# Location and Driver-Specific Vehicle Adaptation Using Crowdsourced Data

Menner, Marcel; Ma, Ziyi; Berntorp, Karl; Di Cairano, Stefano

## Abstract

This paper presents a method that adjusts the operation of advanced driver-assistance systems (ADAS) to a specific location and driver. The method uses crowdsourced data collected from multiple drivers in multiple locations/environments to predict the vehicle behavior of an individual driver in a previously unseen location/environment. This prediction can in turn be used for adapting the calibration of ADAS to the specific location/environment, as well as to the individual driver. We describe an algorithm for making predictions, which uses probabilities and quantile functions of empirical cumulative distribution functions to relate an individual driver to the population. The paper reports a simulation study in SUMO (Simulation of Urban MObility), where an emergency braking system is adapted to individual drivers and to different road surface conditions. The results show that the algorithm is quickly able to make accurate predictions and consequently adjust ADAS to the specific location and drive

# Location and Driver-Specific Vehicle Adaptation Using Crowdsourced Data

Marcel Menner, Ziyi Ma, Karl Berntorp, and Stefano Di Cairano

*Abstract*—This paper presents a method that adjusts the operation of advanced driver-assistance systems (ADAS) to a specific location and driver. The method uses crowd-sourced data collected from multiple drivers in multiple locations/environments to predict the vehicle behavior of an individual driver in a previously unseen location/environment. This prediction can in turn be used for adapting the calibration of ADAS to the specific location/environment, as well as to the individual driver. We describe an algorithm for making predictions, which uses probabilities and quantile functions of empirical cumulative distribution functions to relate an individual driver to the population. The paper reports a simulation study in SUMO (Simulation of Urban MObility), where an emergency braking system is adapted to individual drivers and to different road surface conditions. The results show that the algorithm is quickly able to make accurate predictions and consequently adjust ADAS to the specific location and driver.

## I. INTRODUCTION

An advanced driver-assistance system (ADAS) aims to increase safety of a human-operated vehicle by partially taking control of the vehicle. This increase in safety may come at the expense of comfort if the ADAS is configured too conservatively. Consequently, ADAS aims to increase safety while not decreasing the comfort of the human driver. However, the feeling of comfort may vary between individual human drivers. Further, driving conditions such as traffic or environmental conditions may vary, while ADAS is typically calibrated at production of the vehicle and not adjusted online. This paper presents a method to adapt the calibration of ADAS to the human driver, and to a specific location or environmental condition, using crowdsourced data.

Vehicle connectivity and crowdsourcing offer a new potential for storing and processing data of past and current driving patterns collected from multiple drivers [1]. These data can be utilized to determine patterns of behavior across drivers, e.g., patterns that impact all drivers, such as the weather or accidents. Furthermore, by providing a vehicle with such data, the operation of ADAS can be individualized by contrasting the individual driver with a group of drivers, while taking into consideration effects that impact all traffic participants. In such crowdsourcing applications, anonymity is key to ensure privacy, which is achieved by data aggregation and lack of identifier-based data labeling.

This paper presents a procedure for predicting the behavior of an individual driver using crowdsourced data. The algorithm uses anonymized population data provided, e.g., by a cloud storage as well as driver-specific data, see Fig. 1 for

All authors are with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, 02139, USA (e-mail: menner@ieee.org).
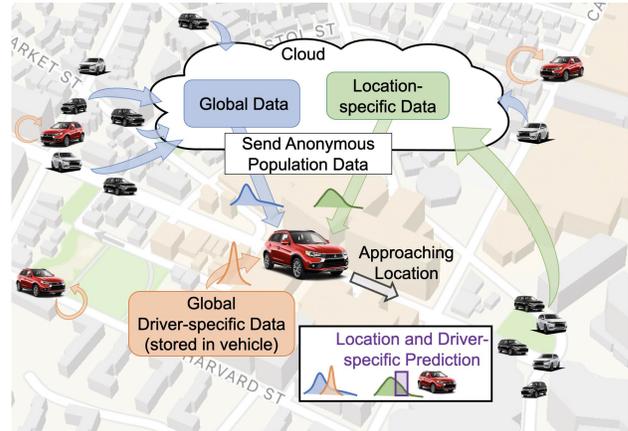
Fig. 1. Illustration of location and driver-specific prediction-making. The global data (blue) reflect behaviors of multiple drivers in different locations. The location-specific data (green) reflect behaviors of multiple drivers in a specific location of interest for the ego vehicle (displayed in red). The global driver-specific data (orange) reflect behaviors of the ego vehicle collected in different locations. These three data sets are used to make location and driver-specific predictions, which are used to adapt ADAS calibration.

