a preference for maintaining a longer time gap when following a truck. This behavior might stem from the lower speeds offering more time for reaction, reducing the perceived risk associated with following larger vehicles, or due to the prevailing traffic density at these speeds constraining drivers' ability to maintain larger gaps.

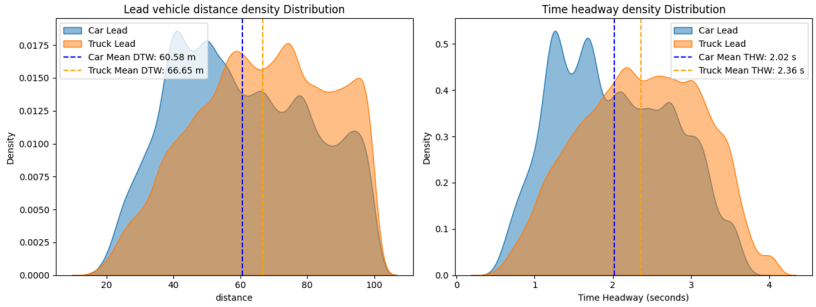
(2) 25 – 30 m/s (Fig. 4.5b): The mean THW for the truck-lead group is approximately 0.3 s longer than the mean THW for the car-lead group. This divergence is expected as traffic conditions generally become less congested at these slightly higher speeds, allowing drivers more flexibility in choosing their following distances. Human drivers tend to prefer leaving a longer time gap when following a truck due to cautious reasons: trucks often have a larger blind spot, their braking capabilities can differ significantly from passenger cars, and their sheer size can obscure visibility of the road ahead, all of which necessitate a greater safety margin and more time to react to sudden events.

(3) 30 – 35 m/s (Fig. 4.5c): It can be seen that the mean THW for the truck-lead group is approximately 0.15 s longer than that for the car-lead group. This further supports the reasoning that human drivers, for safety considerations, prefer to maintain a slightly longer time gap when following a truck, especially as speeds increase. The larger mass, potential for varied braking dynamics, and reduced forward visibility associated with trucks all contribute to this cautious behavior, allowing for more reaction time.

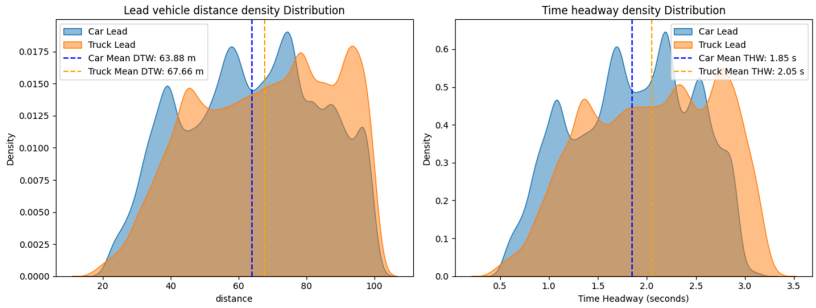
(4) 35 – 40 m/s (Fig. 4.5d): The THW distributions for car-lead and truck-lead groups overlap again and their mean THW values are relatively similar. It is important to note that at high speeds, the number of observed car-following scenarios and the data volume is considerably lower compared to other traffic groups, which might influence the statistical significance of subtle differences. The convergence of THW preferences for car and truck leads at high speeds can be explained by several factors. In well-structured highway environments at higher velocities, traffic conditions are often very good, with minimal congestion. This reduced complexity and the absence of foreseeable emergency braking scenarios or sudden traffic disruptions might lead drivers to feel safer maintaining similar time gaps regardless of the lead vehicle type. The high speeds themselves might also naturally encourage a



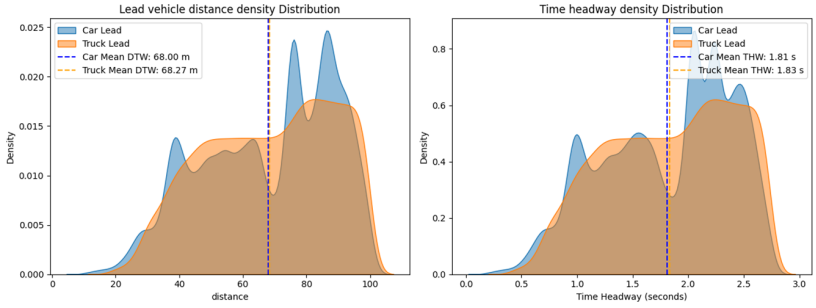
(a) 20 – 25 m/s, outermost lane



(b) 25 – 30 m/s, outermost lane



(c) 30 – 35 m/s, outermost lane



(d) 35 – 40 m/s, outermost lane

Figure 4.5: Comparison of Lead Vehicle Distance and Time Headway Distributions by Lead Vehicle Type for Different Speed Ranges on Motorways by Volvo Cars sample dataset

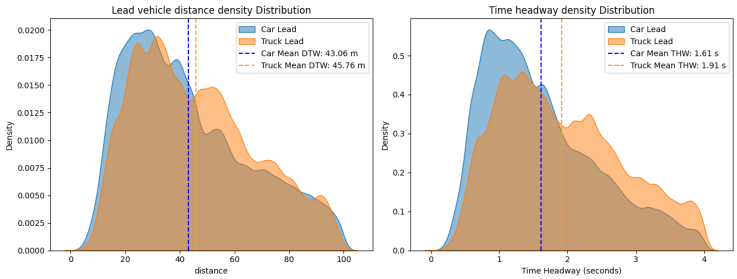
more uniform, often shorter, gap preference as drivers maintain a steady flow, with the perceived risk associated with truck size being less pronounced than at lower, potentially more dynamic, speeds.
Overall, these findings suggest that for an ACC system aiming to better mimic human driving preferences, it would be advisable to maintain a time gap that is approximately 0.1 to 0.3 s longer when following a truck compared to following a car, particularly in medium-speed traffic.

4.3.2 Analysis of HighD Dataset

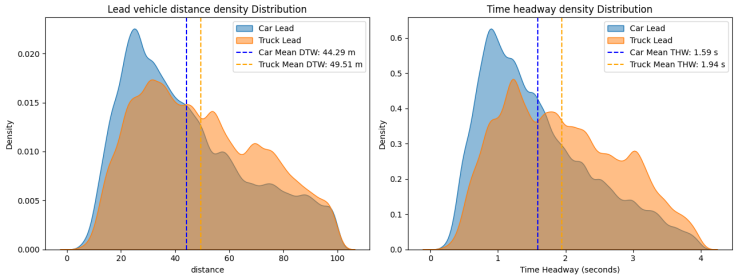
Figure 4.6 illustrates the lead vehicle distance and time headway distributions for car and truck leads across different speed ranges in the outermost lane of the HighD dataset.
Unlike trends observed in the Volvo Cars sample dataset, the findings from the HighD dataset regarding lead vehicle type reveal some consistent characteristics across different traffic speed groups: Firstly, there isn't a significant change across different speed groups in the HighD dataset. The distributions for different speed ranges and mean time headway look relatively similar for both car and truck leads. Secondly, THW for car-following scenarios is consistently smaller than for truck-following scenarios. Specifically, the time headway for the car lead group is approximately 0.3 to 0.5 s smaller than the truck lead group across all examined speed ranges. Thirdly, as the speed range increases, the mean time headway for car leads tends to become slightly smaller, similar to the finding on the Volvo Cars sample dataset. For truck leads, the trend is less pronounced, with mean THW values remaining relatively the same.
This behavior, particularly the relatively stable THW across speed groups and the smaller gap to cars, can be attributed to the driving culture and infrastructure characteristics of German highways, where the HighD dataset was collected. Drivers on these highways often exhibit more unified and highly adapted following behaviors, especially given the prevalence of higher speeds and the emphasis on maintaining flow.
Overall, even though the time headway distribution differ from different dataset and speed group, and the mean time headway between truck and car can differ from almost 0 to 0.5 s, we can observe the general preference to leave a longer time gap when following a heavier vehicle (truck) compared to a car, likely driven by safety considerations due to longer braking distances and reduced visibility behind trucks, which demonstrate that lead vehicle type does have a consistent influence on time headway and it's recommend to include lead vehicle influence on car following time gap in ACC design.

4.4 Infrastructure: Number of Lanes and Lane selection

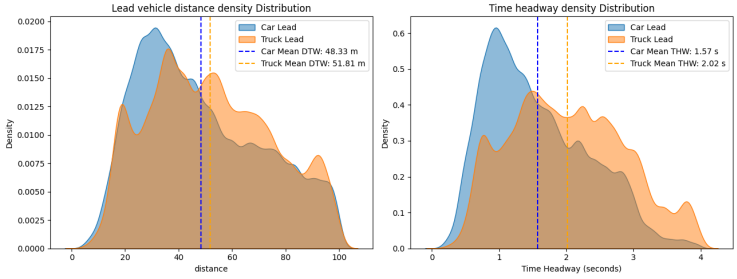
Here we present our results on the Number of Lanes and lane selection utilizing highD and the Volvo Cars sample dataset. The section is divided in analysis on 2



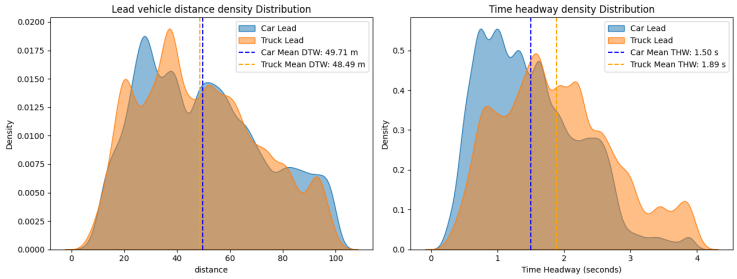
(a) 20 – 25 m/s, outermost lane.



(b) 25 – 30 m/s, outermost lane.



(c) 30 – 35 m/s, outermost lane.



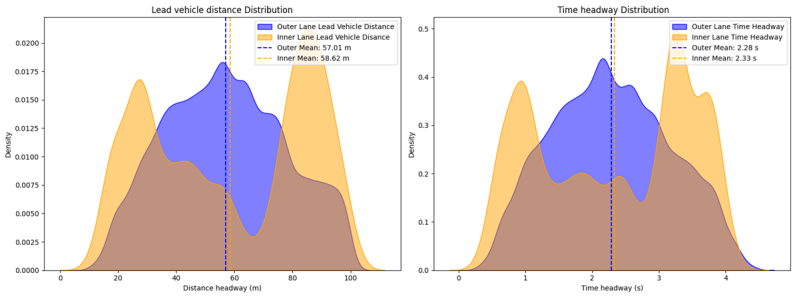
(d) 35 – 40 m/s, outermost lane.

Figure 4.6: Comparison of Lead Vehicle Distance and Time Headway Distributions by Lead Vehicle Type (Car or Truck) for Different Speed Ranges on Motorways by HighD dataset

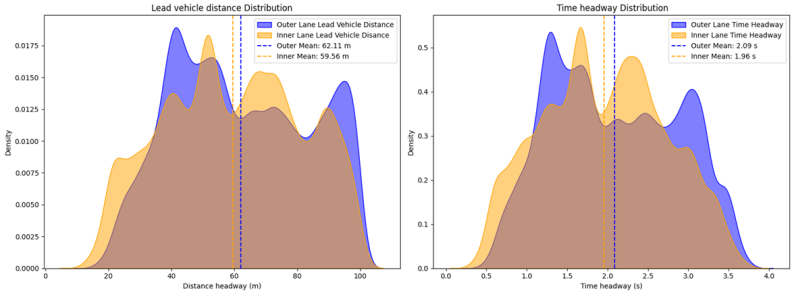
lanes and 3 lanes conditions. We applied the same filter criteria to both dataset as the lead vehicle type study.

