How Much Data Are Enough? A Statistical Approach With Case Study on Longitudinal Driving Behavior

Article · June 2017

DOI: 10.1109/TIV.2017.2720459

3 authors:


Wenshuo Wang
University of California, Berkeley


Chang Liu
Cornell University


Ding Zhao
University of Michigan

How Much Data is Enough? A Statistical Approach with Case Study on Longitudinal Driving Behavior

Wenshuo Wang, *Studen Member, IEEE,* Chang Liu, *Student Member, IEEE,* Ding Zhao

***Abstract*—Big data has shown its uniquely powerful ability to reveal, model, and understand driver behaviors. The amount of data affects the experiment cost and conclusions in the analysis. Insufficient data may lead to inaccurate models while excessive data waste resources. For projects that cost millions of dollars, it is critical to determine the right amount of data needed. However, how to decide the appropriate amount has not been fully studied in the realm of driver behaviors. This paper systematically investigates this issue to estimate how much naturalistic driving data (NDD) is needed for understanding driver behaviors from a statistical point of view. A general assessment method is proposed using a Gaussian kernel density estimation to catch the underlying characteristics of driver behaviors. We then apply the Kullback-Liebler divergence method to measure the similarity between density functions with differing amounts of NDD. A max-minimum approach is used to compute the appropriate amount of NDD. To validate our proposed method, we investigated the car-following case using NDD collected from the University of Michigan Safety Pilot Model Deployment (SPMD) program. We demonstrate that from a statistical perspective, the proposed approach can provide an appropriate amount of NDD capable of capturing most features of the normal car-following behavior, which is consistent with the experiment settings in many literatures.**

***Index Terms*—Naturalistic driving data, modeling driver behaviors, kernel density estimation, Kullback-Liebler divergence, car-following behaviors**

I. Introduction

NATURALISTIC driving studies have shown great potential in smart city [1], [2], transportation energy efficiency [3]–[5], and driver behaviors [6]–[8], in which data are collected from a number of equipped vehicles driven under naturalistic conditions over an extended period of time [6]. Research institutes around the world have spent great efforts and recourses collecting naturalistic driving data (NDD). For example, the major projects of naturalistic driving study from countries around the world such as the United States, the European Union, Australia, Japan, and China are listed in Table I. From Table I, these naturalistic driving studies vary greatly in research topics, the number of participant drivers ranging from 11 to over 2,700, and the duration of experiments ranging from 1 to 6 years. What has not been fully studied, however, is how much driving data is sufficient to address problems such as the cause of accidents, distraction and inattention, eco-driving styles, modeling driver behavior, and the effects of driver assistance systems on driver behavior. Similar problem concerning "*How much data is enough?*" have been asked in other fields [9]–[12] such as sociology, biology, and oceanography, but not yet in the fields of analyzing/modeling human driving behaviors and traffic safety. Therefore, to avoid the issues of insufficient or excessive data and offer a guideline for primary experiment design, we need to develop an efficient way to estimate the appropriate amount of NDD for a variety of problems.

The required amount of NDD depends on the problem to be solved, the way the problem is formulated, and the dataset to be analyzed (e.g., NDD or driving simulator-based data). For example, a traffic accident analysis usually requires the data with longer driving period than that of modeling driver behaviors, because the reasons for traffic accidents are diverse and reflect a small probability event, compared to common driving behavior. Therefore, to answer this question asked by "*How much naturalistic driving data is enough in understanding driving behaviors?*", we make a further discussion and analysis for different cases and propose a general assessment approach to determine the appropriate amount of NDD from a statistical perspective.

In this paper, our main contributions are: (1) we introduce the problem of the amount of driving data; (2) we propose a general assessment approach to compute an appropriate amount of the required naturalistic driving data; (3) a case of modeling car-following behaviors using naturalistic driving data is conducted to validate our proposed method.

This paper is organized as follows. Section II reviews the related work and analyzes the reasons for diversity in the amount of NDD appearing in the literature. Section III presents a general assessment approach to determine the critical value for the required amount of data. Section IV presents the experiments and the results of a case study for modeling driver behavior. Section V concludes this paper with a discussion, final remarks, and future research directions.

W. Wang is with the Department of Mechanical Engineering, Beijing Institute of Technology, Beijing, China 100081 and also with the Department of Mechanical Engineering, University of California at Berkeley, Berkeley, CA, 94720 USA. e-mail: wwsbit@gmail.com; wwsvdc2015@berkeley.edu.

C. Liu is with the Department of Mechanical Engineering, University of California at Berkeley, Berkeley, CA, 94720 USA. e-mail: changliu@berkeley.edu.

D. Zhao (corresponding author) is with the Department of Mechanical Engineering of the University of Michigan, Ann Arbor, MI 48109 USA. e-mail: zhaoding@umich.edu

II. Analysis of Data Size Used in Existing Studies

As shown in Table I, the number of driver participants and the duration used to collect data vary significantly. The differing data amount appearing in the published papers depends greatly on the financial/equipment capabilities of the experiments, the topics focused on, and the the methods employed.

TABLE I: MAJOR PROJECTS OF NATURALISTIC DRIVING STUDY IN THE WORLD

Project name	Conductor	Period	Mileage [mile]	Vehicle	Sensor	Drivers	Research topic
100 Car Naturalistic Driving Study [6]	Virginia Tech.	2001–2009	2×10^6	100 sedans	camera	109 primary drivers, 132 secondary drivers	Rear end collision
Automotive Collision Avoidance System [13]	University of Michigan	2004-2005	1.37×10^5	11 sedans	camera, radar	96 drivers	Forward collision warning (FCW)
Road Departure Crash Warning [14]	University of Michigan	2005–2006	8.3×10^4	11 sedans	camera, radar	11 drivers	Lane departure warning (LDW)
Sweden-Michigan Naturalistic Field Operational Test (SeMiFOT) [15]	University of Michigan	2008–2009	1.07×10^5	10 sedans, 4 trucks	camera, radar	39 drivers	FCW, LDW, blind spot information system, electronic stability control, and impairment warning
Integrated Vehicle-Based Safety Systems [16]	University of Michigan	2010–2011	sedans: 213&309; trucks: 601&944	16 sedans 10 heavy trucks	camera, radar	108 drivers for sedans; 18 professional truck drivers	Integrated warning
Safety Pilot Model Deployment [17]	University of Michigan	2012–2014	more than 3.4×10^7	2,800 various types of vehicles	camera, radar	2,700 volunteer drivers and several professional bus and truck drivers	Connected vehicle
Google driverless car [18]	Google	2012–present	more than 1.3×10^6	At least 50 sedans and SUVs	lidar, camera, radar	Google technicians and volunteers	Fully self-driven vehicle
Australian Naturalistic Driving Study or Australian 400-car Naturalistic Driving Study [19], [20]	Led by University of New South Wales	2015–present	4 months	400 vehicles	camera, CAN data, GPS	360 participants (180 in New South Wales and 180 in Victoria)	Safety at intersections; Speed choice; Interactions with vulnerable road users; Fatigue; Distraction and inattention; Crashes and near-crashes; Interactions with ITS
European naturalistic Driving and Riding for Infrastructure & Vehicle safety and Environ-ment(UDRIVE) [21]	the 7th EU Framework Programme and 20 partners	2012–2017	On going	200 vehicles (cars, trucks, and scooters)	cameras, IMU sensors, GPS, Mobil Eye smart camera, CAN data, and Sound level	On going	Crash causation and risk; Everyday driving; Distraction and inattention; Vulnerable road users; Eco-driving
China Naturalistic Driving Study	Tongji University; VTTI; General Motors	2012–2015	more than 1.0×10^5	5 vehicles	–	90 drivers; each drove vehicle for 2 months	Exploring Chinese moped-vehicle conflict configurations; Examining car driver responses to moped-vehicle conflicts
Japan Naturalistic Driving Study [22]	Ministry of Land, Infras-tructure, Transport and Tourism	2006–2008	–	60 vehicles (35 wagons & 25 sedans)	GPS, CAN data, acceleration sensor, camera	60 drivers (58 males & 2 females)	Accident causation research

Fig. 1 and Table II show the differences in experimental time[1] of data collection for research on between traffic accident analysis and modeling driver behaviors. The "Total time" includes the time of collecting the raw data or purified data. We do not separate them out, as some references did not clearly distinguish them. The data in Fig. 1 is collected from 26 published papers. We note that research related to traffic accident analysis generally requires a longer period of time for data collection (about 3 years on average) than research on modeling driver behaviors (about 288 minutes on average). The factors that influence the required amount of NDD for traffic accident analysis are analyzed and discussed. We mainly focus on the required amount of NDD for modeling common driver behavior.

[1]Experiment time is the duration for conducting an experiment, which differs from the lasting time of driving events. Data collected from the entire period of experiments is called raw data; the data extracted from the raw data is called purified data. The purified data is usually used to model or analyze driver behaviors.

A. Traffic accident analysis

Traffic accident analysis covers a wide range of topics such as analysis of traffic accident injury severity [37]–[39], relationship analysis between personality and traffic accident [40]–[42], accident hotpots detection or prediction [43], [44],

TABLE II: The Amount of Naturalistic Driving Data in Different Studies on Modeling Driver Behaviors†

References	drivers	vehicles	events	Total time t	Driving tasks	Data type
[23]	5	1	*	300 [min]	Car following	In-vehicle sensors
[24]	*	3	229	$t \approx$190.8 min	Car following	Camera/video data
[25]	20	*	392	$t >$ 196 min	Car following	In-vehicle sensors
[26]	13	*	*	$t >$ 1,200 min	Car following	In-vehicle sensor
[27]	*	*	54	$t \approx$ 1172.8 min	Car following	In-vehicle sensors
[28]	3	*	*	*	Signalized Intersections	Camera/video data
[29]	41	*	*	$49 \lesssim t \lesssim 184$ min	Driver distraction	In-vehicle sensors
[30]	*	*	*	$t \approx$ 720 min	Mirror-checking actions	In-vehicle sensors
[31]	18	26	*	*	Lane change	In-vehicle sensors
[32]	3	1	*	4,947 min	Lane change	In-vehicle sensors
[33]	*	*	$>$ 5,700	$>$ 1,140 min	Lane change	Multisensor data
[34]	*	698 (179 trucks, 519 cars)	*	Extract from 4-month data	Modeling drivers' dynamic decision-making behavior	video-based
[35]	20	1	*	$\approx$ 4,200 min	Lane departure	DS
[36]	24 (20 male, 4 female)	2	*	300 min	Car following and cut-in behavior	Field test

†All the data listed in this table are from the published papers, where $*$ means that we did not find the accurate information in the references. The driving time t is the length of experiment time.