an illustration. It uses (i) global data collected from multiple drivers in multiple locations (or environments), (ii) location-specific data collected from multiple drivers in a specific location for which we want to make predictions, and (iii) global driver-specific data collected from an individual driver in multiple locations. The algorithm proposed in this paper uses such data sources to construct three empirical cumulative distribution functions, and to predict the distribution of the individual driver in the specific location. This can in turn be used for adapting the calibration of ADAS. The proposed algorithm is based on the underlying assumption that an individual driver ranking at a certain percentile in multiple locations compared to the population, also ranks at the same percentile in the specific location of interest, compared to the population. Although the distribution functions of data sets (i) and (ii) are available for every driver to use, the data are not labeled. Hence, the data cannot be dissected and assigned to a driver, which protects their privacy. The distribution function in (iii) is only stored locally in the vehicle of the individual driver. The algorithm is verified using SUMO (Simulation of Urban MObility) [2], which is a commonly-used microscopic traffic simulation tool, and it has been extended to simulate connected vehicle scenarios [3]–[5]. We implement a case study in SUMO, where crowdsourced data are used to predict a location-specific and driver-specific distribution of the braking distance, which is then used for

adapting the calibration of an automated emergency braking (AEB) system. This ensures that AEB is activated at a distance where the driver does not perceive the AEB as false position, i.e., braking too early, or false negative, i.e., braking too late, while still ensuring safety.

*1) Related work:* Vehicle connectivity has been extensively investigated for road safety [6]–[8], congestion prediction or mitigation [9], [10], and traffic signal control [11]–[14]. Further, there is an increased interest in utilizing connectivity for emerging technologies and progress in autonomous driving and electric vehicles. For example, [10] uses autonomous vehicles to indirectly control human-operated vehicles using connectivity. In [15], the impact of vehicle connectivity for improved energy management of plug-in hybrid vehicles is discussed. Other applications include pavement condition monitoring [16] and travel time estimation [17], [18]. In addition, there is research in designing communication protocols between vehicles, infrastructures, and pedestrian, see, e.g., [19], [20].

Related crowdsourcing methods use consensus-type algorithms, decentralized blockchain technology, and location obfuscation to ensure the integrity and privacy of data [21]–[25]. For example, in [21] a privacy-preserving system is proposed that guarantees message trustworthiness in vehicle-to-vehicle communications, and [22] uses decentralized blockchain technology for privacy-preserving spatial crowdsourcing. In [23], a location obfuscation strategy is proposed to minimize the information loss due to obfuscation using geo-indistinguishability. In [24], [25], crowdsourced data are used for the routing of connected vehicles. In this work, we propose an algorithm for making location and driver-specific predictions and using such predictions for adaptive ADAS calibration, which appears not to have been investigated yet.

*2) Notation:* $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean $\mu$ and variance $\sigma^2$. The notation $x \sim \mathcal{N}(\mu, \sigma^2)$ means $x$ sampled from $\mathcal{N}(\mu, \sigma^2)$. The cumulative distribution function $F_X$ of a distribution $p(X)$ is defined as

$$\tilde{p} = \text{Prob}(X \leq \tilde{x}) = F_X(\tilde{x}) = \int_{-\infty}^{\tilde{x}} p(X = x)\mathrm{d}x,$$

where $\tilde{p}$ is the probability that the random variable $X$ is less than or equal to $\tilde{x}$, and $\tilde{x} = F_X^{-1}(\tilde{p})$ is the inverse cumulative distribution function. For the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$,

$$\tilde{p} = F_X(\tilde{x}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\tilde{x} - \mu}{\sqrt{2}\sigma}\right)\right]$$

and $\tilde{x} = F_X^{-1}(\tilde{p}) = \mu + \sqrt{2}\sigma\,\text{erf}^{-1}(2\tilde{p} - 1)$, where erf and $\text{erf}^{-1}$ are the error function and inverse error function.

Given $n$ data points $\{x_i\}_{i=1}^n$ with $x_i \in \mathbf{R}$, let $\hat{F}(x)$ be an empirical cumulative distribution function,

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i \leq x} \quad \text{with } \mathbf{1}_{x_i \leq x} = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{else} \end{cases} \quad (1)$$

where $n$ is the number of data samples, $x_i$, and $\mathbf{1}_{x_i \leq x}$ denotes the indicator function, i.e., $\hat{F}(x)$ is a step function increasing by $1/N$ at each data sample point, $x_i$.

## II. MATHEMATICAL PROBLEM FORMULATION

In this paper, we use probability distributions to predict the location-specific and driver-specific behavior of a vehicle. Probability distributions are suitable in this context as data are expected to be noisy due to sensor uncertainty, disturbances, as well as human factors. The goal of the method in this paper is to predict a quantity or the range of a quantity, $X$, which may vary dependent on the location, $L$, of the vehicle as well as on the driver, $D$. For instance, $X$ could denote the fuel/energy consumption of a vehicle, which varies based on the driver due to different vehicles or individual driving styles [26], and based on the location due to congestion, temperature, weather, etc. Hence, we want to predict the probability distribution of $X \in \mathbf{R}$ for a specific driver, $D$, at a specific location, $L$,

$$p(X|D, L). \quad (2)$$

To predict (2), we assume access to three data sets. First, we have a data set with $n_{x|d}$ data points, $\mathcal{X}_{X|D} = \{x_{X|D}^1, ..., x_{X|D}^{n_{x|d}}\}$, for a specific driver, $D$, obtained at multiple locations, i.e., samples from the distribution