4.4.1 2 Lanes

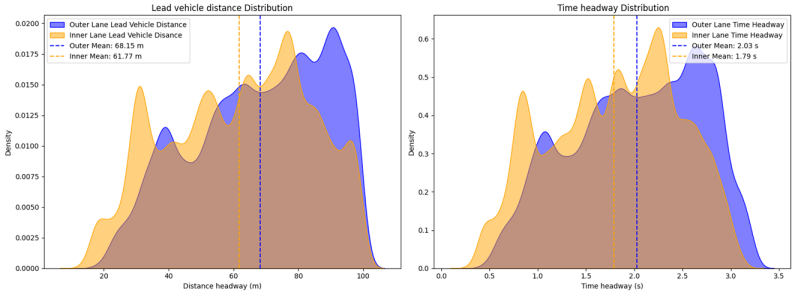
Here we will present our findings on the influence of the number of lanes and the ego vehicle's lane selection on car-following time gaps on a 2-lane motorway using the Volvo Cars sample dataset first.
(1) 20 – 25 m/s (Fig. 4.7a): We can see that the THW distributions for both the outer and inner lanes are highly overlapping, and their mean THW values are nearly identical. This similarity suggests that under low-speed, the overall traffic dynamics and perceived road capacity are relatively consistent across both lanes of a 2-lane motorway. Drivers' preferences for time gaps do not significantly differentiate based on lane choice in such environments where the limited road capacity often dictates more uniform following behaviors.
(2) 25 – 30 m/s (Fig. 4.7b): We can see the different between the inner and outer lane group. This differentiation is expected as the inner lane typically offers better traffic conditions and is often perceived as reserved for higher-speed travel or overtaking maneuvers, leading drivers to maintain shorter time gaps. Conversely, heavier vehicles like trucks often prefer to use the outer lanes, contributing to different driving dynamics and preferences in THW. Drivers in the inner lane might prefer slightly shorter time gaps due to a higher expectation of consistent flow and fewer large vehicles.
(3) 30 – 35 m/s (Fig. 4.7c): We can observe the discernible gap between the THW distributions of the outer and inner lane groups. This slightly wider difference compared to the 25-30 m/s group is expected and reinforces the previous reasoning: as speeds increase, drivers in the inner lane often aim for more efficient, higher-speed travel, leading to preferences for shorter time gaps, while outer lanes may accommodate a mix of vehicle types and driving intentions.
(4) 35 – 40 m/s (Fig. 4.7d): The THW distributions for the outer and inner lanes become highly overlapped again, with very close mean THW values. This convergence can be attributed to the extremely good traffic conditions typically encountered at such high speeds on motorways. Factors like slower trucks or varying overtaking behaviors, which may differentiate lane preferences at lower speeds, tend to have a less pronounced influence. Drivers in both lanes likely feel safe to maintain similar, relatively consistent, and optimal time gaps without significant traffic disruptions or perceived risks.
Here we present our findings on the influence of lane selection for 2-lane roads by the HighD dataset. Figure 4.8 displays the lead vehicle distance and time headway distributions for both outer and inner lanes across different speed ranges on a 2-lane motorway segment in the HighD dataset.
We can observe that: firstly, across speed ranges, the overall shape and characteristics of the time headway distributions for both inner and outer lanes remain relatively consistent similar as what we seen in the lead vehicle type study 4.3. This suggests a stable driving culture within these speed bands on German highways. An exception is observed in the highest speed range (35 – 40 m/s), particularly for



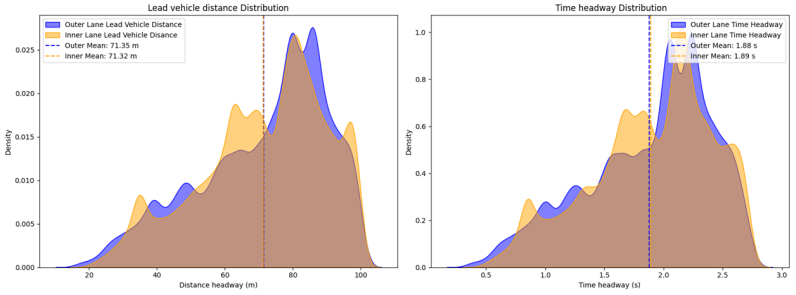
(a) 20 – 25 m/s, 2-lane motorway.



(b) 25 – 30 m/s, 2-lane motorway.

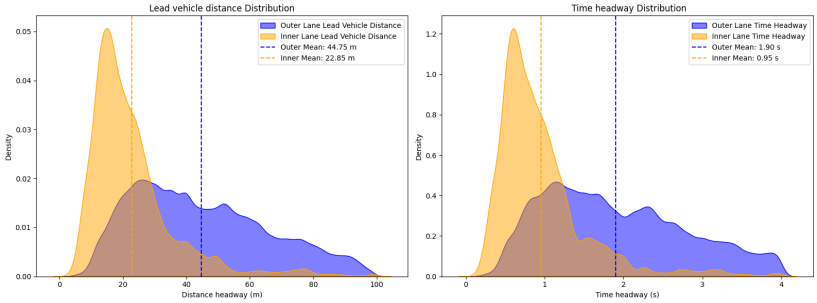


(c) 30 – 35 m/s, 2-lane motorway.

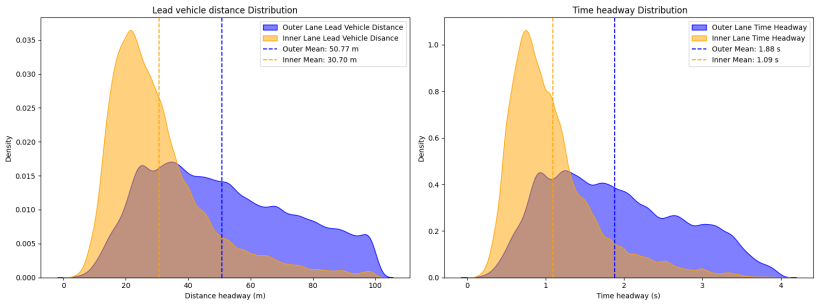


(d) 35 – 40 m/s, 2-lane motorway.

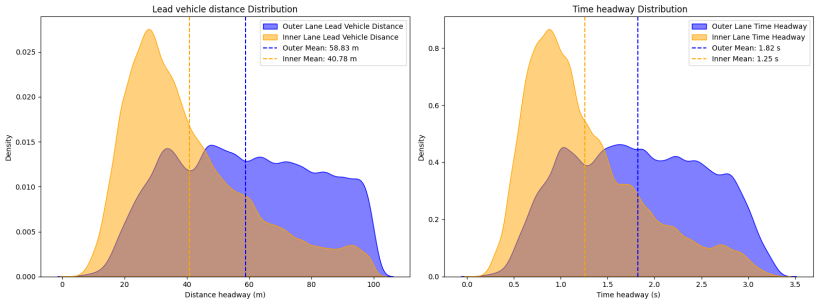
Figure 4.7: Comparison of Lead Vehicle Distance and Time Headway Distributions by Lane for Different Speed Ranges on 2-Lane Motorways by Volvo Cars sample dataset



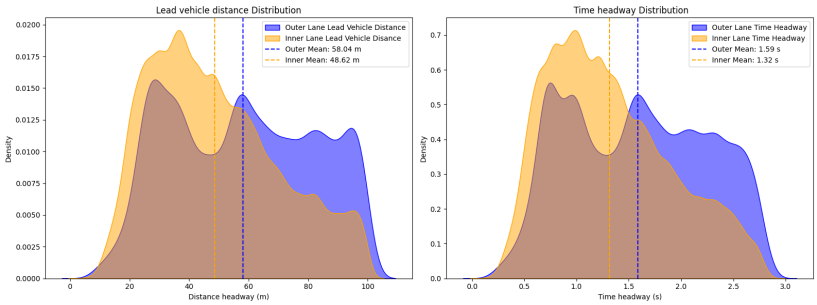
(a) Lead vehicle distance and time headway density distributions for outer and inner lanes (20 – 25 m/s, 2-lane motorway).



(b) Lead vehicle distance and time headway density distributions for outer and inner lanes (25 – 30 m/s, 2-lane motorway).



(c) Lead vehicle distance and time headway density distributions for outer and inner lanes (30 – 35 m/s, 2-lane motorway).



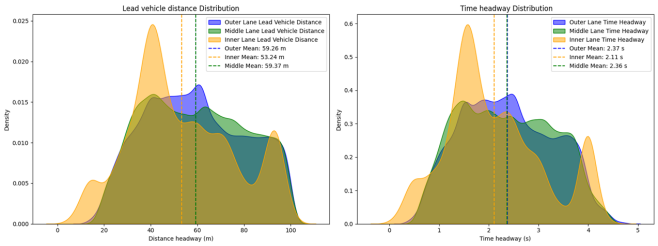
(d) Lead vehicle distance and time headway density distributions for outer and inner lanes (35 – 40 m/s, 2-lane motorway).

Figure 4.8: Comparison of Lead Vehicle Distance and Time Headway Distributions by Lane for Different Speed Ranges on 2-Lane Motorways by highD dataset

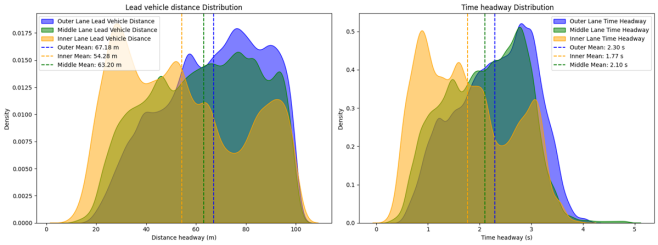
the outer lane. The distribution appears less smooth and more uneven compared to lower speed ranges, since there's a lack of data. Secondly, A striking difference is evident in THW distributions between the inner and outer lanes. The inner lane (typically the faster, overtaking lane) consistently shows a significantly smaller mean THW compared to the outer lane. For instance, at 20 – 25 m/s, the Inner Lane Mean THW is 1.54 s, while the Outer Lane Mean THW is 2.38 s. This considerable difference is attributed to the generally better traffic conditions and the dynamic role of the inner lane, which allows drivers to maintain closer following distances. The outer lane, often used for cruising or by heavier vehicles, exhibits a broader and more diverse distribution of THW choices. This suggests that drivers in the outer lane may have more varied reasons for their chosen headway.
Overall, Consistent with general traffic flow principles and driver expectations, drivers universally prefer to keep a smaller time headway, approximately 0.1 to 0.8 s longer in the inner (faster) lane than in the outer (slower) lane.

4.4.2 3 Lanes

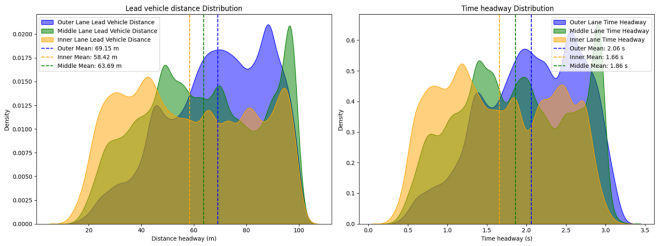
(1) 20 – 25 m/s (Fig. 4.9a): It is important to note that there is a lack of data for the inner lane in this speed group for 3-lane motorways. Despite this, we can still observe a similar trend to that found on 2-lane roads in this low-speed group: the gap between the middle and outer lane THW distributions is relatively small. This suggests that under low-speed or slightly congested conditions, the traffic flow and driver behavior are relatively homogeneous across these lanes, as the road capacity is shared and drivers may not aggressively differentiate their following gaps based solely on lane choice. The limited space and lower speeds constrain the potential for significant behavioral divergence.
(2) 25 – 35 m/s (Fig. 4.9b and Fig. 4.9c): The mean THW values clearly show the progression. As middle lanes accommodate a mix of speeds and lead to intermediate time gaps. Besides, Outer lanes experience heavier vehicle traffic (e.g., trucks) and are more influenced by on- and off-ramp infrastructure, leading drivers to maintain longer, more cautious time headway. The reduced traffic disruptions and more stable flow at these speeds, especially in inner lanes, allow drivers to choose more consistent gaps tailored to the lane's specific function.
(3) 35 – 40 m/s (Fig. 4.9d): It is important to note that at this highest speed range, there is a distinct lack of data, particularly for the outermost lane, which contributes to its unusual and potentially less representative THW distribution shape. However, focusing on the inner and middle lanes where data is more robust, we observe that the THW gap between them begins to close or converge once again. This behavior is consistent with our previous findings for lead vehicle type and 2-lane analysis at very high speeds. Under high-speed traffic conditions, drivers often perceive minimal risk regardless of the lane, leading to a convergence of preferred time gaps. There is less need for exaggerated safety margins, and drivers feel safe to maintain efficient time headway across these lanes, as significant traffic disruptions are less probable.
Overall, we can observe and draw similar findings from both the lead vehicle type analysis and the number of lanes/lane selection study regarding THW preferences.



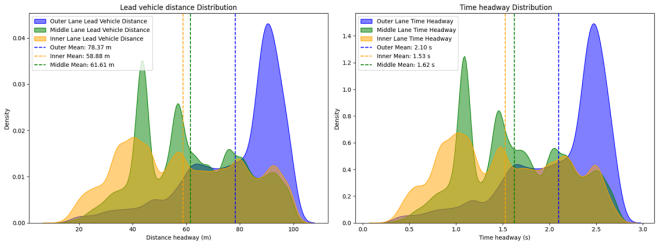
(a) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (20 – 25 m/s, 3-lane motorway).



(b) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (25 – 30 m/s, 3-lane motorway).



(c) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (30 – 35 m/s, 3-lane motorway).



(d) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (35 – 40 m/s, 3-lane motorway).

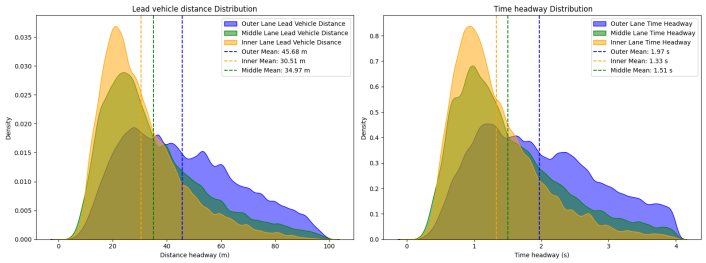
Figure 4.9: Comparison of Lead Vehicle Distance and Time Headway Distributions by Lane for Different Speed Ranges on 3-Lane Motorways (Volvo Cars sample dataset).