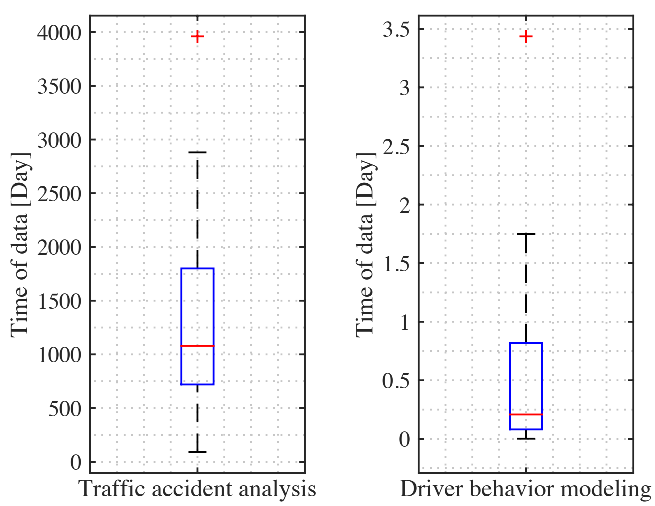


Fig. 1: A comparison between the lasting time of data collection for research topics on traffic accident analysis (left) and modeling driver behaviors (right).

risk factors analysis [45], and traffic accident classification [46]. As shown in reference [47], nearly about thirty approaches were applied to traffic accident analysis. Most data in the traffic accident analysis are collected from the local traffic department, recorded and reported by the traffic police, and/or using questionnaire investigation, which usually does not cost so much compared to the naturalistic driving study. But if conducting research on the relationship between the driving styles and traffic accidents based on the NDD, the data collection will cost a great deal. Three main reasons for the traffic accident analysis requiring long running experiments are:

1) Heterogeneity: The heterogeneity of traffic accidents is reflected in its discretized property in temporal spatial differences. Traffic accident data is generally represented by discrete categories from a variety perspectives. For example, from the viewpoint of injury severity, traffic accident data can be grouped into different levels such as fatal injury or killed, incapacitating injury, non-incapacitating, possible injury, and property damage only [47]. In addition, some heterogeneities of traffic accidents are unobserved, which means that model parameters may vary across observations of traffic accidents. For example, injury severity is likely to exist among the population of crash-involved road users [47] because of differences such as risk-taking behaviors or physiological factors. Therefore, to improve the model accuracy and predict the potential a traffic accident, a huge amount of traffic accident data is normally required.

2) Scarcity: Even though the total number of road traffic crashes is high, the rate of these traffic crashes is low in comparison with the number of miles that people drive. Americans drive nearly 3 trillion miles per year [57], but a failure rate of only 77 per 100 million miles was reported for injuries in 2013. In addition, the diversity in traffic accidents and/or crashes makes a lower rate for a specific kind of traffic accident. For example, the frequency of rear-end crash at the signalized intersection and traffic rush hour will be totally different with the case on the highway. And, different road features and driver's personalities will also cause the diversity in traffic accidents. Therefore, the total number of traffic accidents is high per ten thousands of miles, but for a special or defined case of traffic accident, it is too less to analyze and model this kind of traffic accidents. Thus, to analyze traffic accidents and improve model accuracy, the duration of traffic data should be long enough (usually about 3 years as shown in Fig. 1) and cover more kinds of traffic accident events.

3) Diversity: Traffic accidents can be classified based on criteria such as accident type, age, atmospheric factors, and causes, etc., as shown in Fig. 2, and also depend greatly on these criteria. Thus, a more accurate and comprehensive analysis should be based on a great deal of data that would

TABLE III: AMOUNT OF NATURALISTIC DRIVING DATA FOR RESEARCH ON CAR-FOLLOWING (CF) BEHAVIOR‡

Ref.	Driver	Event	Time	Methods	Topic
[23]	5	(600)	300 min	Gaussian mixture regression & HMM	Modeling CF behaviors
[27]	∗	54	1173 min	Model-based (Steady-State CF Model)	Modeling CF behaviors
[48]	∗	5196	(45 min)	Latent class model structure	Modeling CF behaviors
[24]	∗	229	191 min	Model-based	Interdriver difference
[25]	20	392	196 min	Clustering method	Segment driving patterns
[26]	13	∗	1200 min	Modified latent Dirichlet allocation	Driving style analysis
[49]	∗	6101	(45 min)	Neural networks	Modeling CF behaviors
[50]	∗	(5000)	45 min	Model-based (Newell' CF model)	Capturing traffic oscillations
[51]	276	∗	6 min	GMM and optimal velocity model	Modeling CF behaviors
[52]	∗	∗	6 min	Neural networks	Modeling CF behaviors
[53]	25	35	45	Proposed a new CF model	Explore features of CF and platoon
[54]	1	∗	4.2–5 min	Model-based (Intelligent driver model)	Regime Classification and Calibration
[55]	∗	5687	45 min	Optimization method	Calibrating CF models
[56]	∗	∗	6 min	Model-based (Gazis-Herman-Rothery model)	CF behaviors of individual drivers

‡ All the data is collected from published papers. A value with a bracket indicates that we did not find an accurate value, but we estimated the value using the SPMD datasets. An asterisk ∗ means the reference did not provide any information that can be used to infer the missing value.

Fig. 2: Examples of classifying accident severity based on a variety of criteria.

be able to cover nearly all traffic cases yet be sufficient for accounting for all cases of traffic accidents.

Generally, the heterogeneity, scarcity, and diversity of traffic accidents require that the data collection used for traffic accident analysis should cover a long period of time. The time span for collecting data for traffic accident analysis is much longer than that used for understanding and modeling driver behaviors. On the other hand, the cost of data collection for traffic accident analysis is usually lower than the cost related to understanding and modeling driver behaviors because of the different ways of obtaining data. Therefore, in the following section, we discuss and analyze the causes of diversity in the amount of data for modeling driver behaviors.

B. Modeling Driver Behaviors

Modeling driver behaviors covers a wide range of topics, including, for instance, car following, lane change, left/right turn, U-turn, distraction/inattention, secondary tasks, or brake behaviors. From Fig. 1, we know that data for modeling driver behaviors ranges widely from under 50 minutes (e.g, references [50]–[52]) to more than 5,000 minutes (e.g., reference [32]). We present and analyze the reasons for these big differences in terms of *research topic*, *problem formulation method*, and *data collection methods*. To facilitate the discussion and analysis, we use the car-following behaviors as an example, because car-following behavior is the most common event in driver behaviors.

1) Different Research Topics: Table III shows the wide variation in the amount of NDD across research topics on car-following behaviors. For example, some work focused on the microscopic car-following behavior or traffic flow analysis and collected thousands of car-following events [48], [49], while some others focused on individual car-following behavior and applied hundreds of car-following events to research [23], [25]. Moreover, a special case of car-following behavior, i.e., platoon car-following, required more vehicles in the experiment and a higher dimension of driving data for analysis.

We also found that even for a single kind of research topic, the amount of NDD still varies greatly. For instance, the researchers in [49] and [52] used the same method (i.e., neural networks) to model drivers' car-following behaviors, but varied greatly in the amount of data used.

2) Problem Formulation Methods: The approach to formulating problems can result in diversity in the amount of NDD. Modeling and analyzing drivers' car-following behaviors, generally involves either a *physically-based* or a *learning-based* method.

(a) Physically-based methods: Physically-based method usually describes driver behavior in the form of equations with physical meanings, in which parameters are used to fit the individual driver's characteristics via parameter estimation or calibration methods [54], [55]. For example, the

Gazis-Herman-Rothery (GHR) model describes a driver's car-following behavior by taking current vehicle speed, relative vehicle speed between two adjacent vehicles in the same lane, acceleration, driver reaction time into consideration (see 1).

$$a_n(t) = c \cdot v_n^r(t) \frac{\Delta v(t-T)}{\Delta x^l(t-T)} \tag{1}$$

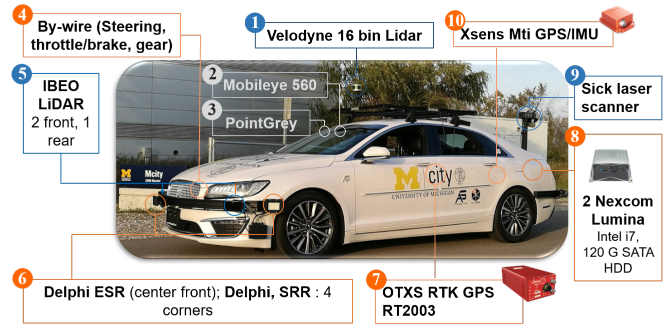
where a_n is the acceleration of vehicle n; v_n^r is the speed of the nth vehicle, Δx and Δv are the relative spacing and speeds, respectively, between the nth and $n-1$ vehicle (the vehicle immediately in front) at an earlier time $t-T$; T is the driver reaction time; r, l and c are the constants to be determined. Most popular car-following models, including the GHR model, intelligent driver model, optimal velocity model, and collision avoidance models, were compared and evaluated in [58], [59]. Thus, the requisite amount of data depends on a number of unknown parameters in physical models. Generally speaking, a physical model with many unknown parameters requires more driving data to fit driver behaviors. In addition, the amount of required data also depends on the method used to calibrate car-following models. For example, a calibration method using statistical techniques usually requires more data than that without considering the statistical features.