$$x_{X|D}^i \sim p(X|D), \quad (3a)$$

which are stored locally on the vehicle to protect the privacy of the driver. Referring back to Fig. 1, this data set is illustrated in orange. Second, we have a crowdsourced data set with $n_x$ data points, $\mathcal{X}_X = \{x_X^1, ..., x_X^{n_x}\}$, e.g., sent from a cloud, collected from multiple drivers at multiple locations, i.e., samples from the distribution

$$x_X^i \sim p(X). \quad (3b)$$

This data set is shown in blue in Fig. 1. Third, we have a crowdsourced data set with $n_{x|l}$ data points, $\mathcal{X}_{X|L} = \{x_{X|L}^1, ..., x_{X|L}^{n_{x|l}}\}$, collected from multiple drivers at a specific location, $L$,

$$x_{X|L}^i \sim p(X|L). \quad (3c)$$

This data set is shown in green in Fig. 1. In (3), $x_{X|D}^i$, $x_X^i$, and $x_{X|L}^i$ denote the $i$th samples of the same quantity, but originating from different distributions, which are indicated by subscripts.

In summary, the goal is to predict the range of the quantity, $X$, as in (2) using the three data sets $\mathcal{X}_{X|D}$ (stored in vehicle), $\mathcal{X}_X$ (stored in cloud), and $\mathcal{X}_{X|L}$ (stored in cloud).

## III. PREDICTION ALGORITHM

In this section, we present an algorithm that uses cumulative distributions functions to predict $X$ for a specific location, $L$, and a specific driver, $D$, as in (2) based on samples of the distributions in (3). The main modeling assumption for the algorithm is that individual human drivers behave "always in the same way" relative to the population of drivers, which is detailed in the following. Mathematically, for any percentage value $\tilde{p} \in (0, 1)$ of the driver distribution with associated value $\tilde{x}_{X|D} = F_{X|D}^{-1}(\tilde{p})$, we compute a

"ranking" of the individual driver (distribution (2)) with respect to the population (distribution (3b)) as

$$\tilde{p}_X = F_X\left(\tilde{x}_{X|D}\right) \tag{4a}$$

$$\tilde{x}_{X|D} = F_{X|D}^{-1}(\tilde{p}), \tag{4b}$$

where $\tilde{x}_{X|D}$ is both the $\tilde{p}$th percentile of the global driver-specific distribution and the $\tilde{p}_X$th percentile of the global population distribution. Hence, we associate the percentile $\tilde{p}$ of the global driver-specific distribution with a percentile $\tilde{p}_X$ of the global population distribution.

To predict (2), we model the association of percentiles of $\tilde{p}$ and $\tilde{p}_X$ in (4) to be similar for the location-specific distributions,

$$\tilde{p}_{X|L} = F_{X|L}\left(\tilde{x}_{X|D,L}\right) \tag{5a}$$

$$\tilde{x}_{X|D,L} = F_{X|D,L}^{-1}(\tilde{p}), \tag{5b}$$

where $\tilde{x}_{X|D,L}$ is both the $\tilde{p}$th percentile of the location and driver-specific distribution and the $\tilde{p}_{X|L}$th percentile of the location-specific population distribution. However, $F_{X|D,L}$ is unknown. To predict $F_{X|D,L}$, we use

$$\tilde{p}_X = \tilde{p}_{X|L} \tag{6}$$

to connect (4) and (5), thus obtaining

$$x_{X|D,L}^{\text{pred}} = F_{X|L}^{-1}\left(F_X\left(F_{X|D}^{-1}(\tilde{p})\right)\right) \quad \forall \tilde{p} \in (0,1), \tag{7}$$

where $x_{X|D,L}^{\text{pred}}$ is the predicted $\tilde{p}$th percentile of the location and driver-specific distribution and can be used for adaptation of ADAS. In other words, this approach is based on the assumption that the ranking of an individual driver with respect to the population is consistent across all different locations. Eq. (7) requires access to the three cumulative distribution functions $F_{X|L}$, $F_X$, and $F_{X|D}$, which is addressed next.

*Remark 1:* It is easy to see that this ranking assumption holds true in expectation. E.g., a generally more aggressive driver is predicted to be more aggressive in the specific location, whereas a cautious driver is predicted to act cautiously.

### A. Empirical Cumulative Distribution Functions

In practice, the cumulative distribution functions can have very complex shapes and may not be well-represented by a specific class of distributions such as Gaussian, Laplace, etc. As a result, we use empirical cumulative distribution functions, which offer the advantage of avoiding assumptions about probability distribution classes and enable flexible shapes of the distributions. This model-free/data-driven solution approximates an unknown cumulative distribution function $F(x)$ with an empirical cumulative distribution function $\hat{F}(x)$ as in (1). By the law of large numbers, $\hat{F}(x) \xrightarrow{\text{a.s.}} F(x)$, i.e., converges (almost surely) as $n \to \infty$, see [27]. It is easy to see that the prediction $\hat{F}(x)$ improves in expectation as more data are accumulated.