The THW gap exhibits a consistent behavioral pattern across different speed categories: the gap between different traffic groups (e.g., inner vs. outer lane; car vs. truck lead) is negligible in very low (20 – 25 m/s) and very high (35 – 40 m/s) speed ranges. However, a significant and clear gap emerges in the intermediate speed ranges (e.g., 25 – 30 m/s and 30 – 35 m/s) where traffic conditions allow for more driver discretion and lane-specific behaviors become more pronounced.
This dynamic relationship highlights that human driver preferences for car-following time gaps behave differently under varying traffic settings, necessitating a nuanced approach to analysis and ACC design. It underscores the importance of carefully constructing scenarios to derive robust conclusions. Based on our analysis, for ACC system design, we suggest incorporating a variable time gap strategy, potentially adding between 0 to 0.7 s of time headway depending on the specific number of lanes, the ego vehicle's lane selection.
Here we present our findings on the influence of lane selection for 2-lane roads by the HighD dataset. Figure 4.10 presents the lead vehicle distance and time headway density distributions for the outer, middle, and inner lanes across various speed ranges on a 3-lane motorway in the HighD dataset.
The analysis of the 3-lane HighD dataset reveals findings consistent with the 2-lane observations that: Firstly, the time headway distributions for the inner, middle, and outer lanes in the 3-lane motorway sections show only small variations across different speed ranges. This again points to a relatively unified and stable driving behavior on German highways, where drivers maintain consistent headway preferences within a given lane, irrespective of marginal speed increases. Secondly, across all examined speed ranges, a clear and increasing difference in mean time headway is observed from the inner lane to the middle lane, reflecting the functional hierarchy of motorway lanes. This clear differentiation indicates that drivers adapt their time headway choices significantly based on the designated purpose and typical traffic flow characteristics of each lane. Drivers maintain smaller distances in faster lanes and progressively larger distances in slower lanes.
Overall, these findings suggest that for the ACC system to better mimic human driving preferences, it's recommended to maintain a time gap that is approximately 0.2 to 0.6 s longer for different lanes.

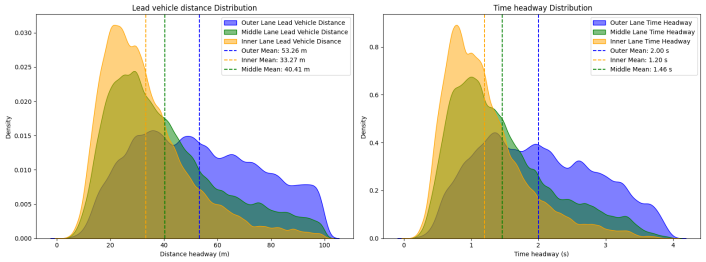
4.5 Lighting Condition

This study specifically leverages the Volvo Cars sample dataset, since it provides a high-quality data for the lighting conditions during each drive, which is not present in the public dataset. This data ranges from 0 to 10000 (arbitrary units). As illustrated in Figure 3.2 in previous sections, a significant amount of the collected data is distributed at the extreme ends of this range. Given this skewed distribution, we examined the influence of lighting conditions on THW from two distinct perspectives:

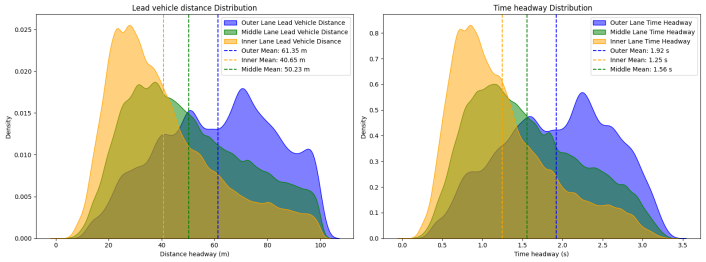
1. We classified traffic data into distinct bright and dark traffic groups to compare their THW distributions.
2. We analyzed the average THW evolution with a granular regard to the continuous range of lighting conditions.



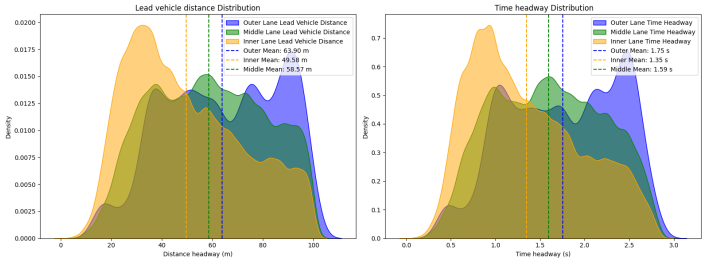
(a) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (20 – 25 m/s, 3-lane motorway).



(b) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (25 – 30 m/s, 3-lane motorway).

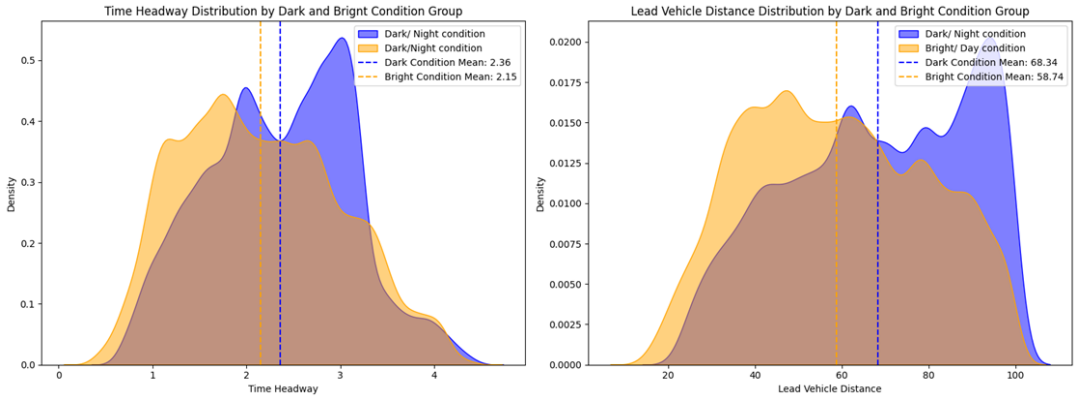


(c) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (30 – 35 m/s, 3-lane motorway).

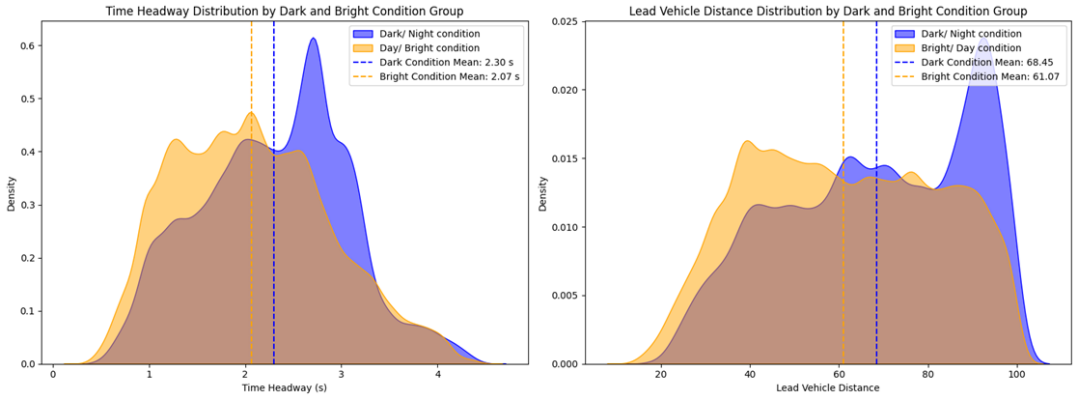


(d) Lead vehicle distance and time headway density distributions for outer, middle, and inner lanes (35 – 40 m/s, 3-lane motorway).

Figure 4.10: Comparison of Lead Vehicle Distance and Time Headway Distributions by Lane for Different Speed Ranges on 3-Lane Motorways by highD dataset



(a) Time Headway and Lead Vehicle Distance Distribution by Dark and Bright Condition Group (Gaussian Filtered Data).



(b) Time Headway and Lead Vehicle Distance Distribution by Dark and Bright Condition Group (Butterworth Filtered Data).

Figure 4.11: Comparison of Time Headway and Lead Vehicle Distance Distributions by Dark and Bright Lighting Conditions, Processed with Gaussian and Butterworth Filters.

4.5.1 Time headway analysis for dark and bright traffic group

Following the signal processing of the lighting condition data, we proceeded to analyze its influence on human drivers' THW preferences. This analysis was conducted using both the Gaussian-filtered and Butterworth-filtered lighting data. Using a similar method like before, we construct two traffic groups: the dark traffic group, referring to drives that occurred under dark conditions (typically at night), and the bright traffic group, referring to drives that occurred under well-lit conditions.

For the data processed with the custom threshold and Gaussian filter, we observed a discernible difference in time headway distributions between the dark and bright lighting conditions. As illustrated in Figure 4.11a, there is approximately a 0.2 s gap in the mean THW between the two groups. Specifically, the time headway distribution for the bright condition is clearly shifted towards smaller values, while the dark condition exhibits a preference for longer time gaps.

This finding aligns with human driving instincts and safety considerations. In dark

conditions, reduced visibility necessitates a longer reaction time and increased safety margins. Drivers instinctively compensate for the diminished visual information by increasing their following distance, thereby allowing more time to perceive and react to potential hazards, such as sudden braking of the lead vehicle or unforeseen obstacles on the road.
Subsequently, we performed a similar comparison using the data processed with the Butterworth low-pass filter. Figure 4.11b presents the time headway and lead vehicle distance distributions for dark and bright conditions based on this filtered data. The distributions derived from the Butterworth-filtered data exhibit a similar pattern. We can draw analogous conclusions regarding driver behavior, where the darker conditions correlate with a preference for longer time headways. This consistency across both filtering methods reinforces the robustness of our finding that lighting conditions significantly influence human drivers' time gap choices, with a clear tendency to increase safety margins when visibility is reduced. The similar outcomes from both filtering approaches validate that the choice of filter, while having minor technical differences as discussed previously, does not fundamentally alter the behavioral insights gained from the dataset.

4.5.2 Analysis of Time Headway Evolution with Lighting Condition

We also examined the continuous relationship between average THW and varying lighting conditions. This provides another perspective on how driver behavior adapts across the full spectrum of ambient brightness.
Figure 4.12 presents a scatter plot depicting the average time headway against the average lighting condition for every 500 data points. This visualization includes a regression line to show the overall trend, a rolling mean to illustrate local variations, and a 95% prediction band to capture the expected range of human driver time headway choices.
As evident from Figure 4.12, despite a clear overarching trend, human drivers exhibit a dynamic and diverse range of time headway choices even under similar lighting conditions. This variability is expected, as numerous other factors influence individual driving styles, including driver personality, psychological state (e.g., fatigue, stress), and the specific traffic context at any given moment (e.g., immediate necessity for acceleration or braking, presence of vulnerable road users). These background traffic conditions frequently necessitate different THW choices, contributing to the broad spread of data points within the prediction band.
Nevertheless, a clear descending trend is observed: as the average lighting condition improves (i.e., brightness increases), the average time headway chosen by drivers distinctly decreases. The regression line visually confirms this inverse relationship. The average/mean THW selection is consistent with our previous findings from the grouped analysis, showing an approximate 0.2 s decrease as conditions transition from dark to bright.
In conclusion, our analysis from two distinct angles—both categorical grouping and continuous trend evaluation—consistently confirms that human drivers prefer to maintain a longer time gap in darker or worsening light conditions. We recommend

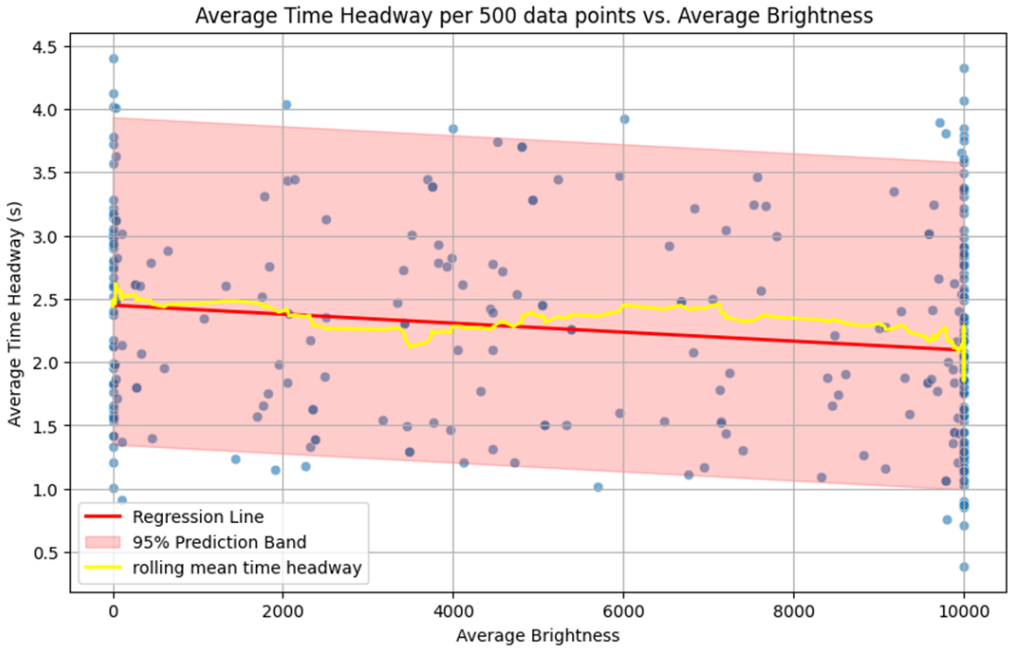


Figure 4.12: Average Time Headway per 500 Data Points vs. Average Brightness, with Regression Line, Rolling Mean, and 95% Prediction Band.

adapting a 0.2 s longer car following time gap for the design of ACC systems.

4.6 Scenarios influenced by infrastructure and surrounding traffic

4.6.1 Scenario 1: Cut-in intention of other vehicles

To investigate the ego vehicle's reaction to cut-in intentions from adjacent vehicles, the first step is to accurately identify periods within the highD dataset where such intentions were present. For this, we employed an HMM to cluster driving behaviors, specifically focusing on detecting when drivers exhibit lateral movements indicative of a cut-in.