(b) Learning-based methods: Learning-based methods utilize machine learning techniques, without considering the physical meaning of the model parameters, to describe more complex and underlying nonlinear relationships between different kinds of surrounding traffic information and driver behaviors. Due to the complexity and diversity of drivers' car-following behaviors, it is generally difficult to capture the stochastic features of drivers using physically-based model. A learning-based method is therefore introduced to solve these kinds of issues. For example, neural networks [49], [52], a Gaussian mixture regression–hidden Markov model [23], [60] and recurrent neural networks [61] have been applied to modeling, analyzing and characterizing driver behaviors. Therefore, different types of problem formulation require different amount of data.

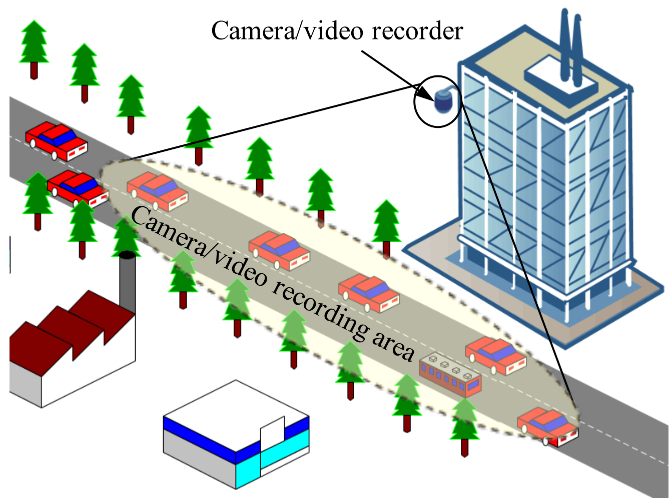
3) Data Collection Approaches: The approach to collecting driving data varies across research topics. Past data collection approaches included: *in-vehicle sensor data* and *video/camera data with a fixed field* (Fig. 3).

(a) In-vehicle sensor data: The NDD collected from in-vehicle sensors, such as cameras and/or radar that can sense information about adjacent vehicles in the same lane and driver's personality, is referred as in-vehicle sensor data. Fig. 3(a) shows an example of an in-vehicle data acquisition system developed by the University of Michigan which consists of an array of sensors such as laser scanners, cameras, and Lidars. For example, Wang [62] *et al.* used cameras to monitor the road, the driver's foot as well as steering hands and analyzed a driver's car-following characteristics. Higgs and Abbas [25] collected the NDD based on in-vehicle cameras, radars, and CAN-Bus signals to analyze a driver's car-following patterns. In addition, the high-precision difference in GPS devices (e.g., Multi-functional Satellite Augmentation System, a product from Japan) can also be directly used to record vehicle speed and position, which can be applied to a pair of cars or car-platoon behaviors [53]. Currently, most data acquisition systems on the market, such as Mobileye used in SPMD program [17] and the data acquisition system in SHRP 2 program developed by VTTI [63] , can be reliably used to collect driving data. This kind of in-vehicle equipment or data acquisition system costs are high, and thus most researchers can not afford a complete set of data acquisition system. Data obtained via the in-vehicle data acquisition system may include data of driver actions/behaviors (e.g., eyes detection, hands detection, and foot action), road features (e.g., road curvature, road/lane width), information of front vehicles (e.g., relative distance, relative speed) and ego vehicle data through CAN-Bus (e.g., acceleration, vehicle speed, throttle opening, steering angle). Thus, for an individual driver, a vehicle with this kind of data acquisition system can be used to built driver behavior models, analyze driver distraction/inattention, ascertain the decision-making process and personal characteristics, and drivers' visual-cognitive, physical and psychomotor capabilities. If many drivers were involved, studies on the difference across individuals could also be conducted, but at a much higher cost.

(b) Video/camera data with a fixed field: A lower cost alternative but efficient way is to install a video recorder at a fixed position, obtaining video-based data (e.g., vehicle



(a) Example of in-vehicle data acquisition systems developed by University of Michigan.



(b) Illustration of data acquisition systems for car-following behaviors using a camera/video recorder with a fixed position.

Fig. 3: Illustrations of two different data collection methods.

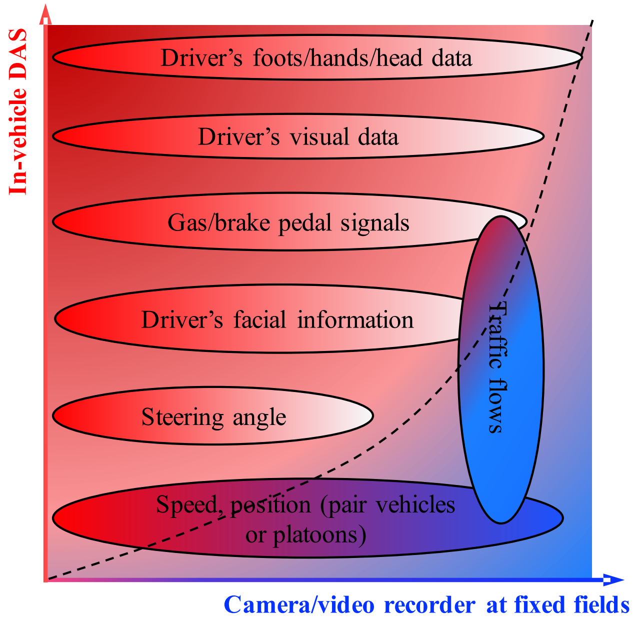


Fig. 4: The illustration of information that could be collected from two different methods.

trajectories and positions) to analyze driver behaviors, as shown in Fig. 3(b). This approach has been widely used to collect vehicle trajectory data and analyze traffic flows or build the car-following model. For example, Yu [64] *et al.* collected the car-following data by installing a video recorder on the windowsill of a tall building adjacent to the intersection, and then utilized these data to analyze the influencing factors of car-following behaviors at urban signalized intersections, determining the structure of an extended car-following model. Some researchers also fixed the camera/video recorder on a helicopter [24], [65], traffic light signal poles and structures to collect driving data. This kind of data collection method allows researchers to obtain a huge amount of driving data for many vehicles at a lower cost and with less time, though tracking a single driver's other behaviors, such as steering angle, head movement, and eye information, is difficult. For instance, more than 6 thousand vehicle trajectories in [55] take the researchers only about 45 minutes to obtain using this method, but included no data on steering angle, head movement. While the method based on an in-vehicle data acquisition system records high-dimension data (Fig. 4), it is very difficult to obtain so many vehicle trajectories of car-following events in a short period of time. As such, this method is usually used for developing a car-following model and analyzing car-following behaviors from a general viewpoint.

Fig. 4 summarizes and presents the comparisons between two approaches of data collection. We note that the collection approach using in-vehicle data acquisition systems, compared to camera/video recorder at a fixed field, can collect a wide range of data from the driver's foot movement to vehicle velocity. The method based on a fixed field camera/video recorder, is best used for collecting a large amount of driving data (i.e., different vehicles) but covering fewer types data.

A video/camera in a fixed field can collect a great amount of driving data at a lower cost, but the diversity of data limits its application in deeply understanding and modeling driver behavior. Thus, most researchers would prefer to utilize multivariate in-vehicle sensors even if it costs more. In the next section, we propose and show a general approach to determine the appropriate amount of NDD for modeling driver behaviors based on an in-vehicle data acquisition system.

III. Proposed Methods

We present an analysis tool to determining how much NDD collected from in-vehicle sensors is sufficient from a statistical point of view. Our proposed methods focus mainly on determining how much NDD is enough to cover the features of driver behaviors rather than assessing which method is better for modeling driver behaviors.

A. Why a Statistical Method?

As discussed in Section II, the amount of NDD varies greatly due to the diversity of research topics, data collection methods, and problem formulation approaches. To develop a flexible approach, we make two assumptions as follows:

- A better driver model or an analysis of driver behavior characteristics should be based on a set of NDD that can cover almost all of the driver's basic characteristics. As such, a driver model built on, or driving characteristics inferred from, an insufficient data set are not suitable for applications.
- Driver behavior is highly affected by uncertainty caused by the surroundings (e.g., other road users) and the driver themselves (e.g., their emotions and mental states), but over the long period of time of driving, the statistical characteristics of driving behavior for an individual driver will be convergent [66], [67]. Namely, a driver will adapt to himself/herself driving styles and then finally shape a stable driving style according to his/her internal model after a long-time period of driving.

In line with the above assumptions, we estimate the appropriate amount of data by finding the convergent point of the density function of collected data from a statistical perspective. The distribution of the NDD sequence $\boldsymbol{x} = \{x_i\}_{i=1}^{n}$ is estimated and denoted as $\hat{F}(x;n)$, and its density is $\hat{f}(x;n) = \frac{d}{dx}\hat{F}(x;n)$ under n observations. For different observation amounts n, the density of observations $\hat{f}(x;n)$ will be different. If an adequate amount of data is provided, the density of observations $\hat{f}(x;n)$ should change slightly with m additional observations, i.e.,

$$\hat{f}(x;n) \sim \hat{f}(x;n+m), \text{ with } n \to \infty, \ m \in \mathbb{R}^{+} \tag{2}$$

If adding more observations does not change the distribution, we consider the additional data is redundant. Thus, we treat the n amount of data as suitable from the statistical perspective, because: (1) the n amount of data can cover almost all of the underlying characteristics of driver behaviors and (2) adding more data can not provide more useful information. The estimated method of density $\hat{f}(x;n)$ is presented formally below.