Thus, our algorithm works as follows. For any $\tilde{p}$, we predict the associated value using the three empirical cumulative distribution functions according to (7) resulting in

$$x_{X|D,L}^{\text{pred}} = \hat{F}_{X|D,L}^{-1}(\tilde{p}) = \hat{F}_{X|L}^{-1}\left(\hat{F}_X\left(\hat{F}_{X|D}^{-1}(\tilde{p})\right)\right),$$

where $\hat{F}_{X|L}$, $\hat{F}_X$, and $\hat{F}_{X|D}$ are provided by the data sets $\mathcal{X}_{X|L}$, $\mathcal{X}_X$, and $\mathcal{X}_{X|D}$, respectively. Algorithm 1 shows how to construct the empirical cumulative distribution function $\hat{F}_{X|D,L}$ by means of a set of predicted data points, $\mathcal{X}_{X|D,L}$.

---

**Algorithm 1** Predict location/driver-specific $\hat{F}_{X|D,L}$

---

**Require:** Data sets $\mathcal{X}_{X|D}$, $\mathcal{X}_X$, $\mathcal{X}_{X|L}$
1: **for** $\tilde{p} = 1\%$, ..., $99\%$ **do**
2: $\quad x_{X|D,L}^{\text{pred}} = \hat{F}_{X|L}^{-1}\left(\hat{F}_X\left(\hat{F}_{X|D}^{-1}(\tilde{p})\right)\right)$
3: $\quad \mathcal{X}_{X|D,L} \leftarrow \{\mathcal{X}_{X|D,L}, x_{X|D,L}^{\text{pred}}\}$ $\quad \triangleright$ Add to data set
4: **end for**
5: **return** $\mathcal{X}_{X|D,L}$ $\quad \triangleright$ Data for predicted distribution

---

### B. Selected Theoretical Properties

For certain distributions in (2) and (3), the proposed estimator is consistent, i.e., the prediction can reconstruct the underlying ground-truth distribution exactly as $n \to \infty$. E.g., let (2) and (3) be Gaussian distributions with

$$p(X|D) = \mathcal{N}(\mu_{x|d}, \sigma_{x|d}^2) \tag{8a}$$

$$p(X) = \mathcal{N}(\mu_x, \sigma_x^2) \tag{8b}$$

$$p(X|L) = \mathcal{N}(\mu_{x|l}, \sigma_{x|l}^2) \tag{8c}$$

$$p(X|D,L) = \mathcal{N}\left(\mu_{x|d,l}, \sigma_{x|d,l}^2\right), \tag{8d}$$

$\mu_{x|d,l} = \mu_{x|l} + \frac{\sigma_{x|l}}{\sigma_x}(\mu_{x|d} - \mu_x)$, $\sigma_{x|d,l}^2 = \left(\frac{\sigma_{x|l}\sigma_{x|d}}{\sigma_x}\right)^2$. Then,

$$F_{X|D,L}^{-1}(\tilde{p}) = x_{X|D,L}^{\text{pred}} = F_{X|L}^{-1}\left(F_X\left(F_{X|D}^{-1}(\tilde{p})\right)\right), \tag{9}$$

for all $\tilde{p} \in (0,1)$, i.e., $F_{X|D,L}$ is reconstructed exactly, which is formally proven in Theorem 1.

*Theorem 1:* Let (2) and (3) be as in (8). Then, (9) holds for all $\tilde{p} \in (0,1)$.

*Proof:* Using the definitions of the cumulative distribution function and its inverse,

$$x_{X|D} = F_{X|D}^{-1}(p) = \mu_{x|d} + \sigma_{x|d}\sqrt{2}\text{erf}^{-1}(2p-1) \tag{10a}$$

$$p_X = F_X(x_{X|D}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x_{X|D} - \mu_x}{\sqrt{2}\sigma_x}\right)\right] \tag{10b}$$

$$x_{X|D,L} = F_{X|L}^{-1}(p_X) = \mu_{x|l} + \sigma_{x|l}\sqrt{2}\text{erf}^{-1}(2p_X - 1), \tag{10c}$$

where erf is the error function. Inserting (10b) into (10c),

$$x_{X|D,L} = \mu_{x|l} + \sigma_{x|l}\sqrt{2}\left(\frac{x_{X|D} - \mu_x}{\sqrt{2}\sigma_x}\right). \tag{11}$$

Using $x_{X|D}$ in (10a), (11) can be reformulated as

$$x_{X|D,L} = \mu_{x|l} + \sigma_{x|l}\sqrt{2}\left(\frac{\mu_{x|d} + \sigma_{x|d}\sqrt{2}\text{erf}^{-1}(2p-1) - \mu_x}{\sqrt{2}\sigma_x}\right)$$
$$= \mu_{x|l} + \frac{\sigma_{x|l}}{\sigma_x}(\mu_{x|d} - \mu_x) + \frac{\sigma_{x|l}\sigma_{x|d}}{\sigma_x}\sqrt{2}\text{erf}^{-1}(2p-1).$$

Hence, by definition of the cumulative distribution function,

$$p(X|D,L) = \mathcal{N}(\mu_{x|d,l}, \sigma_{x|d,l}^2)$$

with $\mu_{x|d,l} = \mu_{x|l} + \frac{\sigma_{x|l}}{\sigma_x}(\mu_{x|d} - \mu_x)$, $\sigma_{x|d,l} = \frac{\sigma_{x|l}\sigma_{x|d}}{\sigma_x}$, which show (9) for all percentages $p$ and distributions in (8). $\blacksquare$

*Remark 2:* The Gaussian assumption makes theoretical investigations easier. However, the presented method does not require Gaussian distributions. Studying more general theoretical properties exceeds the scope of this paper.