The HMM was trained using the following features: longitudinal acceleration (xAcceleration), longitudinal velocity (xVelocity), lateral acceleration (yAcceleration), and lateral velocity (yVelocity). These features comprehensively describe a vehicle's motion and are highly relevant for capturing lane-changing dynamics. Figure 4.13 illustrates the distribution of these features across the different clusters identified by the HMM.

Based on the analysis of these feature distributions, we determined that Cluster 0 (Purple), Cluster 2 (Light Blue), and Cluster 4 (Yellow) are most likely associated with cut-in behaviors, as they predominantly reflect periods where the driver is performing significant lateral movement. While these clusters are distributed differently across the feature space, they all show notable activity in either lateral acceleration (yAcceleration) or lateral velocity (yVelocity), confirming their associ-

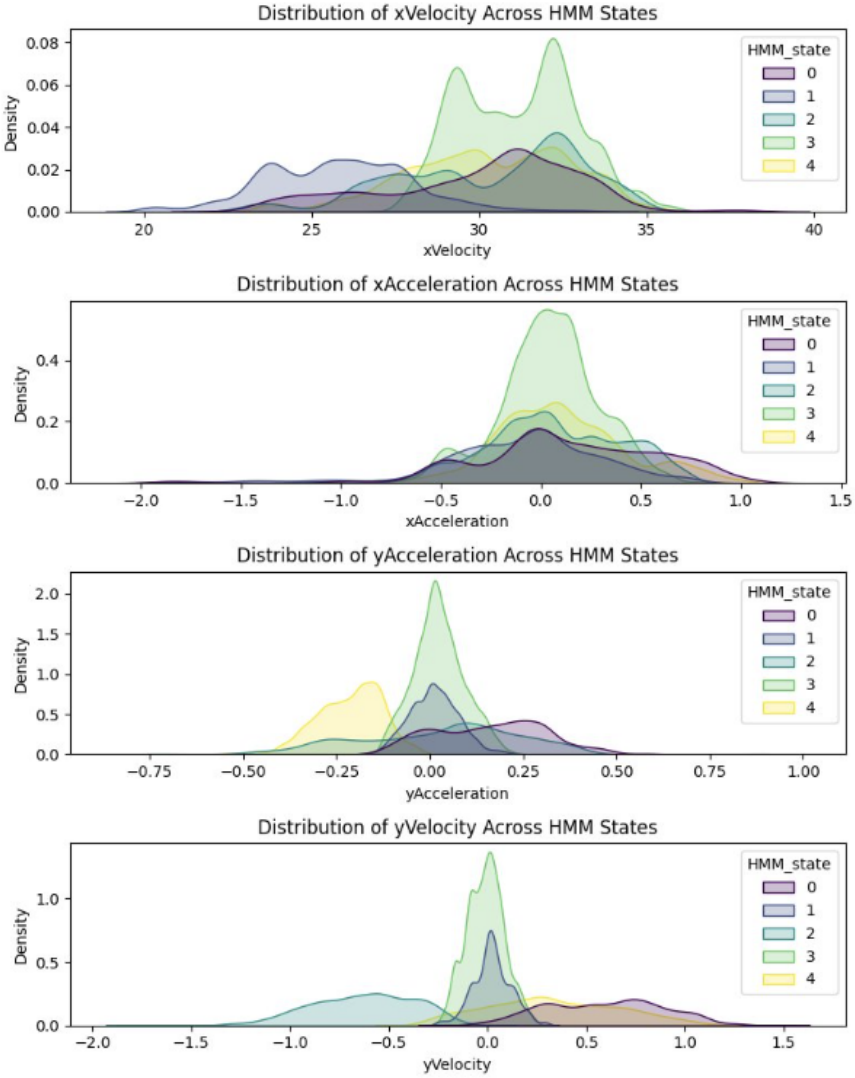


Figure 4.13: Distribution of velocity and acceleration in x and y direction under four HMM states

ation with lane-changing maneuvers.
To further validate these clusters as indicators of cut-in events, we investigated whether all recorded lane change events in the highD dataset were indeed classified into these identified clusters (0, 2, or 4) during the actual cut-in period. The results are overwhelmingly positive:

- Total lane changes identified in the dataset: 1363
- Total lane change events classified into states 0, 2, or 4 by the HMM: 1363
- Percentage of lane changes occurring within states 0, 2, or 4: 100.00%

This 100% capture rate strongly confirms that the HMM effectively identifies all instances of lane changes within these three clusters. Furthermore, the distribution of lane change events among these states provides insight into the prevalence of each specific lateral movement pattern during a cut-in:

- **State 0:** Lane change events: 620 (45.49% of total lane changes)
- **State 2:** Lane change events: 597 (43.80% of total lane changes)
- **State 4:** Lane change events: 146 (10.71% of total lane changes)

Following the identification of cut-in periods using the HMM, our next step was to investigate the reaction of the "being cut in" vehicle during these events. This study focuses on how drivers respond to an encroaching vehicle just before it fully cuts into their lane.
We train a K-Means clustering model on features representing the ego vehicle's braking behavior during these cut-in events: maximum deceleration, deceleration duration, and mean deceleration. We filter out drives where the maximum deceleration is smaller than 0.2 m/s^2, since most cut-ins occur at low speeds, and a previous study determined that accelerations smaller than this threshold represent normal steady car-following behavior. This allows us to categorize different types of driver reactions into three distinct clusters. Figure 4.14 presents a 3D scatter plot visualizing the distribution of these clusters across the three aforementioned features. Each point in the plot represents a distinct driver's reaction instance.
The K-Means clustering resulted in the following distribution of reaction patterns:

- **Cluster 0 (Purple): Short and Intense Braking (**15% **of data points).** This cluster reflects situations where drivers perform sudden and intense deceleration over a short period. Characterized by a deceleration time of 3 s and a high average decelerate speed of 2.29 m/s, this reaction suggests a more urgent or reactive braking response to a perceived immediate threat from the cutting-in vehicle.
- **Cluster 1 (Green): Long and Moderate Braking (**22% **of data points).** Drivers in this cluster engage in a more prolonged, but still moderate, braking action. This pattern is defined by a longer deceleration time of approximately 7.8 s and a decelerate speed of 1.76 m/s. This indicate a more gradual, sustained reaction to a cut-in, or a situation where a longer adjustment period is available.
- **Cluster 2 (Yellow): Short and Moderate Braking (**61% **of data points).** This cluster represents the majority of driver reactions. Drivers in this cluster exhibit short and moderate braking behavior, characterized by a typical deceleration time of 3 s and a relatively low average decelerate speed of 0.72 m/s. This suggests a common, perhaps anticipated, response to a cut-in that allows

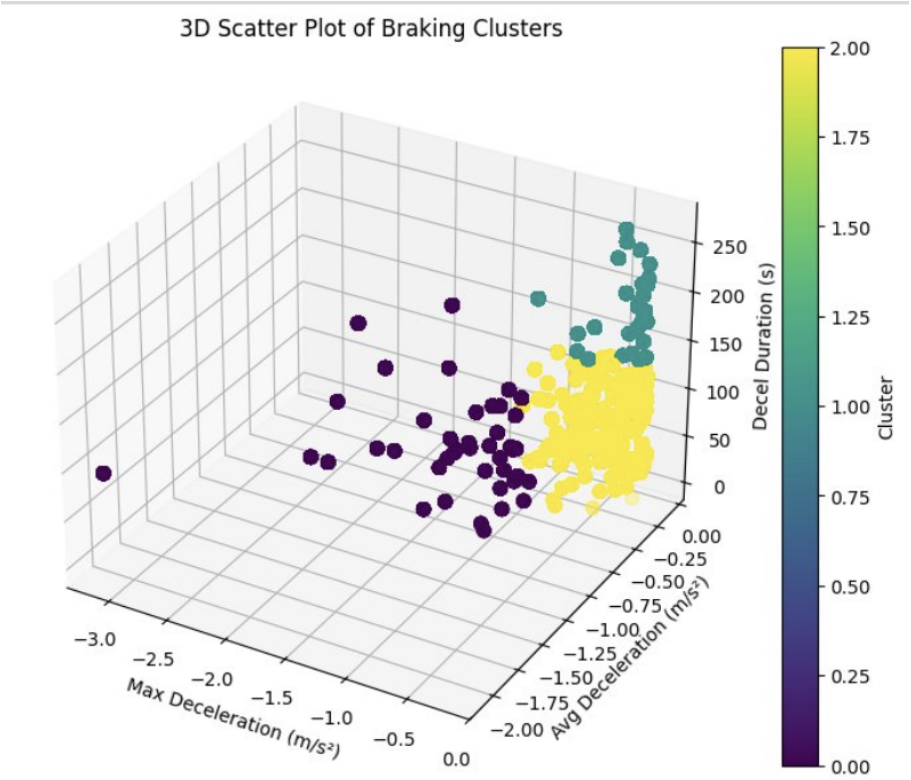


Figure 4.14: 3D scatter plot for different braking clusters across different features

for gentle adjustment.

To understand the underlying situational factors that cause a driver to react in a certain way (i.e., belong to one of the K-Means reaction clusters), we investigated the relationship between various pre-cut-in scenario features and the resulting braking behavior cluster. Our objective was to identify what specific traffic and interaction conditions compel a driver to react in a particular manner.

For this purpose, we trained a Random Forest Classifier using the following scenario-specific features as inputs: Average speed difference between the ego vehicle and the cut-in vehicle, Average distance headway between the ego vehicle and the cut-in vehicle, Average lateral acceleration of the cut-in vehicle, Average lateral velocity of the cut-in vehicle., Traffic flow and traffic density.

The performance of the Random Forest Classifier was robust with 100% precision, recall, and F1-score across all clusters. This high performance suggests a strong relationship between the chosen features and the observed reaction types.

To understand which of these scenario features were most influential in determining the driver's reaction, we analyzed the Feature Importance from the trained Random Forest model:

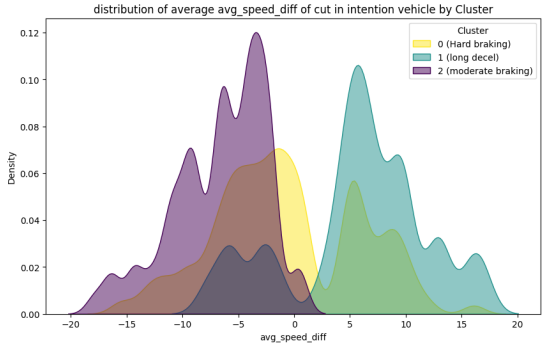
From the table 4.5, it is evident that several factors related to the immediate interaction with the cutting-in vehicle have a large influence on the driver's reaction behavior: average speed difference, average distance headway, average lateral acceleration, and lateral velocity.

to gain a deeper understanding of how these scenario factors influence specific driver reactions, we visualize the distributions of the most influential features for each

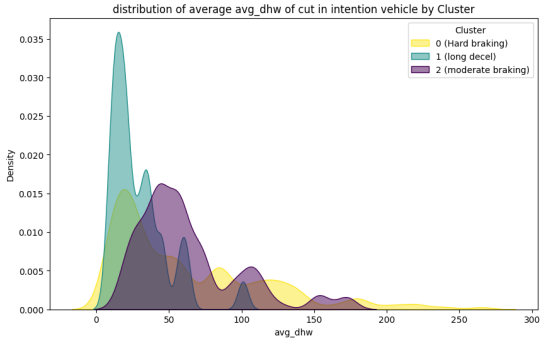
Feature	Importance Score
Average Speed Difference (m/s)	0.2694
Average Lateral Velocity (m/s)	0.2589
Average Lateral Acceleration (m/s^2)	0.2187
Average Distance Headway (m)	0.1913
Density (vehicles/km)	0.0487
Flow (vehicles/hour)	0.0131

Table 4.5: Feature Importance from Random Forest Classifier

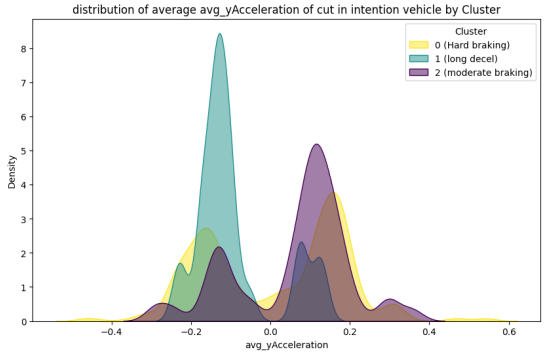
braking cluster. This allows for a direct comparison of the conditions under which different reaction types occur. We particularly focus on Cluster 0 (Hard Braking), as it indicates an intensive reaction likely driven by immediate safety concerns, which is critical for ACC to understand and adapt to. Below Figure show the distributions of average speed difference, average distance headway, average lateral acceleration, and average lateral velocity for the cut-in intention vehicle.



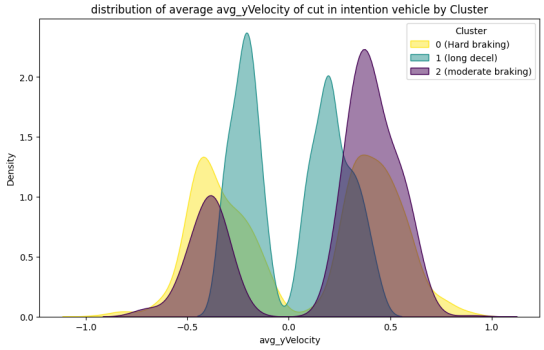
(a) Speed difference density distribution between cut-in intention vehicle and the ego vehicle across different clusters



(b) Average distance headway distribution between cut-in intention vehicle and the ego vehicle across different clusters



(c) Lateral average acceleration distribution for cut-in intention vehicle across different clusters



(d) Lateral average velocity distribution for cut-in intention vehicle across different clusters.