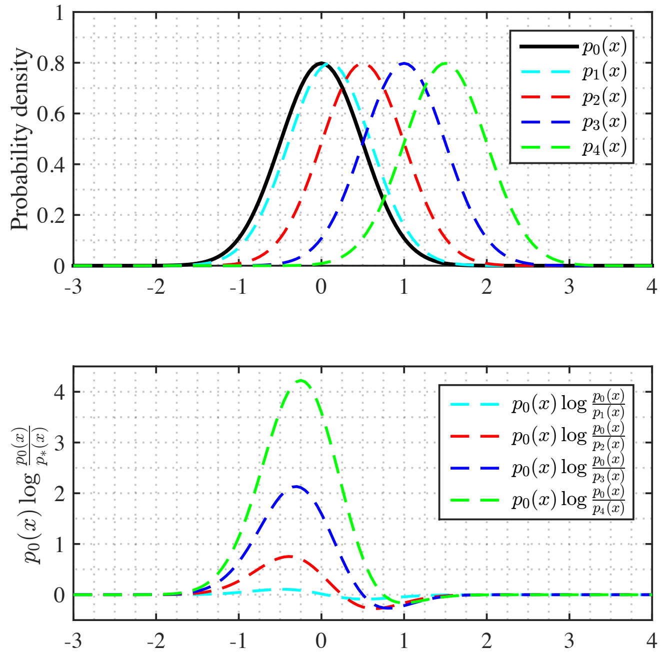


Fig. 5: Illustrations of the integral term ($f\log(\frac{f}{g})$). Top: different density distributions. Bottom: the values of integral terms for different distributions.

B. Univariate Kernel Density Estimation

Driver behavior data can be formulated using a parametric method such as a multivariate Gaussian mixture model (GMM) [23], [32], [51]. It is difficult, however, to directly assess the similarity of two multivariate GMMs, particularly when the number of GMM components is big. In this paper, we utilize a non-parametric method, that is, kernel density estimation (KDE) method, to estimate the density for a given data sequence.

Given a sampling dataset $\{x_i\}_{i=1}^n$ with density function $f(x)$, the estimated density from the data sample x can be formulated by [68]

$$\hat{f}(x;n) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^D}\cdot\kappa\left(\frac{x-x_i}{h}\right) \tag{3}$$

where h is the bandwidth, $\kappa(u)$ is the kernel function and a Gaussian kernel function is selected, i.e., $\kappa(u) = 1/\sqrt{2\pi}\cdot\exp(-u^2/2)$. Thus, we can generate a density function $\hat{f}(x;n)$ on the basis of a given data sample $\boldsymbol{x}$ with n observations. During the kernel density estimation, the kernel bandwidth h has a great influence on the estimated kernel function. A large kernel bandwidth h will result in an over-smooth issue and inversely a small kernel bandwidth h will cause an under-smooth issue. In this paper, we applied a Gaussian kernel function with the bandwidth can be estimated by $h = 1.06\cdot\hat{\sigma}\cdot n^{-1/5}$ [69], where $\hat{\sigma}$ is the standard deviation of the training data $\{x_i\}_{i=1}^n$.

C. Kullback-Liebler Divergence

We will assess the similarity between two adjacent kernel functions estimated from n and $n+m$ data observations. To achieve this, we employ the Kullback-Liebler (KL) divergence index [68] to test the similarity between the distribution of two adjacent data sets, defined by

$$\begin{aligned} KL\left(\hat{f}(x;n+m)||\hat{f}(x;n)\right) = \int\Bigg[\hat{f}(x;n+m) \\ \times\log\frac{\hat{f}(x;n+m)}{\hat{f}(x;n)}\Bigg] \end{aligned} \tag{4}$$

The KL can quantify the level of similarity between two density functions as follows:

1) when $KL(\hat{f}(x;n+m)||\hat{f}(x;n))$ approaches 0, it indicates that $\hat{f}(x;n)$ is extremely close to $\hat{f}(x;n+m)$, meaning that additional data would not provide more useful information to the density function;
2) when $KL(\hat{f}(x;n+m)||\hat{f}(x;n))$ becomes large, it indicates that $\hat{f}(x;n)$ is different from $\hat{f}(x;n+m)$, indicating that more data is needed.

Fig. 5 provides an example to illustrate the KL divergence between different normal density functions. The top picture shows five normal density distributions with different center values, where the black line represents the basic density function. The bottom picture shows the values of the integral term in (4) between the other four density functions and the basic density function. We note that (1) when the probability density $p_0(x)$ is close to $p_1(x)$, the sum value of $p_0(x)\log(\frac{p_0(x)}{p_1})$ approaching to zero and (2) when the probability density $p_0(x)$ is different from $p_4(x)$, the sum value of $p_0(x)\log(\frac{p_0(x)}{p_4})$ becomes larger.

We thus determine the proper amount of driving data so that $KL(\hat{f}(x;n+m)||\hat{f}(x;n))$ change very slightly, even if more data samples were to be added, i.e.,

$$\begin{aligned} &\Big|KL(\hat{f}(x;n+m)||\hat{f}(x;n)) - \\ &KL(\hat{f}(x;n+2m)||\hat{f}(x;n+m))\Big| \leq \epsilon,\ \ \epsilon\in\mathbb{R}^+ \end{aligned} \tag{5}$$

where ϵ is a small positive value. It is obvious that a larger value of ϵ can lead to a small amount of the required NDD. In this paper, to obtain a more conservative result, we set $\epsilon = 10^{-4}$.

IV. Case Study of Modeling Driver Behaviors

The NDD has been widely used to extract, model, and understand driver behaviors or their internal mechanisms, as a new way to design vehicles that transition from automated to manual driving [70], to develop personalized driver assistance systems [28], [32], [60], [71], and to improve fuel efficiency [72] as well as vehicle/road/traffic safety [66]. However, the stochastic features and nonlinearity of driver behaviors make it difficult to directly model and analyze driver behaviors as dynamical systems [32]. A more efficient way is to treat driver behaviors as a stochastic process and fit a model or extract features from a large quantity of data, called the data-driven method. Driving data can be collected using four different testing approaches [73]: (1) driving simulators, (2) quasi-experimental field studies, (3) field operational tests and

Fig. 6: An example of the data collection equipment: (a) Experiment vehicle; (b) Mobileye; (c) Data acquisition system.

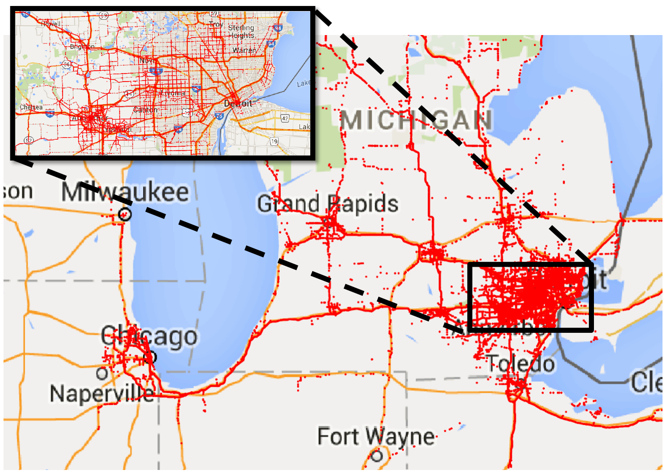


Fig. 7: The trajectories of all car-following data.

(4) naturalistic driving studies. Compared to the first three methods, driving data collected from the fourth method (i.e., NDD) can more accurately reflect a driver's natural traits, but they are very costly and time intensive [73], [74]. An appropriate amount of NDD is required to avoid insufficient or excessive data to save time and money and to improve model accuracy. In this section, we investigate and answer the question "*How much naturalistic driving data is enough to model drivers' behaviors?*" by taking the case of modeling car-following behaviors as an example.

A. Experiments

The NDD used in this research was extracted from the SPMD database. It recorded the naturalistic driving of 2,842 equipped vehicles in Ann Arbor, Michigan, for more than two years. As of April 2016, 34.9 million miles were logged, making the SPMD one of the largest public naturalistic fields of test databases ever. We used 98 sedans to run experiments and collect the real on-road data. The experiment vehicles were equipped with a data acquisition system and MobilEye, as shown in Fig. 6. The in-vehicle data includes vehicle speed, acceleration, and GPS signal from the CAN-bus. The lateral

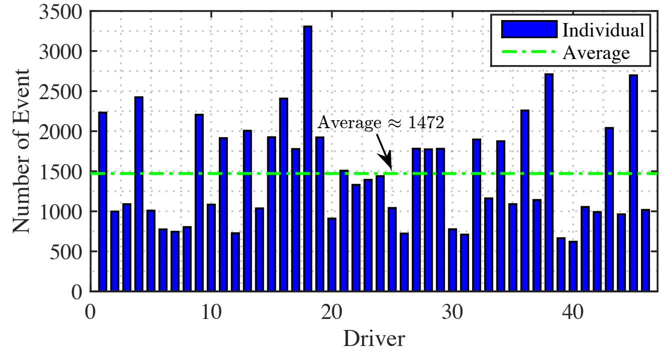


Fig. 8: Statistical information of NDD for 46 drivers.

position with respect to lane or road edges were recorded by MobilEye. All driver participants had an opportunity to drive in rural, urban, and highways situations without any specific restrictions or requirements, as shown in Fig. 7. The NDD were recorded at the rate of 10 Hz or 10 samples per second.

B. Driving Scenarios Definition

We define the following variables to describe drivers' car-following behavior between two adjacent vehicles in the same lane. The ego vehicle is the vehicle we model. The preceding vehicle is the adjacent vehicle located ahead in the same lane as the ego vehicle. To extract the data from the entire database, we define the car-following scenario as follows:

1) Ego vehicle is close to the preceding vehicle in the same lane. The relative distance between the ego vehicle and the preceding vehicle must be longer than 120 m [25]. If the relative distance between the two vehicles is larger than 120 m, this driver behavior was treated as a free-following case.
2) The speed of the ego vehicle is larger than 5 m/s. The limitation is placed on speed to separate the car-following data from the traffic jam data and Stop&Go data.
3) The cut-in behavior of surrounding vehicles or lane change behavior of the ego vehicle is also not involved. When a car cut-in from the neighboring lane to the gap between the current preceding vehicle and the ego vehicle, or the ego vehicle makes a lane change behavior, the car-following event will end.
4) The length of the car-following period must be greater than 30 s [25], and the number of car-following events for each driver should be larger than 300. The two limitations ensure that the NDD is sufficiently large for determining the appropriate amount.