## IV. SIMULATION RESULTS

### A. Convergence of Predicted Distribution

First, we investigate how quickly the predicted distribution converges to the ground-truth distribution as data are gathered. Here, let $p(X|D)$ and $p(X)$ as in (8a)–(8b) be known. This amounts to having observed sufficiently many data points such that the empirical cumulative distribution functions approximate (8a)–(8b) well. In practice, this means having constructed the distributions using multiple trips of multiple drivers and multiple trips of the target driver. Then, we collect samples of the distribution (8c) and the goal of the proposed algorithm is to predict (8d).

We performed 2000 Monte Carlo simulations where at each iteration, we first obtained the means and variances of (8a), (8b), (8c) as $\mu_x \sim \mathcal{N}(0, 10)$, $\sigma_x \sim \mathcal{N}(5, 25)$, $\mu_{x|d} \sim \mathcal{N}(0, 5)$, $\sigma_{x|d} \sim \mathcal{N}(3, 9)$, $\mu_{x|l} \sim \mathcal{N}(0, 10)$, $\sigma_{x|l} \sim \mathcal{N}(5, 25)$. This ensures that a wide range of different scenarios are simulated. For each of the 2000 trials, the means and variances remain constant while samples are drawn. During each trial, we draw samples from (8c) to build the empirical cumulative distribution function, $\hat{F}_{X|L}$. Then, we predict (8d) using Algorithm 1 after each sample has been drawn and compute the prediction error as

$$\text{err} = \frac{\sum_{\tilde{p}=1\%}^{99\%}\left(F_{X|D,L}^{-1}(\tilde{p}) - \hat{F}_{X|L}^{-1}\left(F_X\left(F_{X|D}^{-1}(\tilde{p})\right)\right)\right)^2}{\sum_{\tilde{p}=1\%}^{99\%}\left(F_{X|D,L}^{-1}(\tilde{p})\right)^2}.$$

Fig. 2 shows the convergence results of the error, by the median, the 20th, and 80th percentile of the 2000 Monte-Carlo runs. The algorithm reduces the error quickly, e.g., see median of err $= 1$ after the first sample and err $= 0.02$ after 100 samples. As a comparison, Fig. 2 also shows the convergence of the error that we would obtain if we could directly observe samples from $F_{X|D,L}$ (rather than $F_{X|L}$). Since we cannot observe samples from $F_{X|D,L}$, this is an oracle and is expected to have faster convergence. The simulation results show that the proposed algorithm converges only slightly slower than the ideal case of the oracle, which indicates that the algorithm is able to quickly predict the unknown distribution.

### B. ADAS Adaptation Case Study using SUMO

Next, we present a case study using SUMO [2], where drivers are traveling on a straight road with a speed limit of 50km/h and need to stop at a traffic light. The vehicles' behaviors are simulated using the Krauss car-following model [28]–[30], which is calibrated using a range of different parameters to simulate different drivers. For this case study, we generate 100 different drivers with the parameters sampled from distributions and value ranges in Table I. The parameters `accel` and `decel` define the acceleration and
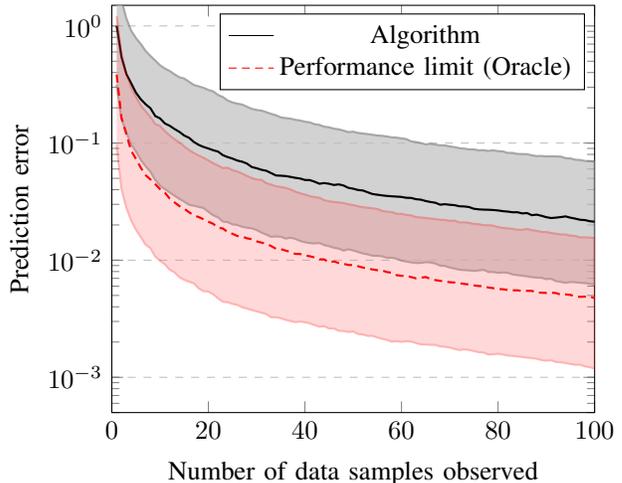


Fig. 2. Convergence of prediction error. The prediction error (displayed in black) reduces quickly from a median of err $= 1$ after the first sample and err $= 0.02$ after 100 samples. The hypothetical convergence rate of an oracle observing samples directly (displayed in red) is only 9.5% faster, which can be expected to be the performance limit.