Figure 4.15: Behavioral features of the cut-in intention vehicle clustered by driver reaction.

By examining these distributions from Figure 4.15, we can clearly observe the distinct conditions that lead to each reaction cluster.

- **Cluster 0:(Hard Braking - Yellow)** : For hard braking, drivers tend to react in this manner when the cutting-in vehicle's speed is significantly smaller than the ego vehicle's speed by Figure 4.15a ,when the distance between the ego vehicle and the cutting-in vehicle is small by Figure 4.15b, when the cutting-in vehicle performs aggressive and abrupt lateral movements since the distributions in both Acceleration in Fig 4.15c and Velocity in Figure 4.15dby show peaks at larger absolute values. This is reasonable because these factors combine to present a dangerous scenario: a sudden and aggressive lateral movement of a cutting-in vehicle into a small gap, coupled with a high relative speed, leads to a short time-to-collision or rapid decrease in headway. Such conditions contribute to an intense braking response from the ego vehicle driver to ensure safety and avoid potential collisions.
- **Cluster 1 (Long and Moderate Braking - Green):** Drivers tend to exhibit this type of reaction when the average distance headway is relatively small, but the cutting-in vehicle's speed is higher than the ego vehicle's speed. While the distance headway might be small, the positive speed difference indicates a diverging scenario with little collision possibility. In this case, drivers are likely braking not due to an immediate collision threat, but rather to increase their following distance and adjust to a comfortable and safe headway with the now faster lead vehicle.
- **Cluster 2 (Short and Moderate Braking - Purple):** This represents the most frequent reaction. Drivers tend to perform short and moderate braking when the distance headway is relatively large, and the cutting-in vehicle's speed is smaller than the ego vehicle's speed, indicating that the ego vehicle needs to slow down. However, with a generous distance headway, the driver has plenty of room and time to adjust comfortably, allowing for a less aggressive and more controlled deceleration.

In conclusion, it is recommended that the ACC algorithm introduce the braking mechanism for when the there's a cut-in intention vehicle, the cut-in vehicle speed is significantly smaller than the ego vehicle, and the distance between the ego vehicle and the cut-in vehicle is small.

4.6.2 Scenario 2: Passing on-ramp area without merging vehicles

Lane change behavior analysis For this study, we selected a map from the exiD dataset that contains an on-ramp, enabling a detailed investigation of the infrastructure on car-following behavior. Within this map, $5,453$ vehicles were analyzed. Among these, 62.3% ($3,399$ vehicles) maintained their lane throughout the observed trajectory, 27.6% ($1,505$ vehicles) executed a single lane change, and 10.1% (549 vehicles) performed multiple lane changes.

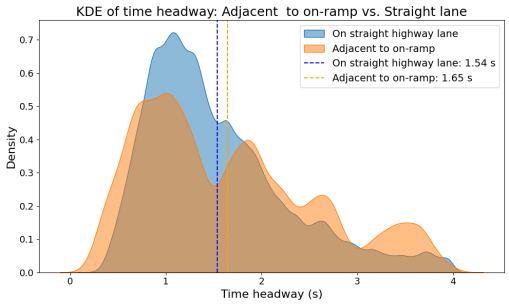
As shown in Figure 3.11, the on-ramp lane refers to the lane containing the acceleration lane; lane 0 is the lane directly adjacent to the on-ramp, and lane 1 is the innermost lane adjacent to lane 0.

	Number of tracks	**Time headway mean** (s)	**Velocity mean** (m/s)	**Acceleration mean** (m/s^2)
Straight lane	542	1.54	24.03	0.09
Adjacent to on-ramp	12	1.65	24.95	0.13

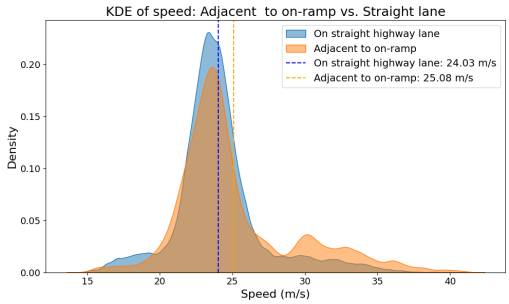
Table 4.6: Comparison of mean values in Scenario 2

For vehicles entering from the on-ramp, all merged into one of the two main lanes, with the predominant maneuver being a merge from the on-ramp into lane 0 within the acceleration lane region.
For vehicles starting in lane 0, 82.1% maintained their lane, 15.0% made a single lane change, and 3.0% made multiple changes. The most frequent maneuver was a shift from lane 0 to lane 1 during the acceleration lane section, likely to avoid merging traffic from the on-ramp.
For vehicles starting in lane 1, 86.2% remained in the same lane, 13.5% made a single lane change, and 0.3% made multiple changes. The lower incidence of lane changes compared with lane 0 suggests a reduced direct influence from the on-ramp. The most common maneuver from lane 1 was a merge into lane 0 shortly after the end of the acceleration lane, possibly to position for an upcoming exit or due to local traffic flow patterns.
A small number of vehicles performed multiple lane changes within the approximately 1 km study segment. Notably, some vehicles from the on-ramp merged into inner lanes twice, while 24 vehicles starting in lane 0 shifted into an inner lane during the acceleration lane and then returned to lane 0.
These observations point to further discussion on the on-ramp area: Impact of the on-ramp on lane 0 behavior. The existence of on-ramp increases the likelihood of lane changes during this lane segment, either to accommodate merging vehicles or to avoid potential slowdowns.

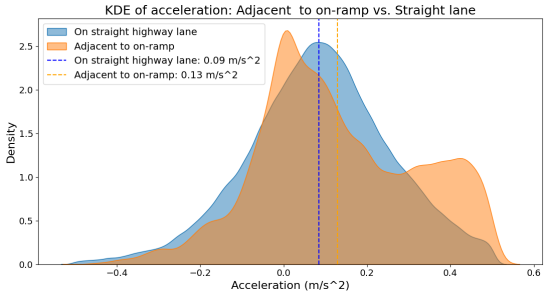
Scenario 2 analysis When drivers pass an on-ramp section and no vehicle is present in the adjacent acceleration lane, the observed differences in time headway between the straight lane and the lane adjacent to the on-ramp are statistically significant ($p < 0.001$), but the magnitude is small (0.08s) and unlikely to have a practical impact. The 80% HDI for vehicles in the straight lane is 0.5 ˘ 2.2 s, while for vehicles in the lane adjacent to the on-ramp it is 0.3 ˘ 2.5 s, which has been shown in Table 4.6.
As shown in Figure 4.16, although the mean time headway in the on-ramp-adjacent lane is slightly longer (1.65 s vs. 1.54 s), speed and acceleration distributions remain comparable across the two lane types. This suggests that in the absence of merging activity, the influence of the on-ramp is minimal, and drivers maintain car-following patterns largely consistent with those in uninterrupted lanes. The small observed difference may be attributed to a cautious driving tendency in proximity to an on-ramp, even when no merging occurs.



(a) Time headway



(b) Speed



(c) Acceleration

Figure 4.16: Distributions of (a) time headway, (b) speed, and (c) acceleration for vehicles on straight lanes and on lanes adjacent to on-ramps. Each subplot compares the two lane types.

4.6.3 Scenario 3: Passing on-ramp area with merging vehicles

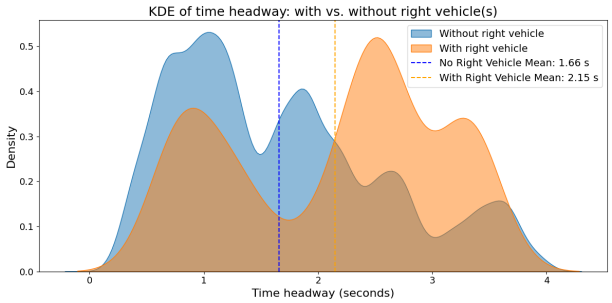
When vehicles are present in the adjacent acceleration lane, drivers in the on-ramp-adjacent lane adopt longer time headways, with a mean difference of 0.4 s compared to the "without right vehicle" case ($p < 0.001$). The 80% HDI shifts upward, reflecting a consistent pattern of increased spacing. Additionally, mean acceleration decreases (0.04 m/s^2 vs. 0.13 m/s^2), indicating milder longitudinal control in the presence of potential merging, which has been shown in Table 4.7.

As shown in Figure 4.17, this anticipatory spacing behavior appears regardless of whether the vehicle in the acceleration lane actually merges ahead of the ego vehicle. The adjustment is likely a proactive safety measure, allowing space for a possible merge and reducing the need for abrupt deceleration. The lower average speed in the "with right vehicle" condition may also reflect a conscious moderation of pace to maintain this buffer.

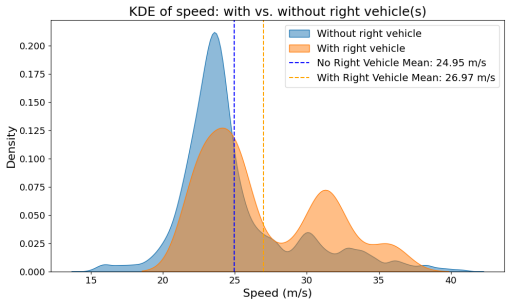
Overall, the comparison between Scenario 2 and Scenario 3 demonstrates that the presence of vehicles in the adjacent acceleration lane is the primary driver of behavioral change in this context.

	Number of tracks	**Time headway mean** (s)	**Velocity mean** (m/s)	**Acceleration mean** (m/s^2)
"Without" right vehicle	542	1.65	26.97	0.13
"With" right vehicle	12	2.15	24.95	0.04

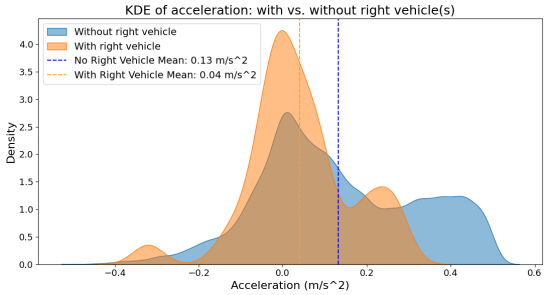
Table 4.7: Comparison of mean values in Scenario 3



(a) Time headway



(b) Speed



(c) Acceleration

Figure 4.17: Distributions of (a) time headway, (b) speed, and (c) acceleration for vehicles in On-Ramp Adjacent Lane (exiD): With vs. Without Right Vehicle(s). Each subplot compares the two lane types.

In this section, we presented the core findings derived from the analysis of both the public datasets and the Volvo Cars sample dataset. Our analysis began with a comprehensive examination of factor importance. Then we quantified the individual influence of specific factors. We found that drivers consistently adjust their time gaps based on lead vehicle type, number of lanes and lane selection, and lightning conditions. Furthermore, our scenario-based analysis provided crucial insights into adaptive behavior in complex situations. We confirmed that human drivers proactively adjust their time gaps in response to a detected cut-in intention and also adapt their following distance when passing an on-ramp with merging vehicles.

5
Conclusion

In this concluding chapter, we summarize the main findings of the thesis, discuss their limitations and implications, and outline promising directions for future research. Finally, we reflect on the ethical considerations relevant to the use of large-scale driving trajectory datasets.

5.1 Summary of Findings

This thesis conducts a comprehensive comparative analysis of how contextual factor influence car following time gap, leveraging data from both a publicly available dataset and a proprietary dataset from Volvo Cars. By applying consistent research methodologies across both sources, we aimed to gain a deeper understanding of driving dynamics in various contexts, providing insights for the development of the ACC System. Our work involved establishing robust car-following filters, analyzing the importance of diverse influencing factors, examining the specific impact of factors such as lead vehicle type, number of lanes, and lighting conditions on driver behavior and investigate specific scenario where the time gap is influenced by infrastructure and road user.
Our comparative study revealed several key similarities and differences between the datasets. Methodologically, common car-following filtering techniques were applied across both datasets. The primary findings of this study offer valuable insights into human driving behavior and actionable recommendations for ACC systems:

- **Refined Car-Following Definition**: Our pattern recognition approach for identifying "steady car-following" behavior led to a recommended refinement in acceleration criteria for rule-based models, suggesting a transition from 0.2 m/s^2 to 0.5 m/s^2 in higher speed for better characterization.
- **Key Influencing Factors Identified**: The analysis of feature importance across datasets underscored the significant role of traffic density in shaping car-following behavior. Furthermore, lighting conditions were confirmed as a crucial factor, particularly in the Volvo Cars sample dataset. This highlights the importance of environmental factor influence in car following time gap. The study also emphasized the critical need to investigate the influence of specific geographical contexts, such as on-ramp areas, on adjacent lane behaviors.
- **Quantified Driver Preferences**: Our research quantified several consistent human driving preferences:
  - Drivers were found to consistently prefer to maintain a longer time gap

(approximately 0.3 – 0.5 s more) when following a truck compared to a car, indicating a higher safety margin for larger vehicles.
  - A clear preference for smaller time headway was observed in the inner (faster) lane compared to the outer (slower) lane, reflecting adaptive behavior to lane function and traffic flow.
  - Under darker light conditions, drivers demonstrated a tendency to maintain a longer time gap (approximately 0.2 s more), suggesting an increased cautionary approach in reduced visibility.
- **Scenario where the time gap is influenced by infrastructure and road user**: We identified and investigated specific scenarios:
  - Drivers were found to have a tendency to adjust to a approximately 0.1-0.3 s longer time gap and react proactively when a clear cut-in intention is detected from an adjacent vehicle, especially when that vehicle's speed is significantly lower than the ego vehicle's and the lateral distance is minimal.
  - Drivers tended to maintain a time gap approximately 0.5 s longer and exhibited proactive reactions when the ego vehicle was passing an on-ramp area where vehicles might be merging from the acceleration lane, thereby increasing safety during these common conflict points.