After data extraction, most typical car-following behavior were included such as data related to constant moving speed of the leading car at various speed, data related to constant acceleration, deceleration, oscillation with various amplitude and frequency, etc.

C. Data Processing

Based on the definition and limitations of the car-following behavior, 46 drivers with 67,754 car-following events were

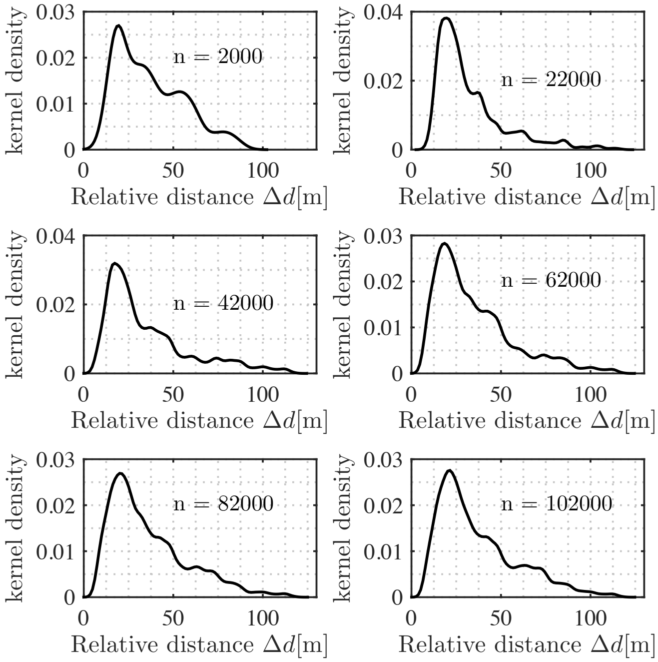


Fig. 9: Example of kernel density of relative distance for driver #12 car-following behavior using different amounts of NDD.

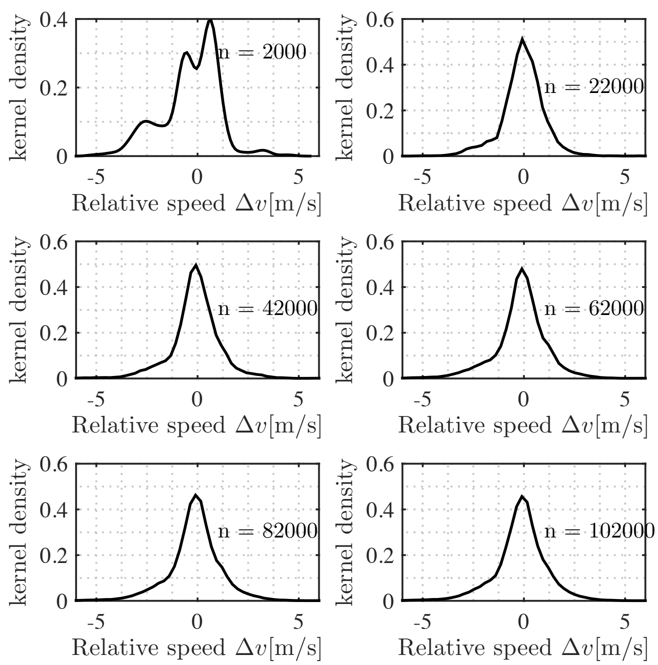


Fig. 10: Example of kernel density of relative speed for driver #12 car-following behavior using different amounts of NDD.

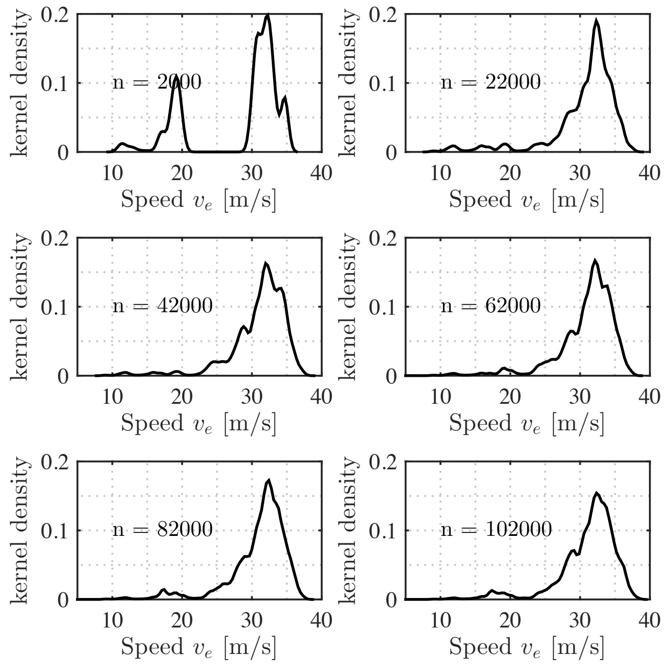


Fig. 11: Example of kernel density of speed for driver #12 car-following behavior using different amounts of NDD.

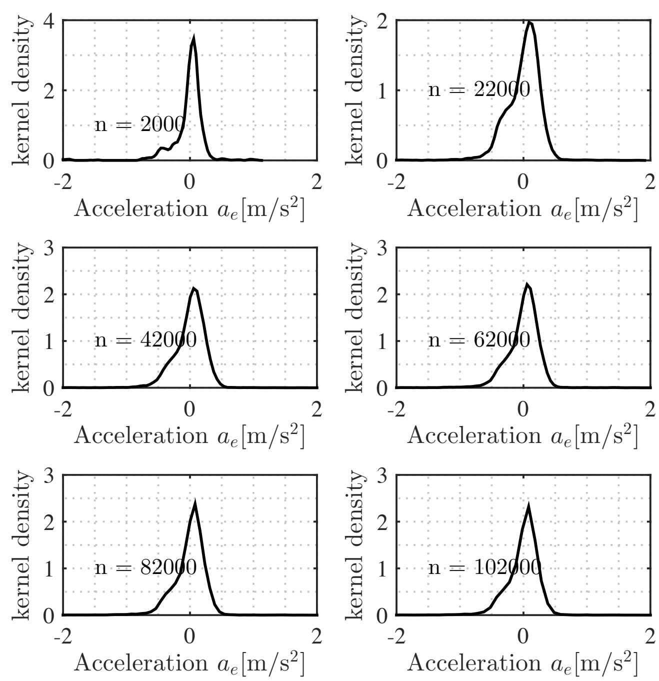
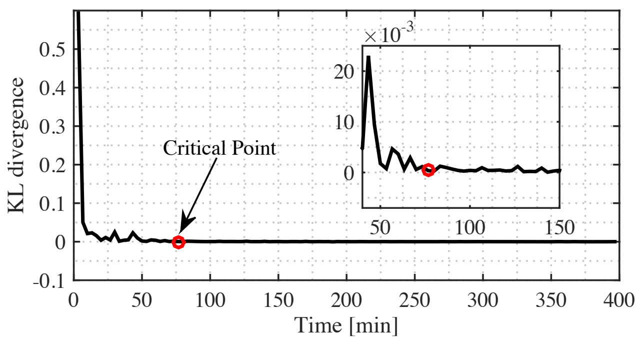


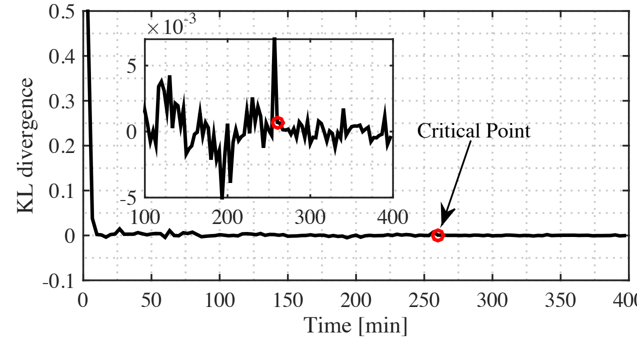
Fig. 12: Example of kernel density of acceleration for driver #12 car-following behavior using different amounts of NDD.

extracted (Fig. 8). For modeling car-following behaviors, the variable selection varies by research topic. Different variable selection requires differing amounts of NDD. In this research, we apply the velocity v_e of the ego vehicle, the acceleration a_e of the ego vehicle, the relative speed Δv, and the relative distance Δd between the ego vehicle and the preceding vehicle to formulate drivers' car-following behaviors, similar to [23]. For each variable, we compute the critical amount of driving data using (5). To make the method more generalizable, we propose a max-minimum method to determine an appropriate amount of NDD. The appropriate amount of driving data that can fully cover driver behavior characteristics for each variable is computed by

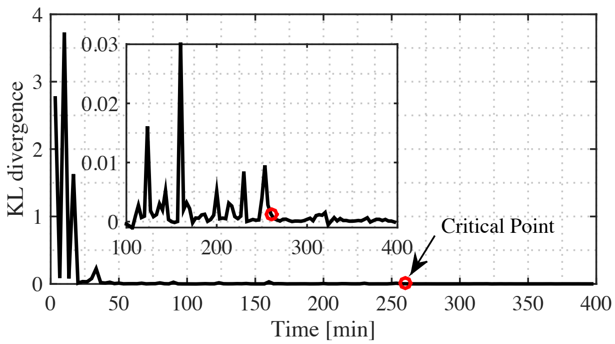
$$n^*_{\{\star\}} = \min\{n|\text{Equation}(5) \text{ is valid}\} \tag{6}$$