TABLE I
VEHICLE PARAMETERS FOR SUMO

| Parameter | Range | Distribution |
|---|---|---|
| `accel` | [0, 9.81] | Beta ($\alpha = 2$, $\beta = 5$) |
| `decel` | [0, 9.81] | Beta ($\alpha = 2$, $\beta = 5$) |
| `sigma` | [0, 1] | Truncated Gauss ($\mu = 0.2$, $\sigma = 0.5$) |
| `tau` | [0.5, 1.6] | Truncated Gauss ($\mu = 0.6$, $\sigma = 0.5$) |
| `speedFactor` | [0.8, 1.2] | Truncated Gauss ($\mu = 1.1$, $\sigma = 0.2$) |
| `speedDev` | [0, 0.2] | Truncated Gauss ($\mu = 0.1$, $\sigma = 0.05$) |

deceleration ability of vehicles, respectively; `sigma` defines a driver's imperfection; `tau` defines the driver's desired time headway; `speedFactor` defines the vehicle's expected multiplicator for speed limits; and `speedDev` defines the deviation of the `speedFactor`. For more detail, the reader is referred to the SUMO documentation [31]. For our case study, we modify the parameters `accel` and `decel` to account for different driving behaviors in different environmental conditions. Here, we choose a simple scaling, i.e., `accel`$\leftarrow \mu\cdot$`accel` and `decel`$\leftarrow \mu\cdot$`decel`, where $\mu$ is the friction coefficient. This scaling considers that drivers are more cautious in conditions with lower friction coefficients.

Fig. 3 shows how the algorithm predicts the location-specific and driver-specific cumulative distribution function. In this case study, the "location" refers to snowy road conditions with friction coefficient $\mu = 0.4$, whereas "global" refers to data collected for multiple different friction coefficients, $\mu \in [0.3, 0.9]$. Fig. 3 illustrates the assumption on consistent rankings in (6) for $\tilde{p} = 50\%$ and $\tilde{p}_X = \tilde{p}_{X|L} = 42\%$, i.e., a driver ranking at the 42th percentile w.r.t. the population also ranks at the 42th percentile w.r.t. the population on a snow-covered road. Fig. 3 shows an exemplary trial, where 50 data points have been collected for the location-specific population data, and 1000 data points for both global data sets, i.e, the left column in Fig. 3.
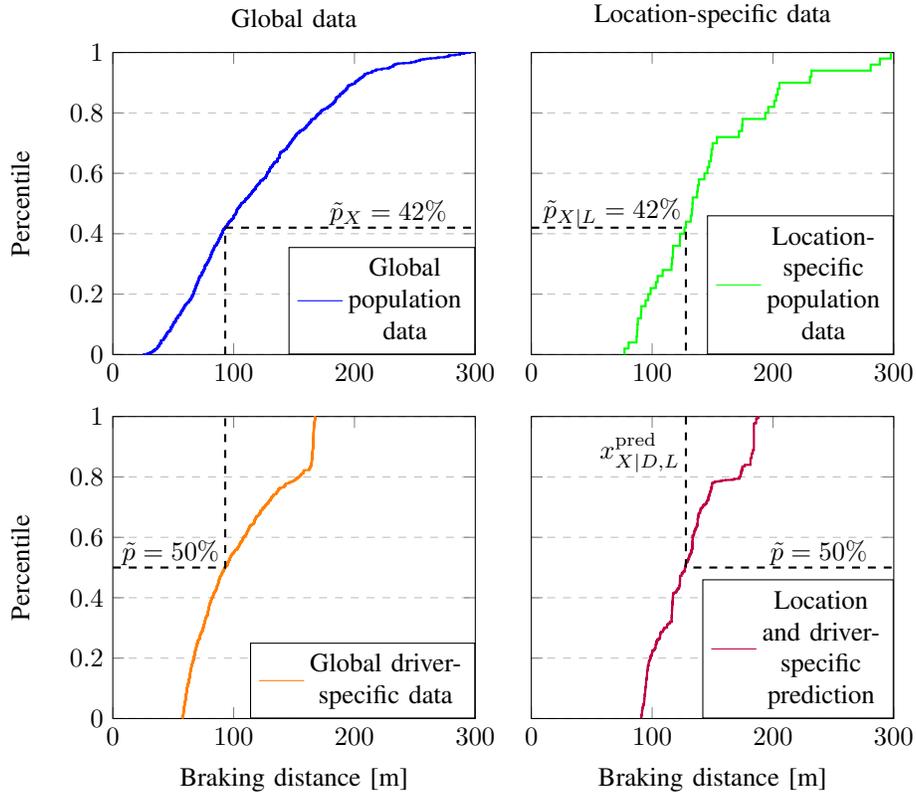
Fig. 3. Illustration of empirical cumulative distribution functions. Top row: population data. Bottom row: driver-specific data or prediction. Left column: global data. Right column: location-specific data or prediction. The dashed lines indicate the procedure of the proposed algorithm.