Our study highlights that car-following time gaps are influenced by a range of contextual factors. Across both the public and Volvo Cars datasets, we found that lead vehicle type, lane configuration, on-ramp infrastructure, lighting conditions, and surrounding vehicles' merging intentions all play a significant role. Drivers tended to extend their following distance in more demanding conditions, when following trucks, in darker environments, when adjacent vehicles displayed cut-in intentions, and when passing an on-ramp area with vehicles in the on-ramp area.
In summary, this thesis provides a robust, data-driven framework for understanding complex human car-following behavior across diverse driving contexts. The insights gained from both public and Volvo datasets offer a valuable foundation for developing more intelligent, context-aware, and human-centric ACC systems.

5.2 Discussion and Implications

While the findings of this thesis provide valuable insights into manual car-following behavior and offer practical recommendations for ACC time gap adaptions, several limitations must be noted.
First, the analysis focuses exclusively on passenger cars and does not include other vehicle types such as trucks or vans; therefore, the findings may not generalize to other classes of vehicles. Second, the public datasets used in this research (highD and exiD) are limited in spatial coverage (typically 400 – 500 m), which may restrict the ability to capture long-term driving patterns of individual drivers. In addition, the Volvo Cars sample dataset consists of test drives rather than fully naturalistic driving, which may introduce bias in driver behavior.
Despite these limitations, the results offer several interpretations. Firstly, by analyzing both the public dataset and the Volvo Cars sample dataset, we found that they present both consistent and unique results, offering a more complete view of

car-following behavior. For instance, both datasets confirmed that speed and traffic density are primary determinants of the time gap. However, the Volvo dataset, with its richer contextual data, allowed us to quantify the significant influence of factors like lighting conditions and traffic density in a way that the public datasets could not. This demonstrates that a comprehensive understanding of car-following behavior requires data that captures both standard kinematic variables and a wide range of contextual details.
This thesis is grounded in classical car-following perspectives in which headway choice is primarily explained by speed-related factors and safety margins. The evidence here supports that foundation, while other contextual factors are found to influence the time gap choice. First, lead-vehicle class, lane selection, and lighting shift the preferred headway even when speeds are matched. Second, scenario structure matters: anticipation effects appear before a merge is completed. The results argue for extensions of standard car-following theory that (i) incorporate contextual factors as explicit inputs or latent regimes, and (ii) model anticipatory adjustments triggered by impending interactions (e.g., cut-ins maneuvers, ramp proximity). In short, practical ACC design benefits from a theory that is context-aware.

Reflection on theory and implementation. Gradient boosting and XGBoost were suitable for capturing non-linearities and interactions, and they provided good predictive performance with heterogeneous features. Their main limitation is the lack of transparency in how multiple factors interact, and there is also a potential risk of overfitting. In our study, the test results did not indicate overfitting. To reduce bias and improve clarity, the models were trained separately for each speed group, and feature importance was interpreted in a relative rather than absolute sense.
Methods like K-Means for clustering and Random Forest for classification are relatively transparent and interpretable. This transparency is crucial for safety-critical applications. If a system trained on these models causes an accident, a transparent model allows for a clear investigation into the cause.

5.3 Further work

Building upon the insights and methodologies developed in this thesis, several contextual factors can be study to further enrich the understanding of car-following behavior and enhance the development of ACC systems.
Our current scenario studies, investigating cut-in intentions and the influence of on-ramp infrastructure, were primarily conducted and validated using the public dataset due to time constraints. A critical next step would involve implementing and validating these scenario studies on the Volvo Cars sample dataset. This cross-dataset validation would assess the reliability of our findings to different driving locations and data collection environments, thereby strengthening the robustness of our conclusions.
Future work could also expand the scope to include off-ramp infrastructure and cut-out scenarios. While on-ramps and cut-ins represent situations where vehicles

merge into the ego lane's path, off-ramps and cut-outs involve vehicles diverging from the ego lane. Understanding how drivers adjust their time headway and react to diverging vehicles near off-ramps, or when a surrounding vehicle performs a cut-out maneuver, is equally crucial for developing comprehensive and robust ACC functionalities that can handle the full spectrum of merging and diverging traffic interactions.
Another promising direction for future research involves a factor analysis of road curvature impact on car following time gap. This is an especially interesting factor, given its direct relevance to vehicle dynamics and driver comfort in real-world driving. Steady car-following data is naturally more common on straight roads than on curves due to the naturalistic driving dynamics in curvy roads. Consequently, after filtering for steady car-following states in Volvo Cars sample dataset, few data in our dataset have curvature values larger than 0.001 m^{-1}, corresponding to curve radii greater than 1000 m. According to Othman et al. [35], road segments with radii greater than 1250 m are typically considered straight, thus providing no high-quality data on time gap adaptation on highly curvy roads. Future work could focus on understanding human drivers precautionary adaptations before entering the curve. This could involve analyzing how drivers modify their speed and time gap on the straight section of road leading into a curve to prepare for the change in dynamics.

5.4 Ethics

The ethical handling of data from naturalistic driving studies is paramount to ensure that research benefits society without compromising individual privacy. Therefore, both the public trajectory datasets and the Volvo Cars sample dataset we used in the study are processed according to the specific compliance; all records are fully anonymized to remove any personally identifiable information. Data were analyzed only within the approved research scope and were not used to identify individuals. Additionally, the primary objective of our study is to contribute to a deeper understanding of human driving behavior, with the ultimate goal of more intelligent, more human-centric, and more comfortable advanced driver assistance systems. Our findings derived from this work, such as thresholds for human-like following distances, are intended to support the development of the ACC system that enhances ride comfort and user acceptance, rather than being used for commercial exploitation or purposes outside of its intended academic scope.

Bibliography

[1] G. Marsden, M. McDonald, and M. Brackstone, "Towards an understanding of adaptive cruise control," *Transportation Research Part C: Emerging Technologies*, vol. 9, no. 1, pp. 33–51, 2001.

[2] L. Xiao and F. Gao, "A comprehensive review of the development of adaptive cruise control systems," *Vehicle system dynamics*, vol. 48, no. 10, pp. 1167–1192, 2010.

[3] S. F. Varotto, C. Mons, J. H. Hogema, M. Christoph, N. van Nes, and M. H. Martens, "Do adaptive cruise control and lane keeping systems make the longitudinal vehicle control safer? insights into speeding and time gaps shorter than one second from a naturalistic driving study with sae level 2 automation," *Transportation Research Part C: Emerging Technologies*, vol. 141, p. 103 756, 2022.

[4] A. Vahidi and A. Eskandarian, "Research advances in intelligent collision avoidance and adaptive cruise control," *IEEE transactions on intelligent transportation systems*, vol. 4, no. 3, pp. 143–153, 2004.

[5] B. Van Arem, C. J. Van Driel, and R. Visser, "The impact of cooperative adaptive cruise control on traffic-flow characteristics," *IEEE Transactions on intelligent transportation systems*, vol. 7, no. 4, pp. 429–436, 2006.

[6] W. J. Schakel, C. M. Gorter, J. C. F. De Winter, and B. Van Arem, "Driving characteristics and adaptive cruise control: A naturalistic driving study," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 2, pp. 17–24, 2017. DOI: 10.1109/MITS.2017.2666582.

[7] M. Van der Hulst, T. Rothengatter, and T. Meijman, "Strategic adaptations to lack of preview in driving," *Transportation research part F: traffic psychology and behaviour*, vol. 1, no. 1, pp. 59–75, 1998.

[8] S. Ossen and S. P. Hoogendoorn, "Heterogeneity in car-following behavior: Theory and empirics," *Transportation research part C: emerging technologies*, vol. 19, no. 2, pp. 182–195, 2011.

[9] A. Dòria-Cerezo, J. M. Olm, E. Benedito, and D. Biel, "A variable structure-based algorithm for adaptive cruise control," in *2018 15th International Workshop on Variable Structure Systems (VSS)*, 2018, pp. 37–42. DOI: 10.1109/VSS.2018.8460313.

[10] L. Yu and R. Wang, "Researches on adaptive cruise control system: A state of the art review," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 236, no. 2-3, pp. 211–240, 2021. DOI: 10.1177/09544070211019254.

[11] A. Kesting, M. Treiber, M. Schönhof, and D. Helbing, "Adaptive cruise control design for active congestion avoidance," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 6, pp. 668–683, 2008.

[12] P. G. Gipps, "A behavioural car-following model for computer simulation," *Transportation research part B: methodological*, vol. 15, no. 2, pp. 105–111, 1981.

[13] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[14] J. K. Hedrick, D. Mcmahnon, and D. Swaroop, "Vehicle modeling and control for automated highway systems," 1993.

[15] A. Khodayari, A. Ghaffari, R. Kazemi, and R. Braunstingl, "A modified car-following model based on a neural network model of the human driver effects," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 6, pp. 1440–1449, 2012.

[16] E. Kamjoo, R. Saedi, A. Zockaie, M. Ghamami, T. Gates, and A. Talebpour, "Developing car-following models for winter maintenance operations incorporating machine learning methods," *Transportation research record*, vol. 2677, no. 2, pp. 519–540, 2023.

[17] T. T. Zhang, P. J. Jin, S. T. McQuade, A. Bayen, and B. Piccoli, "Car-following models: A multidisciplinary review," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 1, pp. 92–116, 2025. DOI: 10.1109/TIV.2024.3409468.

[18] V. Papathanasopoulou and C. Antoniou, "Towards data-driven car-following models," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 496–509, 2015.

[19] M. Karrouchi, I. Nasri, M. Rhiat, *et al.*, "Driving behavior assessment: A practical study and technique for detecting a driver's condition and driving style," *Transportation Engineering*, vol. 14, p. 100 217, 2023.

[20] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[22] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[23] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[24] Z. Su, Q. Liu, C. Zhao, and F. Sun, "A traffic event detection method based on random forest and permutation importance," *Mathematics*, vol. 10, no. 6, p. 873, 2022.

[25] Y. Zhang, Q. Lin, J. Wang, S. Verwer, and J. M. Dolan, "Lane-change intention estimation for car-following control in autonomous driving," *IEEE*

Transactions on Intelligent Vehicles, vol. 3, no. 3, pp. 276–286, 2018. DOI: 10.1109/TIV.2018.2843178.

[26] Y. Zhang, Y. Zou, Y. Wang, L. Wu, and W. Han, "Understanding the merging behavior patterns and evolutionary mechanism at freeway on-ramps," *Journal of Intelligent Transportation Systems*, vol. 27, no. 5, pp. 573–586, 2023.

[27] B. Higgs and M. Abbas, "Segmentation and clustering of car-following behavior: Recognition of driving patterns," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 81–90, 2015. DOI: 10.1109/TITS.2014.2326082.

[28] R. Yu, Y. Zhang, L. Wang, and X. Du, "Time headway distribution analysis of naturalistic road users based on aerial datasets," *Journal of Intelligent and Connected Vehicles*, vol. 5, no. 3, pp. 149–156, 2022. DOI: 10.1108/JICV-01-2022-0004.

[29] A. K. Maurya, S. Das, S. Dey, and S. Nama, "Study on speed and time-headway distributions on two-lane bidirectional road in heterogeneous traffic condition," *Transportation Research Procedia*, vol. 17, pp. 428–437, 2016, International Conference on Transportation Planning and Implementation Methodologies for Developing Countries (12th TPMDC) Selected Proceedings, IIT Bombay, Mumbai, India, 10-12 December 2014, ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2016.11.084. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146516306998.

[30] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st international conference on intelligent transportation systems (ITSC)*, IEEE, 2018, pp. 2118–2125.

[31] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, and L. Eckstein, "The exid dataset: A real-world trajectory dataset of highly interactive highway scenarios in germany," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2022, pp. 958–964.

[32] A. Loulizi, Y. Bichiou, and H. Rakha, "Steady-state car-following time gaps: An empirical study using naturalistic driving data," *Journal of advanced transportation*, vol. 2019, no. 1, p. 7 659 496, 2019.

[33] J. Han, X. Wang, and G. Wang, "Modeling the car-following behavior with consideration of driver, vehicle, and environment factors: A historical review," *Sustainability*, vol. 14, p. 8179, 2022. DOI: 10.3390/su14138179.

[34] H. Kita, "Effects of merging lane length on the merging behavior at expressway on-ramps," *Transportation and Traffic Theory*, pp. 37–51, 1993.

[35] S. Othman, R. Thomson, and G. Lannér, "Identifying critical road geometry parameters affecting crash rate and crash type," in *Annals of advances in automotive medicine/annual scientific conference*, vol. 53, 2009, p. 155.

[36] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[37] J. Schlechtriemen, F. Wirthmueller, A. Wedel, G. Breuel, and K.-D. Kuhnert, "When will it change the lane? a probabilistic regression approach for rarely

occurring events," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 1373–1379. DOI: 10.1109/IVS.2015.7225907.