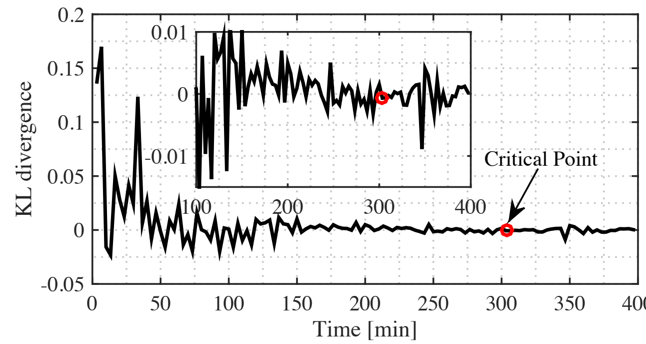
(a) Relative range, $n^*_{\Delta d} \approx 4.6 \times 10^4$



(b) Relative speed, $n^*_{\Delta v} \approx 1.56 \times 10^5$

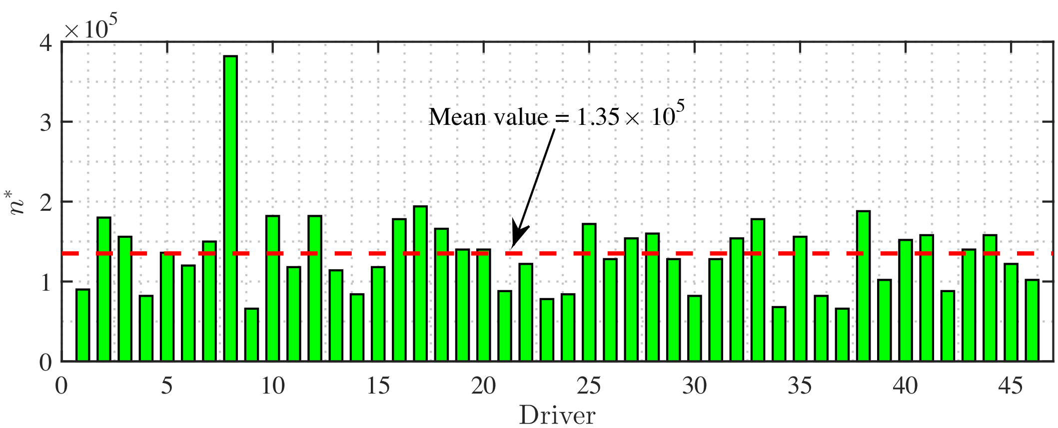


(c) Speed, $n^*_{v_e} \approx 5.8 \times 10^4$

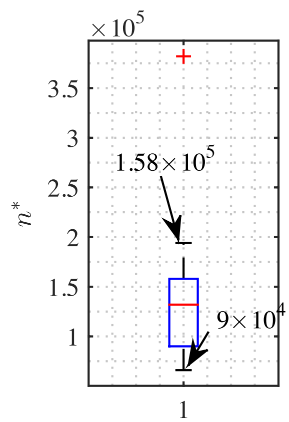


(d) Acceleration, $n^*_{a_e} \approx 1.82 \times 10^5$

Fig. 13: The appropriate data amount of modeling driver's car-following behavior in terms of four variables for driver #12 with $\epsilon = 10^{-4}$.



(a)



(b)

Fig. 14: The appropriate amount of NDD for all the participants in terms of modeling car-following behaviors with $\epsilon = 10^{-4}$.

with $\{\star\} \in \{v_e, a_e, \Delta v, \Delta d\}$ and $m = 2,000$ in (5). According to (5) and (6), for each variable we can find an appropriate amount of NDD to cover the underlying characteristics. If researchers utilize a multivariate model to describe driver behaviors, the minimum amount of required NDD to cover driver behavior characteristics is the maximum value of all appropriate amount of these variables. Taking modeling the car-following behaviors for example, the appropriate amount of NDD using four variables can be computed by

$$n^* = \max\{n^*_{\{\star\}} | \{\star\} \in \{v_e, a_e, \Delta v, \Delta d\}\} \tag{7}$$

Thus, we can obtain the optimal amount of NDD that can most effectively cover all the driving characteristics that we focus on by using the NDD as little data as possible.

D. Results Discussion and Analysis

1) Univariate Kernel Density Estimation: Based on (3), we obtain the kernel density for all variables with different amounts of data, as shown in Fig. 9 – Fig. 12. From the estimated results of kernel density with four variables, we note that when the amount of driving data is limited, the density changes greatly. For example, kernel densities greatly differ for relative distance, relative speed, speed and acceleration of the

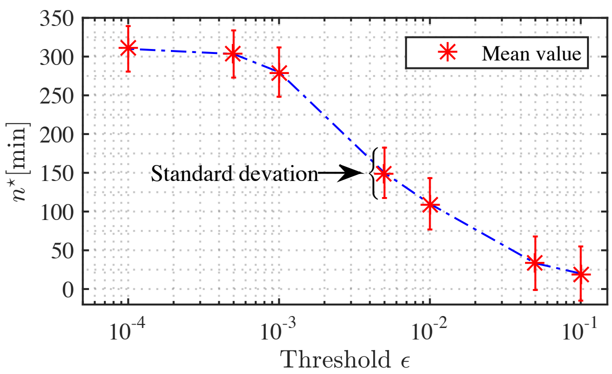


Fig. 15: The statistical results of the influences of threshold ϵ on data size for 46 drivers.

ego vehicle, when comparing $n = 2,000$ and $n = 22,000$, respectively. When the quantity of the data is larger, the divergences between densities with different data amounts are smaller. For example, the kernel densities with $n = 82,000$ and $102,000$ are quite similar for every single variable.

2) Appropriate Amount of NDD: To show the appropriate data amount of data for each variable, examples for driver #12 are given for each single variable. The KL divergences for each variable are computed by (4) and shown in Fig. 13. The red circle represents the critical value for each variable computed via (5). The vertical axis is the KL divergence value and the horizontal axis is the driving time, t, of collecting data, computed by

$$t = \frac{n}{f \cdot 60} \tag{8}$$

where n is the amount of data collected, f is the sample frequency, the unit of t is minute, and $f = 10$ Hz. We can conclude that the appropriate amounts of driving data with respect to Δd, Δv, v_e and a_e are $4.6 \times 10^4 (\approx 76.7 \text{ min})$, $1.56 \times 10^5 (\approx 260 \text{ min})$, $1.56 \times 10^5 (\approx 260 \text{ min})$, and $1.82 \times 10^5 (\approx 303.3 \text{ min})$, respectively. Based on the results in Fig. 13, the appropriate amount of data for modeling the car-following behaviors of driver #12 using four variables can be computed by (7) and obtained as $n^* = 1.82 \times 10^5$.

Fig. 14 shows the statistical results of the appropriate amount of NDD to model drivers' car-following behavior for all driver participants. We note that the appropriate amount of NDD to model the driver's car-following behavior using four variables is about 1.35×10^5 ($\approx$ 225.5 min). The suitable amount of NDD for modeling driver's car-following behavior ranges from $9.0 \times 10^4 (\approx 150 \text{ min})$ to $1.58 \times 10^5 (\approx 263.3 \text{ min})$, as shown in Fig. 14(b).

3) Influence of Threshold ϵ on Data Size: According to (5) we know that the threshold ϵ will affect the estimated data amount for understanding driver behavior. Fig. 15 presents the influences of threshold ϵ on the estimated amount of NDD. We conclude that a larger threshold results in a smaller amount of NDD, and vice versa. When the threshold is less than 5×10^{-4}, the amount of required NDD is convergent to a constant ($\approx$ 300 min) for the car-following behaviors. Therefore, to obtain a conservative result, the threshold was set $\epsilon < 10^{-3}$. When

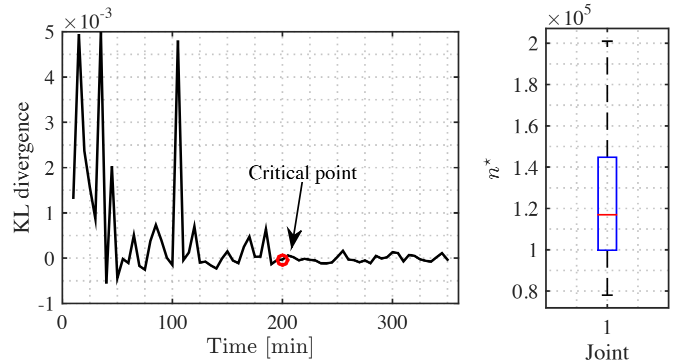


Fig. 16: The KL divergence using the multivariate kernel density estimation method. Left: a case example; right: the statistical results for the critical point with $\epsilon = 10^{-4}$.

TABLE IV: The Optimal Amount of Demanded NDD Using Univariate and Multivariate KDE Methods.

	Median	Maximum	Minimum
Univariate KDE	225.5 min	263.3 min	150.0 min
Multivariate KDE	195.0 min	335.0 min	130.0 min

$\epsilon = 5 \times 10^{-4}$, the results ($n^* \approx 300$ min in Fig. 14 and Fig. 15) from the methodology we propose in this paper are consistent with the results collected from the published papers ($n^* \approx 288$ min in Fig. 1), which also support the claims based on our proposed methods.