To increase the statistical significance of the results, we conduct 1000 trials similar to Fig. 3. Fig. 4 reports the mean and standard deviation of braking distances over the 1000 trials. Fig. 4 shows how quickly the prediction of the 50th percentile, i.e., $\tilde{p} = 50\%$, converges for three drivers in two different locations/environment, which simulate driving on snow with friction coefficient $\mu = 0.4$ and driving on dry asphalt, $\mu = 0.9$. First, having observed only one data point, the algorithm will not be able to make driver-specific predictions (see, braking distance of 90 m for $\mu = 0.9$ and 155 m $\mu = 0.4$ for all drivers). However, in expectation, the algorithm is able to correctly predict the impact of the different friction coefficients. Further, after collecting 10 data points at the specific location with either dry asphalt or snow, the algorithm is able to distinguish between the different drivers and the specific location more reliably. After collecting 100 data points, the algorithm is able to distinguish between drivers and locations, as all predicted braking distances are well-separated by more than one standard deviation. Note that 100 data points are collected when 100 (connected) vehicles have gone through the specific location. For roads with at least medium level of traffic, this happens in the order of minutes, which is appropriate for our purposes.

Lastly, we utilize the prediction in Fig. 3 to adapt the calibration of an AEB system in SUMO, where we use the predicted braking distance $x_{X|D,L}^{\mathrm{pred}}$ for percentile $\tilde{p} = 2\%$, i.e., the automatic emergency braking system is triggered

at the distance $x_{X|D,L}^{\mathrm{pred}}$ from the stopping point. Note that other percentiles can be used as well, e.g., $\tilde{p} = 20\%$ can be used to trigger a warning to the driver. One advantage of the proposed procedure for adapting the calibration of such an AEB system is that safety is ensured by design, as long as the data are nonadversarial. This is because the location-specific population data can be viewed as a filter, which only allows predictions that are physically possible, i.e., filtered by actual braking distances that were observed at the specific location or the specific friction coefficient, under the assumption that the AEB is not worse than any driver.

## V. CONCLUSION

This paper presented a concept to make location-specific and driver-specific predictions of a vehicle's behavior. The concept leverages crowdsourced data collected from multiple locations and multiple drivers and can be used in order to adapt the calibration of advanced driver assistance systems to a specific location and a specific driver. In addition, we developed an algorithm/estimator based on three empirical cumulative distribution functions of (i) global population data, (ii) location-specific population data, and (iii) global driver-specific data in order to make predictions. The algorithm is based on a consistent ranking assumption, which models that drivers behave similar with respect to the population for all locations. We showed conditions for which the estimator is consistent, i.e., the estimator reconstructs the underlying unknown distribution exactly. A simulation
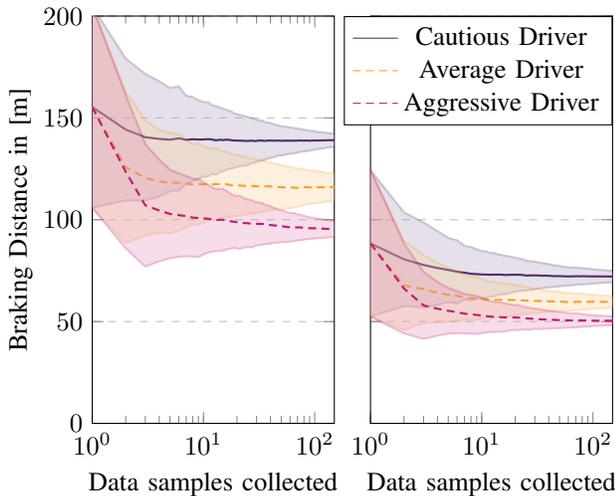
Fig. 4. Convergence of braking distance prediction (50th percentile). Left: Predictions on snowy road with $\mu = 0.4$. Right: Predictions on asphalt with $\mu = 0.9$. Three drivers are compared, where the cautious driver is displayed in purple, the average driver in yellow, and the aggressive driver in red. The plots display mean as well as standard deviation over 1000 trials.

study showed that the algorithm is fast to converge and requires little data. Finally, we presented a case study in SUMO in which an emergency braking system was adapted using location-specific and driver-specific predictions of the braking distance.

## REFERENCES

[1] J. E. Siegel, D. C. Erb, and S. E. Sarma, "A survey of the connected vehicle landscape—architectures, enabling technologies, applications, and development areas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 8, pp. 2391–2406, 2018.

[2] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using SUMO," in *21st IEEE International Conference on Intelligent Transportation Systems*, 2018.

[3] V. Milanés and S. E. Shladover, "Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data," *Transportation Research Part C: Emerging Technologies*, vol. 48, pp. 285–300, 2014.

[4] C. Lei, E. M. van Eenennaam, W. K. Wolterink, J. Ploeg, G. Karagiannis, and G. Heijenk, "Evaluation of CACC string stability using sumo, simulink, and OMNeT++," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–12, 2012.

[5] M. Malinverno, G. Avino, C. Casetti, C.-F. Chiasserini, F. Malandrino, and S. Scarpina, "Performance analysis of C-V2I-based automotive collision avoidance," in *IEEE 19th International Symposium on A World of Wireless, Mobile and Multimedia Networks*, pp. 1–9, 2018.

[6] Y. Wu, M. Abdel-Aty, O. Zheng, Q. Cai, and L. Yue, "Developing a crash warning system for the bike lane area at intersections with connected vehicle technology," *Transportation Research Record: J. of the Transportation Research Board*, vol. 2673, no. 4, pp. 47–58, 2019.

[7] S. Son, J. Jeong, S. Park, and J. Park, "Effects of advanced warning information systems on secondary crash risk under connected vehicle environment," *Accident Analysis & Prevention*, vol. 148, p. 105786, 2020.

[8] M. Islam, M. Rahman, M. Chowdhury, G. Comert, E. D. Sood, and A. Apon, "Vision-based personal safety messages (PSMs) generation for connected vehicles," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9402–9416, 2020.

[9] R. Bauza, J. Gozalvez, and J. Sanchez-Soriano, "Road traffic congestion detection through cooperative vehicle-to-vehicle communications," in *IEEE Local Computer Network Conference*, pp. 606–612, 2010.

[10] C. Wu, A. M. Bayen, and A. Mehta, "Stabilizing traffic with autonomous vehicles," in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6012–6018, 2018.

[11] S. I. Guler, M. Menendez, and L. Meier, "Using connected vehicle technology to improve the efficiency of intersections," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 121–131, 2014.

[12] Y. Feng, K. L. Head, S. Khoshmagham, and M. Zamanipour, "A real-time adaptive signal control in a connected vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 460–473, 2015.

[13] Q. Guo, L. Li, and X. J. Ban, "Urban traffic signal control with connected and automated vehicles: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 313–334, 2019.

[14] Z. Yao, Y. Jiang, B. Zhao, X. Luo, and B. Peng, "A dynamic optimization method for adaptive signal control in a connected vehicle environment," *Journal of Intelligent Transportation Systems*, vol. 24, no. 2, pp. 184–200, 2019.

[15] C. M. Martinez, X. Hu, D. Cao, E. Velenis, B. Gao, and M. Wellers, "Energy management in plug-in hybrid electric vehicles: Recent progress and a connected vehicles perspective," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 4534–4549, 2017.

[16] E. P. Dennis, Q. Hong, R. Wallace, W. Tansil, and M. Smith, "Pavement condition monitoring with crowdsourced connected vehicle data," *Transportation Research Record*, vol. 2460, no. 1, pp. 31–38, 2014.

[17] L. Lin, W. Li, and S. Peeta, "Efficient data collection and accurate travel time estimation in a connected vehicle environment via real-time compressive sensing," *Journal of Big Data Analytics in Transportation*, vol. 1, no. 2, pp. 95–107, 2019.

[18] J. Zhou, X. Lai, and J. Y. Chow, "Multi-armed bandit on-time arrival algorithms for sequential reliable route selection under uncertainty," *Transportation Research Record*, vol. 2673, no. 10, pp. 673–682, 2019.

[19] X. Yang, L. Liu, N. H. Vaidya, and F. Zhao, "A vehicle-to-vehicle communication protocol for cooperative collision warning," in *The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, pp. 114–123, 2004.

[20] F. Lyu, H. Zhu, N. Cheng, H. Zhou, W. Xu, M. Li, and X. Shen, "Characterizing urban vehicle-to-vehicle communications for reliable safety applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 6, pp. 2586–2602, 2019.

[21] Q. Wu, J. Domingo-Ferrer, and U. González-Nicolás, "Balanced trustworthiness, safety, and privacy in vehicle-to-vehicle communications," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 2, pp. 559–573, 2009.

[22] J. Zhang, F. Yang, Z. Ma, Z. Wang, X. Liu, and J. Ma, "A decentralized location privacy-preserving spatial crowdsourcing for internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[23] C. Qiu, A. C. Squicciarini, C. Pang, N. Wang, and B. Wu, "Location privacy protection in vehicle-based spatial crowdsourcing via geo-indistinguishability," *IEEE Transactions on Mobile Computing*, 2020.

[24] G. Russo, "On the crowdsourcing of behaviors for autonomous agents," *IEEE Control Systems Letters*, vol. 5, no. 4, pp. 1321–1326, 2020.

[25] É. Garrabé and G. Russo, "On the design of autonomous agents from multiple data sources," *IEEE Control Systems Letters*, vol. 6, pp. 698–703, 2021.

[26] M. Menner, K. Berntorp, M. N. Zeilinger, and S. D. Cairano, "Inverse learning for data-driven calibration of model-based statistical path planning," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 131–145, 2021.

[27] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge University Press, 2000.

[28] S. Krauß, P. Wagner, and C. Gawron, "Metastable states in a microscopic model of traffic flow," *Physical Review E*, vol. 55, no. 5, p. 5597, 1997.

[29] S. Krauß, "Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics," 1998.

[30] L. Lücken, E. Mintsis, N. P. Kalliroi, R. Alms, Y.-P. Flötteröd, and D. Koutras, "From automated to manual - modeling control transitions with sumo," in *SUMO User Conference 2019*, vol. 62, pp. 124–144, 2019.

[31] "SUMO online documentation." https://sumo.dlr.de/docs/Car-Following-Models.html. Accessed: 2021-10-18.