[38] X. Qiu, Y. Pan, M. Zhu, L. Yang, and X. Zheng, "Driving style-aware car-following considering cut-in tendencies of adjacent vehicles with inverse reinforcement learning," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 1329–1336. DOI: 10.1109/IV55156.2024.10588574.

[39] Q. Xue, K. Wang, J. J. Lu, and Y. Liu, "Rapid driving style recognition in car-following using machine learning and vehicle trajectory data," *Journal of Advanced Transportation*, vol. 2019, p. 9 085 238, 2019.

[40] B. Filzek and B. Breuer, "Distance behavior on motorways with regard to active safety: A comparison between adaptive-cruise-control (acc) and driver," in *International Technical Conference on Enhanced Safety of Vehicles*, SAE Technical Paper 2001-06-0066, 2001, p. 8.

[41] Y. Li, Y. Liu, D. Ni, A. Ji, L. Li, and Y. Zou, "A re producible approach to merging behavior analysis based on high definition map," *arXiv preprint arXiv:2303.11531*, 2023.

[42] M. Treiber and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin, Heidelberg: Springer, 2013, ISBN: 978-3-642-32459-8. DOI: 10.1007/978-3-642-32460-4.

A

Appendix 1

In here, we present one more group of result for Car following pattern recognition. Similarly we examined the feature distributions across identified HMM clusters for the 30-35 m/s speed group.

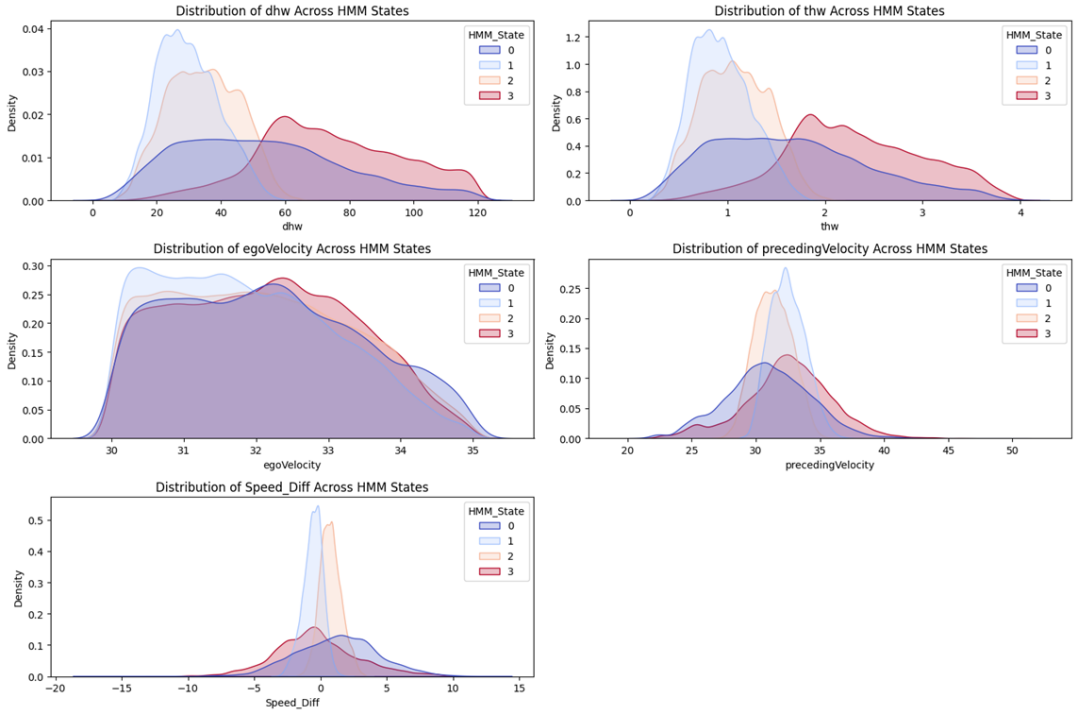


Figure A.1: Distribution of Distance Headway, Time Headway, Ego Velocity, Preceding Velocity, and Speed Difference Across HMM States for 30-35 m/s Speed Group (highD dataset).

our analysis indicates the following interpretations for the clusters:

- **Cluster 1 & Cluster 2: Steady Car-Following Behavior** These two clusters consistently represent steady car-following behavior within this higher speed range.The preceding vehicle speed within these clusters is typically around 30-35 m/s. So the ego vehicle and lead vehicle have similar overall velocity allowing for stable relative velocities. The speed difference for both Cluster 1 and Cluster 2 is largely concentrated around 0, indicating that the ego and lead vehicles maintain very similar speeds.
  - **Cluster 1 (Light Blue)** reflects steady car follow with lead vehicle speed slightly higher than ego vehicle

  - **Cluster 2 (Orange)**, while also representing steady following, reflects steady car follow with lead vehicle speed slightly smaller than ego vehicle
- **Cluster 0 (Purple) : Lead vehicle Driving away behavior** The cluster represent more dynamic and transient car-following states. A cluster where the speed difference is negative indicates a "driving away" scenario, where the preceding vehicle is accelerating relative to the ego vehicle, or the ego vehicle is decelerating.
- **Cluster 3 (Red): Ego vehicle Closing in behavior** Conversely, a cluster with a positive speed difference indicate a "closing in" scenario, where the ego vehicle is gaining on the lead vehicle.

To validate the HMM-based classification for the 30-35 m/s speed group, we again compared the combined data from HMM Clusters 1 and 2 (representing steady car-following) with data classified as steady car-following by the rule-based model for the 30-35 m/s group Fig A.2

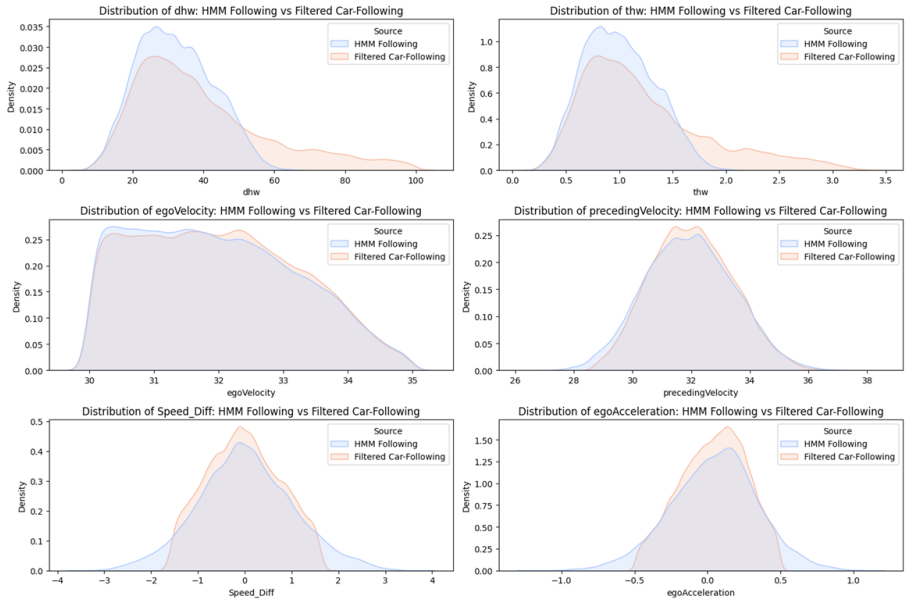


Figure A.2: Distribution Comparison: HMM-identified Car-Following vs. Rule-Based Filtered Car-Following for 20- 25 m/s group

A key adjustment made for this higher speed group was loosening the criterion for ego acceleration in the rule-based model from 0.2 m/s^2 to 0.5 m/s^2. This adjustment was necessary because at higher speeds, vehicle adjustments (even during "steady" following) often involve larger magnitudes of acceleration or deceleration due to increased speed and the need for more responsive maneuvers to maintain a stable state.

Quantitatively, we observed that:

- Percentage of data points that cluster as steady following by HMM model that also meet the rule based model car-following criteria: 57.98%.
- Percentage of data point that cluster as steady car-following by rule based model that are also cluster as steady car following by HMM model: 57.22%.

B

Appendix 2

In here, we present our result for signal factor study for Lead vehicle type, Number of Lanes and Lane selection, Lighting Condition using our majority selection method to these distributions reveals the range of preferred THWs Base on Volvo Cars sample dataset.

B.1 Lead vehicle type

Figure B.1 and Figure B.2 display the majority (90%) THW sections for car-leads and truck-leads, respectively, in 20-25 m/s speed category from Volvo Cars sample dataset.

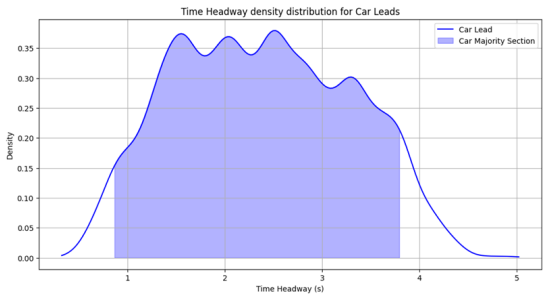


Figure B.1: Time headway density distribution for Car Leads, showing majority section (20-25 m/s, Volvo Cars sample dataset).

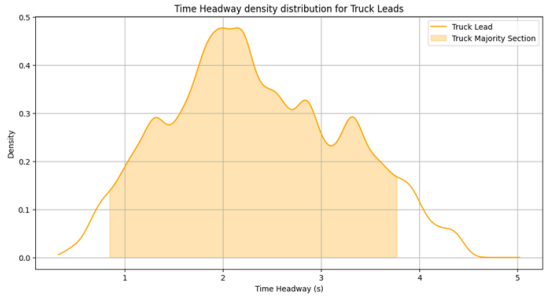


Figure B.2: Time headway density distribution for Truck Leads, showing majority section (20-25 m/s, Volvo Cars sample dataset).

As seen in Figures B.1 and B.2, the time headway range preferred by human drivers is quite wide within this speed category, encompassing values from approximately 0.87 s to 3.8 s for car leads and 0.84 s to 3.77 s for truck leads, reflecting a diverse set of following time gap preferences.

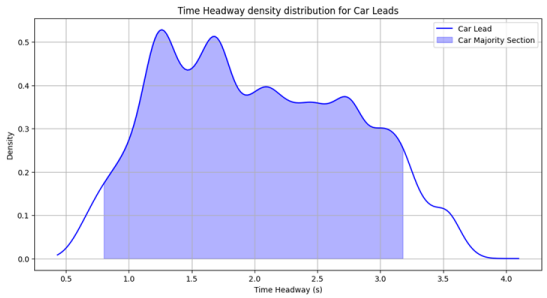


Figure B.3: Time headway density distribution for Car Leads, showing majority section (25-30 m/s, Volvo Cars sample dataset).

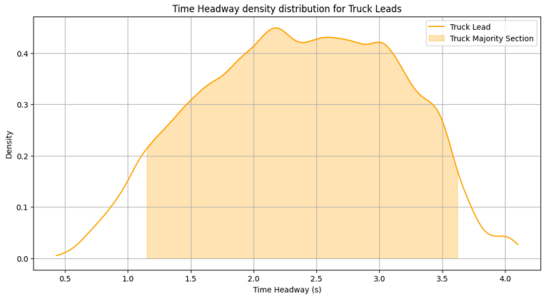


Figure B.4: Time headway density distribution for Truck Leads, showing majority section (25 -30 m/s, Volvo Cars sample dataset).

Furthermore, a comparison with the lower speed group (20-25 m/s) reveals that the majority THW selection in the 25-30 m/s category is generally smaller, and the overall range of the majority section is also narrower. This indicates a more consolidated preference for certain time gaps as speeds increase within this range.

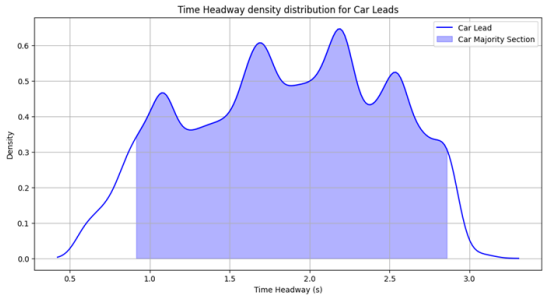


Figure B.5: Time headway density distribution for Car Leads, showing majority section (30-35 m/s, Volvo Cars sample dataset).

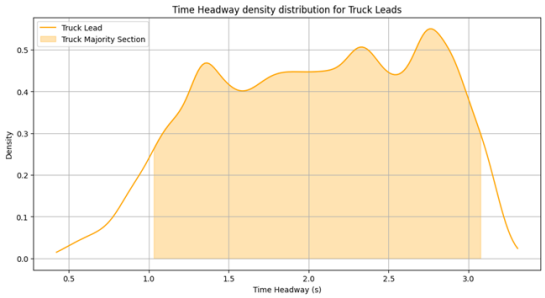


Figure B.6: Time headway density distribution for Truck Leads, showing majority section (30 -35 m/s, Volvo Cars sample dataset).

Consistent with the trend observed in increasing speed categories, the overall data range and the mean THW for both car and truck lead groups continue to be smaller compared to the lower speed groups, indicating a further convergence towards a narrower set of preferred time gaps at higher speeds.

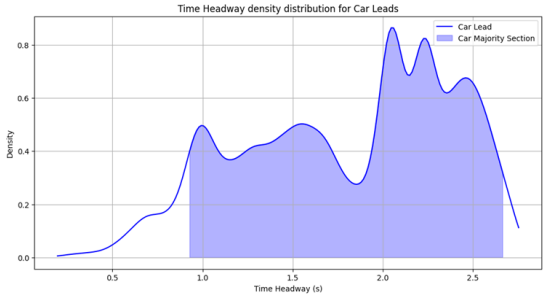


Figure B.7: Time headway density distribution for Car Leads, showing majority section (35-40 m/s, Volvo Cars sample dataset).