E. Multivariate KDE Method

To support the proposed method, we also investigate the joint relationship between different variables using multivariate KDE [2] method [69]. Thus, a multivariate kernel density, $\widehat{\boldsymbol{f}}(\boldsymbol{x}; n)$, with n amount of driving data is estimated, where $\boldsymbol{x} \in \mathbb{R}^{4\times 1}$. To improve computing speed, we select 15 points for each variable as computing points, then obtaining $N = 15^4$ vectors $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^N$ to compare the similarity between two multivariate kernel densities by

$$KL\left(\widehat{\boldsymbol{f}}(\boldsymbol{x}; n+m) || \widehat{\boldsymbol{f}}(\boldsymbol{x}; n)\right) = \sum_{i=1}^{N} \widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}_i; n+m) \log \frac{\widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}_i; n+m)}{\widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}_i; n)} \tag{9}$$

Fig. 16 demonstrates an example of the optimal amount of NDD that is enough to cover driver's car-following characteristics based on multivariate KDE and the statistical results of 21 drivers. We can know that the appropriate amount of driving data to model driver's car-following behavior using four variables is about $n^* = 1.17 \times 10^5$ ($\approx$195 min). The right plot in Fig. 16 demonstrates that the suitable amount of NDD ranges from 7.8×10^4 ($\approx$ 130 min) to 2.01×10^5 ($\approx$ 335 min).

Table IV compares the estimation results of the amount of required driving data for modeling car-following behavior

[2]This can be achieved by using Matlab command `mvksdensity`

using four variables based on univariate KDE and multivariate KDE. We note that the univariate KDE method and the multivariate KDE method obtain the appropriate data amount of 225.5 min and 195.0 min, respectively. The minimum amounts of required NDD using both methods are also similar (150.0 min and 130.0 min), but the univariate KDE method will slightly overestimate the required data amount, compared to the multivariate KDE method.

However, the multivariate KDE method will exponentially increase the computation cost with increasing sampling data points of each variable. In the case with a four-dimension feature $\boldsymbol{x} = [x_1, x_2, x_3, x_4]^T \in \mathbb{R}^{4\times 1}$, M sampling points of each variable are selected, i.e., $\widetilde{x}_i = \{\widetilde{x}_i^1, \cdots, \widetilde{x}_i^M\}$, where $i = 1, 2, 3, 4$, then we will obtain M^4 sampling feature vectors by meshing each variable to compute $KL(\widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}; n+m)||\widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}; n))$ in (9). Compared to the multivariate KDE method, the univariate KDE method only requires $4M$ sampling points in the same condition. For example, when $M = 100$, the univariate KDE method only requires 400 data points, but the multivariate KDE method needs to compute 10^8 feature vectors. Therefore, in our case, the amount of sampling point in each variable is selected as 15 to compute the KL divergence when using the multivariate KDE method. A lower amount of sampling point in multivariate KDE method can shorten computing time but reduce the accuracy of estimating $KL(\widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}; n+m)||\widehat{\boldsymbol{f}}(\widetilde{\boldsymbol{x}}; n))$, which may result in no solutions for convergent condition (5).

V. Further Discussions

In this paper, we point out and discuss the issues concerning the amount of data needed to understand and model driver behaviors, which is, to our best knowledge, the very first time to do so in literature. Question such as "*How much naturalistic driving data is sufficient for understanding and modeling driver behaviors?*" is a basic issue that most researchers face. The methodology included in this paper can be used to assess the amount of data before modeling driver behaviors and designing a data-driven driving simulator. We provide a case study for the longitudinal driving behaviors to demonstrate the advantages of the proposed method. The approach could also be extended to the lateral driving behavior analysis such as lane change behavior. Other attributes are discussed below.

A. Personalized Behavior

In this paper, we focus primarily on modeling driver behaviors using the NDD collected from each single driver. We utilize the individual's driving data to model and understand individual driver behaviors that is also called personalized behaviors. The analysis and investigation based on all drivers' driving data for general driver behaviors were not involved in this paper. The methodology developed in this paper can also be directly applied to determining the requisite amount of data for establishing a general driver model, thus reducing the cost of experiments and resources. We will collect a broader range of driving data covering different ages, driving experience, and genders to investigate the difference in the amount of required data for modeling between individual and general driver behavior.

B. Small Probability Events

The proposed assessment method for determining how much NDD is sufficient is feasible for modeling and understanding common driver behaviors such as car following, lane change, distractions/inattentions, or decision-making behaviors. But we have not investigated its application in research focusing on events at low probability, such as traffic accidents, because the small probability events has their own analysis approach [57] differing from the proposed method in this paper.

C. Feature Variable Selection

As discussed in Section II, different formulation methods, including feature variable selection, lead to variety in the required amount of data. From (7), we know that the proposed method depends greatly on feature variable selection, which renders the proposed method more flexible. Let us take the car-following modeling of driver #12 for example. When four variables are selected as shown in our case study, the appropriate amount of NDD is about 300 min; but when only three variables, e.g., relative distance, relative speed, and vehicle speed, are selected, the appropriate amount of NDD will be about 260 min (Fig. 13).

In this case study, we applied our approach to a limited number of scenarios. For example, stop-and-go scenarios were not included. However, we expect that the proposed methodology for determining how much data is enough to cover the features of driver behavior is relevant for a variety of scenarios, including stop-and-go.

VI. Conclusion

In this paper, we focus on issues concerning the amount of data needed in naturalistic driving studies. To understand the diversity in the amount of data required for modeling driver behavior, we discuss and analyze the factors across different kinds of research. We propose a general method to determine the appropriate amount of driving data used for modeling driver behaviors from a statistical perspective. The Gaussian kernel density estimation approach is utilized and the Kullback-Liebler divergence method is employed to evaluate the similarity between two density functions with differing amounts of data. And then, a max-minimum method is applied to determine the appropriate amount of driving data. Last, a case study for modeling driver car-following behavior using the naturalistic driving data is conducted to demonstrate our proposed method. The proposed method in this paper and the conclusions from our experiment can provide researchers and engineers guidelines to design or conduct a naturalistic driving study.

However, thus far the proposed method does not suffice to reveal the correlated traffic dynamics over space and time. The method allows to determine the appropriate amount of driving data covering most of driving behaviors without considering correlated traffic dynamics and dynamic process in primitive behaviors. The development of a general method based on driving patterns and traffic dynamics to determine the amount of required driving data is our future work.

REFERENCES

[1] I. Vilajosana, J. Llosa, B. Martinez, M. Domingo-Prieto, A. Angles, and X. Vilajosana, "Bootstrapping smart cities through a self-sustainable model based on big data flows," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 128–134, 2013.

[2] A. M. Townsend, *Smart cities: Big data, civic hackers, and the quest for a new utopia*. WW Norton & Company, 2013.

[3] M. B. Arias and S. Bae, "Electric vehicle charging demand forecasting model based on big data technologies," *Applied Energy*, vol. 183, pp. 327–339, 2016.

[4] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 215–225, 2016.

[5] H. Cai, X. Jia, A. S. Chiu, X. Hu, and M. Xu, "Siting public electric vehicle charging stations in beijing using big-data informed travel patterns of the taxi fleet," *Transportation Research Part D: Transport and Environment*, vol. 33, pp. 39–46, 2014.

[6] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, D. J. Ramsey *et al.*, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," 2006.

[7] P. Green, "Integrated vehicle-based safety systems (ivbss): Human factors and driver-vehicle interface (dvi) summary report," 2008.

[8] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE transactions on intelligent transportation systems*, 2016.

[9] R. E. Heyman, B. R. Chaudhry, D. Treboux, J. Crowell, C. Lord, D. Vivian, and E. B. Waters, "How much observational data is enough? an empirical test using marital interaction coding," *Behavior Therapy*, vol. 32, no. 1, pp. 107–122, 2002.

[10] A. H. Wortley, P. J. Rudall, D. J. Harris, and R. W. Scotland, "How much data are needed to resolve a difficult phylogeny? case study in lamiales," *Systematic Biology*, vol. 54, no. 5, pp. 697–709, 2005.

[11] W. Saris, S. Blair, M. Van Baak, S. Eaton, P. Davies, L. Di Pietro, M. Fogelholm, A. Rissanen, D. Schoeller, B. Swinburn *et al.*, "How much physical activity is enough to prevent unhealthy weight gain? outcome of the iaso 1st stock conference and consensus statement," *Obesity reviews*, vol. 4, no. 2, pp. 101–114, 2003.

[12] K. D. Splinter, I. L. Turner, and M. A. Davidson, "How much data is enough? the importance of morphological sampling interval and duration for calibration of empirical shoreline models," *Coastal Engineering*, vol. 77, pp. 14–27, 2013.

[13] R. Ervin, J. Sayer, D. LeBlanc, S. Bogard, M. Mefford, M. Hagan, Z. Bareket, and C. Winkler, "Automotive collision avoidance system field operational test report: Methodology and results," 2005.

[14] D. LeBlanc, J. Sayer, C. Winkler, R. Ervin, S. Bogard, M. Devonshire, J. Mefford, M. Hagan, Z. Bareket, R. Goodsell, and T. Gordon, "Road departure crash warning system field operational test : Methodology and results," 2006.

[15] T. Victor, J. Bärgman, M. Hjälmdahl, and K. Kircher, "Sweden-michigan naturalistic field operational test (SeMiFOT) phase 1: Final report," Tech. Rep., Feb. 2010.

[16] J. Sayer, D. LeBlanc, S. Bogard, D. Funkhouser, S. Bao, M. Buonarosa, and A. Blankespoor, "Integrated vehicle-based safety systems field operational test final program report," Tech. Rep., 2011.

[17] D. Bezzina and J. Sayer, "Safety pilot model deployment: Test conductor team report," *Report No. DOT HS*, vol. 812, p. 171, 2014.

[18] "Google self-driving car project." [Online]. Available: https://www.google.com/selfdrivingcar/

[19] [Online]. Available: http://www.ands.unsw.edu.au/about-study

[20] M. Regan, A. Williamson, R. Grzebieta, J. Charlton, M. Lenneb, B. Watson, N. Haworth, A. Rakotonirainy, J. Woolley, R. Anderson *et al.*, "The australian 400-car naturalistic driving study: Innovation in road safety research and policy," in *Proceedings of the 2013 Australasian road safety research, policing & education conference, Brisbane, Queensland*, 2013.

[21] Y. Barnard, F. Utesch, N. Nes, R. Eenink, and M. Baumann, "The study design of udrive: the naturalistic driving study across europe for cars, trucks and scooters," *European Transport Research Review*, vol. 8, no. 2, pp. 1–10, 2016.

[22] N. Uchida, M. Kawakoshi, T. Tagawa, and T. Mochida, "An investigation of factors contributing to major crash types in japan based on naturalistic driving data," *IATSS research*, vol. 34, no. 1, pp. 22–30, 2010.

[23] S. Lefèvre, A. Carvalho, and F. Borrelli, "A learning-based framework for velocity control in autonomous driving," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 32 – 42, Jan. 2016.

[24] S. Ossen, S. Hoogendoorn, and B. Gorte, "Interdriver differences in car-following: a vehicle trajectory-based study," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1965, pp. 121–129, 2006.

[25] 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.

[26] G. Qi, Y. Du, J. Wu, N. Hounsell, and Y. Jia, "What is the appropriate temporal distance range for driving style analysis?" *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1393–1403, 2016.

[27] L. Pariota, G. N. Bifulco, and M. Brackstone, "A linear dynamic model for driving behavior in car following," *Transportation Science*, vol. 50, no. 3, pp. 1032 – 1042, Aug 2015.

[28] V. A. Butakov and P. Ioannou, "Personalized driver assistance for signalized intersections using V2I communication," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 1910 –1919, Jul. 2016.

[29] T. Liu, Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1108–1120, 2016.

[30] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, 2016.

[31] K. Nobukawa, S. Bao, D. J. LeBlanc, D. Zhao, H. Peng, and C. S. Pan, "Gap acceptance during lane changes by large-truck drivers?an image-based analysis," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 3, pp. 772–781, 2016.

[32] V. A. Butakov and P. Ioannou, "Personalized driver/vehicle lane change models for ADAS," *IEEE Transaction on Intelligent Transportation Systems*, vol. 64, no. 10, pp. 4422 – 4431, Oct. 2015.

[33] H. Zhao, C. Wang, Y. Lin, F. Guillemard, S. Geronimi, and F. Aioun, "On-road vehicle trajectory collection and scene-based lane change analysis: Part I," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 192–205, 2017.

[34] K. Tang, S. Zhu, Y. Xu, and F. Wang, "Modeling drivers' dynamic decision-making behavior during the phase transition period: An analytical approach based on hidden markov model theory," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 206–214, Jan. 2016.

[35] Y. Saito, M. Itoh, and T. Inagaki, "Driver assistance system with a dual control scheme: Effectiveness of identifying driver drowsiness and preventing lane departure accidents," *IEEE Transactions on Human-Machine Systems*, DOI: 10.1109/THMS.2016.2549032.

[36] J. Wang, J. Wu, X. Zheng, D. Ni, and K. Li, "Driving safety field theory modeling and its application in pre-collision warning system," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 306–324, 2016.

[37] J. de Oña, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on spanish rural highways using bayesian networks," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 402–411, 2011.

[38] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 1019–1027, 2006.

[39] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident Analysis & Prevention*, vol. 38, no. 3, pp. 434–444, 2006.

[40] P. Ulleberg, "Personality subtypes of young drivers. relationship to risk-taking preferences, accident involvement, and response to a traffic safety campaign," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 4, no. 4, pp. 279–297, 2001.

[41] N. Sümer, "Personality and behavioral predictors of traffic accidents: testing a contextual mediated model," *Accident Analysis & Prevention*, vol. 35, no. 6, pp. 949–964, 2003.

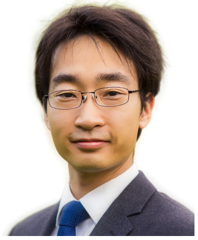
[42] H. Iversen and T. Rundmo, "Personality, risky driving and accident involvement among norwegian drivers," *Personality and individual Differences*, vol. 33, no. 8, pp. 1251–1263, 2002.

[43] T. K. Anderson, "Kernel density estimation and k-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[44] Z. Xie and J. Yan, "Kernel density estimation of traffic accidents in a network space," *Computers, Environment and Urban Systems*, vol. 32, no. 5, pp. 396–406, 2008.
[45] G. Zhang, K. K. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Analysis & Prevention*, vol. 87, pp. 34–42, 2016.
[46] J. de Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using latent class clustering and bayesian networks," *Accident Analysis & Prevention*, vol. 51, pp. 1–10, 2013.
[47] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1666–1676, 2011.
[48] H. N. Koutsopoulos and H. Farah, "Latent class model for car following behavior," *Transportation research part B: methodological*, vol. 46, no. 5, pp. 563–578, 2012.
[49] 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, Cybernetics: Systems and Humans*, vol. 42, no. 6, pp. 1440 –1449, Nov. 2012.
[50] D. Chen, J. Laval, Z. Zheng, and S. Ahn, "A behavioral car-following model that captures traffic oscillations," *Transportation research part B: methodological*, vol. 46, no. 6, pp. 744–761, 2012.
[51] C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, and F. Itakura, "Driver modeling based on driving behavior and its evaluation in driver identification," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 427–437, 2007.
[52] S. Panwai and H. Dia, "Neural agent car-following models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 60–70, 2007.
[53] R. Jiang, M.-B. Hu, H. Zhang, Z.-Y. Gao, B. Jia, and Q.-S. Wu, "On some experimental features of car-following behavior and to model them," *Transportation Research Part B: Methodological*, vol. 80, pp. 338–354, 2015.
[54] A. B. Zaky, W. Gomaa, and M. A. Khamis, "Car following markov regime classification and calibration," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 1013–1018.
[55] P. J. Jin, D. Yang, and B. Ran, "Reducing the error accumulation in car-following models calibrated with vehicle trajectory data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 148–157, 2014.
[56] S. Ossen and S. Hoogendoorn, "Car-following behavior analysis from microscopic trajectory data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1934, pp. 13–21, 2005.
[57] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.
[58] M. Brackstone and M. McDonald, "Car-following: a historical review," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 2, no. 4, pp. 181–196, 1999.
[59] S. Panwai and H. Dia, "Comparative evaluation of microscopic car-following behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 3, pp. 314–325, 2005.
[60] W. Wang, D. Zhao, J. Xi, and W. Han, "A learning-based approach for lane departure warning systems with a personalized driver model," *arXiv preprint arXiv:1702.01228*, 2017.
[61] J. Morton, T. A. Wheeler, and M. J. Kochenderfer, "Analysis of recurrent neural networks for probabilistic modeling of driver behavior," *IEEE Transactions on Intelligent Transportation Systems*, DOI: 10.1109/TITS.2016.2603007.
[62] J. Wang, L. Zhang, D. Zhang, and K. Li, "An adaptive longitudinal driving assistance system based on driver characteristics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 1–12, 2013.
[63] K. L. Campbell, "The shrp 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety," *TR News*, no. 282, 2012.
[64] S. Yu and Z. Shi, "An extended car-following model at signalized intersections," *Physica A: Statistical Mechanics And Its Applications*, vol. 407, pp. 152–159, 2014.
[65] H. Ozaki, "Reaction and anticipation in the car-following behavior." *Transportation and traffic theory*, 1993.
[66] F. Sagberg, G. F. B. Piccinini, J. Engström *et al.*, "A review of research on driving styles and road safety," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 7, pp. 1248–1275, 2015.
[67] S. Häkkinen, "Traffic accidents and driver characteristics: A statistical and psychological study." DTIC Document, Tech. Rep., 1958.
[68] C. Bishop, "Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn," 2007.
[69] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.
[70] H. E. Russell, L. K. Harbott, I. Nisky, S. Pan, A. M. Okamura, and J. C. Gerdes, "Motor learning affects car-to-driver handover in automated vehicles," *Science Robotics*, vol. 1, no. 1, p. eaah5682, 2016.
[71] W. Wang, J. Xi, C. Liu, and X. Li, "Human-centered feed-forward control of a vehicle steering system based on a driver's path-following characteristics," *IEEE Transactions on Intelligent Transportation Systems*, DOI: 10.1109/TITS.2016.2606347.
[72] J. C. Ferreira, J. de Almeida, and A. R. da Silva, "The impact of driving styles on fuel consumption: a data-warehouse-and-data-mining-based discovery process," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2653–2662, 2015.
[73] I. Karl, G. Berg, F. Ruger, and B. Farber, "Driving behavior and simulator sickness while driving the vehicle in the loop: validation of longitudinal driving behavior," *IEEE intelligent transportation systems magazine*, vol. 5, no. 1, pp. 42–57, 2013.
[74] M. Akamatsu, P. Green, and K. Bengler, "Automotive technology and human factors research: Past, present, and future," *International journal of vehicular technology*, vol. 2013, 2013.

Wenshuo Wang (S'15) received his B.S. in Transportation Engineering from ShanDong University of Technology, Shandong, China, in 2012. He is a Ph.D. candidate for Mechanical Engineering, Beijing Institute of Technology (BIT). Now he is a visiting scholar studying in the School of Mechanical Engineering, University of California at Berkeley (UCB). His work focuses on modeling and recognizing drivers behavior, making intelligent control systems between human driver and vehicle.

Chang Liu (S'15) received the B.S. degree in Electrical Engineering and B.S. degree in Applied Mathematics from Peking University, China, in 2011. He received the M.S. degree in Mechanical Engineering from the University of California at Berkeley, CA, USA, in 2014, where he is also currently working toward the Ph.D. degree in Mechanical Engineering. He is a Graduate Student Researcher with the Vehicle Dynamics and Control Laboratory headed by Prof. J Karl Hedrick. His research interests include robot path planning, distributed estimation and human-robot collaboration.



Ding Zhao received his Ph.D. degree in 2016 from the University of Michigan, Ann Arbor. He is currently an Assistant Research Scientist at Mechanical Engineering of the University of Michigan. His research interests include the autonomous vehicles, intelligent transportation, connected vehicles, dynamics and control, human-machine interaction, machine learning, and big data analysis.