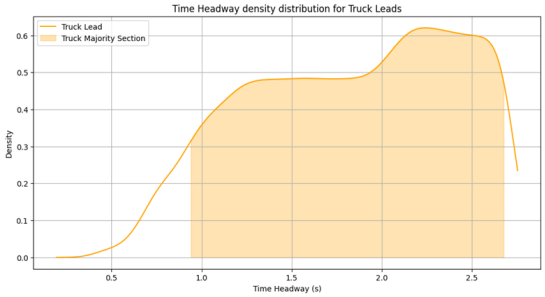


Figure B.8: Time headway density distribution for Truck Leads, showing majority section (35-40 m/s, Volvo Cars sample dataset).

Furthermore, the majority data range for THW in this speed group is slightly smaller compared to the 30-35 m/s group, indicating a continued trend towards a more refined and consistent time gap selection among human drivers under these high-speed, stable conditions.

B.2 Number of Lanes and Lane selection

Figures B.9 and B.10 further detail the majority THW ranges for the outer and inner lanes, respectively.

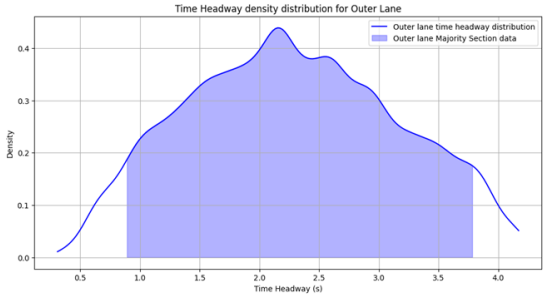


Figure B.9: Time headway density distribution for Outer Lane, showing majority section (20-25 m/s, 2-lane motorway, Volvo Cars sample dataset).

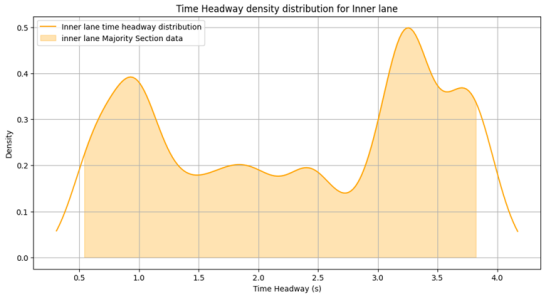


Figure B.10: Time headway density distribution for Inner Lane, showing majority section (20-25 m/s, 2-lane motorway, Volvo Cars sample dataset).

The majority data range is notably wide and diverse for both groups, reflecting a broad spectrum of human driver preferences even under these specific conditions. It is worth noting that the data volume for the inner lane in this speed category is comparatively lower. Additionally, the THW distribution for the inner lane presents a somewhat unique, possibly bimodal, pattern with two discernible peaks at both the lower and higher ends of the distribution. This might suggest a subset of inner-lane drivers either following very closely for quick maneuvers or maintaining larger gaps to avoid frequent lane changes, or it could be an artifact of the smaller sample size influencing the KDE.

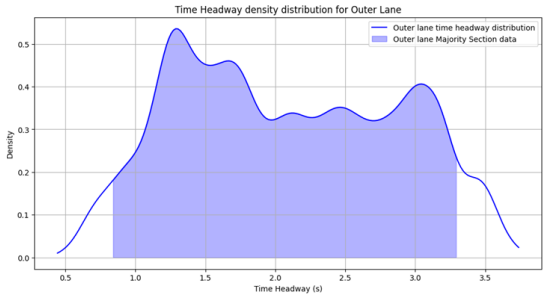


Figure B.11: Time headway density distribution for Outer Lane, showing majority section (25-30 m/s, 2-lane motorway, Volvo Cars sample dataset).

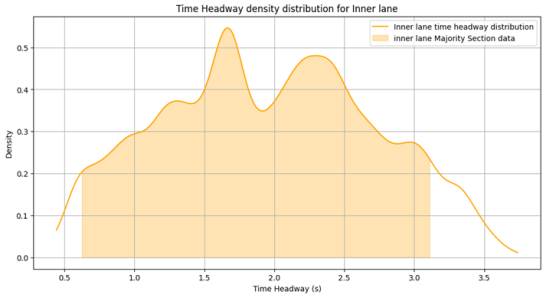


Figure B.12: Time headway density distribution for Inner Lane, showing majority section (25-30 m/s, 2-lane motorway, Volvo Cars sample dataset).

Additionally, consistent with trends observed at increasing speeds, the majority data range for THW in this speed group is also notably smaller and narrower, indicating a more converged and potentially optimized time gap preference as traffic conditions improve and speeds increase.

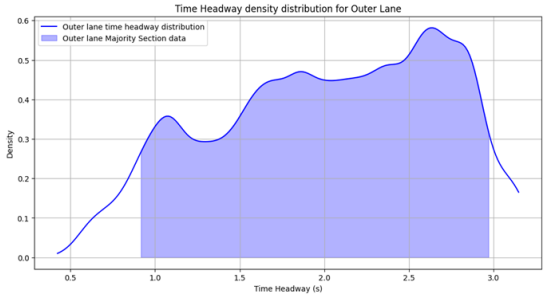


Figure B.13: Time headway density distribution for Outer Lane, showing majority section (30-35 m/s, 2-lane motorway, Volvo Cars sample dataset).

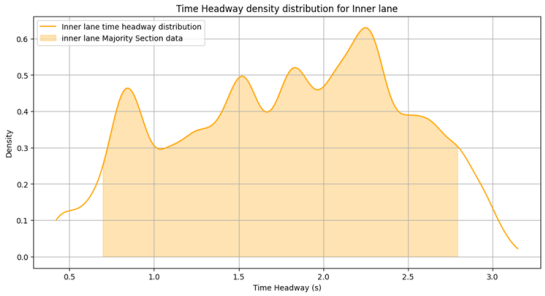


Figure B.14: Time headway density distribution for Inner Lane, showing majority section (30-35 m/s, 2-lane motorway, Volvo Cars sample dataset).

Figures B.13 and B.14 display the majority (90%) THW sections for the outer and inner lanes, respectively, in this speed category. Regarding the majority selection, a similar trend to higher speed ranges is observed: the range of preferred THWs is narrower. Specifically, the maximum THW in the majority section is smaller, while the minimum THW remains relatively consistent, often around 0.8 seconds. This lower bound is critical for safety considerations, as a time headway below this value is generally considered a very close gap that significantly increases collision risk. The reduction in the upper range of preferred time gaps at higher speeds indicates that drivers tend to maintain more compact and consistent following distances in faster, often less congested, traffic flows.
Figures B.15 and B.16 illustrate the majority (90%) THW sections for the outer and inner lanes, respectively, in this highest speed category. For the majority selection, we observe that the range is even narrower compared to the 30-35 m/s group, signifying a highly refined and consistent set of preferred time gaps by human drivers

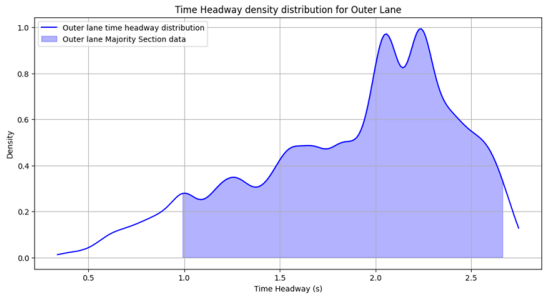


Figure B.15: Time headway density distribution for Outer Lane, showing majority section (35-40 m/s, 2-lane motorway, Volvo Cars sample dataset).

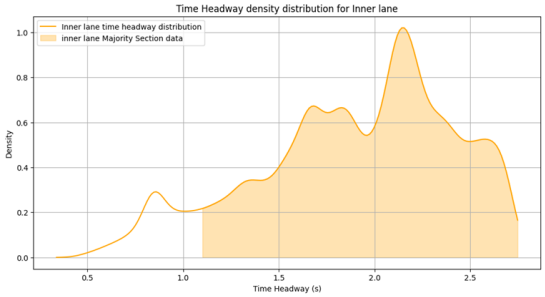


Figure B.16: Time headway density distribution for Inner Lane, showing majority section (35-40 m/s, 2-lane motorway, Volvo Cars sample dataset).

under these ideal conditions. The majority of THWs are concentrated within a very compact interval, generally spanning from approximately 1.0 s to 2.7 s.
Similarly for 3 lanes analysis:
Figures B.17, B.18, and B.19 display the majority (90%) THW sections for the outer, middle, and inner lanes, respectively.

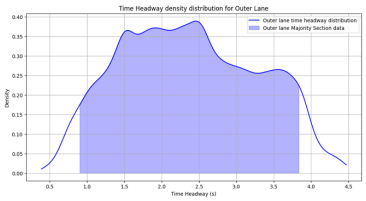


Figure B.17: THW density for Outer Lane, majority section (20-25 m/s, 3-lane).

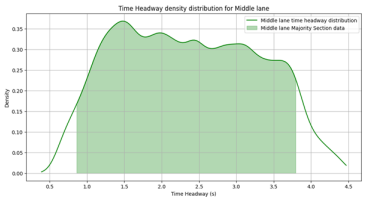


Figure B.18: THW density for Middle Lane, majority section (20-25 m/s, 3-lane).

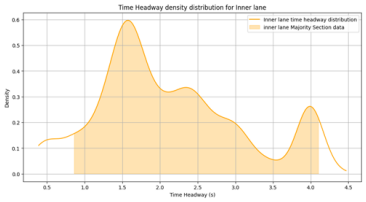


Figure B.19: THW density for Inner Lane, majority section (20-25 m/s, 3-lane).

The majority THW distribution is broadly similar to what was observed in the 2-lane group for this speed category, indicating that the number of lanes does not significantly alter the range of preferred time gaps under these particular low-speed conditions.

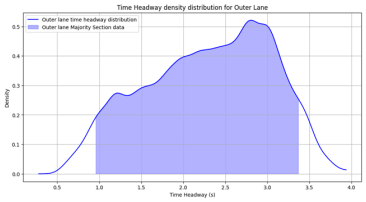


Figure B.20: THW density for Outer Lane, majority section (25-30 m/s, 3-lane).

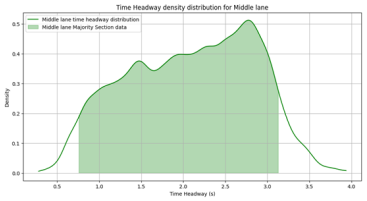


Figure B.21: THW density for Middle Lane, majority section (25-30 m/s, 3-lane).

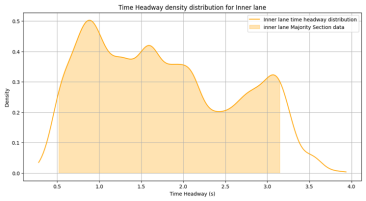


Figure B.22: THW density for Inner Lane, majority section (25-30 m/s, 3-lane).

While the overall distributions show clear distinctions in means, the majority THW

ranges for each lane exhibit trends consistent with those observed in previous analyses, generally showing a narrower spread for higher speeds, reflecting a more refined preference by human drivers in less constrained environments. For instance, the inner lane's majority selection shows a concentration of data points, albeit with some spread, indicating a set of preferred time gaps optimized for faster, less interrupted driving.

B.3 Lighting Condition

Here's majority selection analysis for both the dark and bright lighting condition groups. This analysis identifies the range of THW values most frequently adopted by human drivers within each lighting scenario. Figure B.23 illustrates the majority section for the dark condition group.

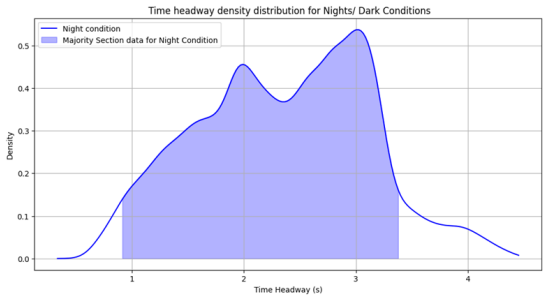


Figure B.23: Time Headway density distribution for Nights/Dark Conditions, showing majority section.

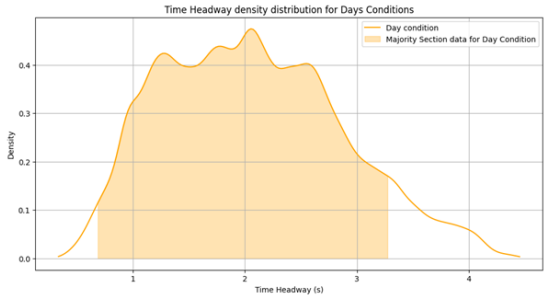


Figure B.24: Time Headway density distribution for Days/Bright Conditions, showing majority section.

From Figure B.24 and Figure B.23, we can clearly see a distinction in the preferred time headway ranges. For the dark lighting condition group, the majority of time headway are concentrated at larger time headway. for the bright lighting condition group, the The peak of the distribution is also shifted towards smaller time headway values compared to the dark condition. This tighter distribution reflects drivers' increased confidence and reduced need for extended reaction times in clear, well-lit conditions. The more compact range signifies a higher degree of consensus among drivers regarding appropriate following distances when visibility is optimal.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY