Delft University of Technology

Delft Center for Systems and Control

Technical report 19-019

Safety assessment of automated vehicles: How to determine whether we have collected enough field data?*

E. de Gelder, J.-P. Paardekooper, O. Op den Camp, and B. De Schutter

If you want to cite this report, please use the following reference instead:

E. de Gelder, J.-P. Paardekooper, O. Op den Camp, and B. De Schutter, "Safety assessment of automated vehicles: How to determine whether we have collected enough field data?," *Traffic Injury Prevention*, vol. 20, pp. S162–S170, 2019. doi:10.1080/15389588.2019.1602727

Delft Center for Systems and Control
Delft University of Technology
Mekelweg 2, 2628 CD Delft
The Netherlands
phone: +31-15-278.24.73 (secretary)
URL: https://www.dcsc.tudelft.nl

* This report can also be downloaded via https://pub.bartdeschutter.org/abs/19_019.html

Safety assessment of automated vehicles: how to determine whether we have collected enough field data?

Erwin de Gelder[a,b], Jan-Pieter Paardekooper[a], Olaf Op den Camp[a], and Bart De Schutter[c]

Abstract— Objective: **The amount of collected field data from naturalistic driving studies is quickly increasing. The data are used for, amongst others, developing automated driving technologies (such as crash avoidance systems), studying driver interaction with such technologies, and gaining insights into the variety of scenarios in real-world traffic. Since the collection of data is time consuming and requires high investments and resources, questions like "do we have enough data?", "how much more information can we gain when obtaining more data?", and "how far are we from obtaining completeness?" are highly relevant. In fact, deducing safety claims based on collected data, e.g., through testing scenarios based on collected data, requires knowledge about the degree of completeness of the data used. We propose a method for quantifying the completeness of the so-called activities in a dataset. This enables us to partly answer the aforementioned questions.**

Method: **In this paper, the (traffic) data are interpreted as a sequence of different so-called scenarios that can be grouped into a finite set of scenario classes. The building blocks of scenarios are the activities. For every activity, there exists a parametrization that encodes all information in the data of each recorded activity. For each type of activity, we estimate a probability density function (pdf) of the associated parameters. Our proposed method quantifies the degree of completeness of a data set using the estimated pdfs.**

Results: **To illustrate the proposed method, two different case studies are presented. First, a case study with an artificial dataset, of which the underlying pdfs are known, is carried out to illustrate that the proposed method correctly quantifies the completeness of the activities. Next, a case study with real-world data is performed to quantify the degree of completeness of the acquired data for which the true pdfs are unknown.**

Conclusion: **The presented case studies illustrate that the proposed method is able to quantify the degree of completeness of a small set of field data and can be used to deduce whether sufficient data have been collected for the purpose of the field study. Future work will focus on applying the proposed method to larger datasets. The proposed method will be used to evaluate the level of completeness of the data collection on Singaporean roads, aimed at defining relevant test cases for the autonomous vehicles' road-approval procedure that is being developed in Singapore.**

I. Introduction

The amount of collected field data from driving studies is increasing rapidly and these data are extensively used for the research, development, assessment, and evaluation of driving-related topics; for example, see Broggi et al. (2013), Dingus et al. (2016), Elrofai et al. (2018), Gelder and Paardekooper (2017), Klauer et al. (2006), Krajewski et al. (2018), Ploeg et al. (2018), Pütz et al. (2017), Sadigh et al. (2014), Williamson et al. (2011), and Zofka et al. (2015). For any work that depends on data, it is important to know how complete the data are. As mentioned by various authors (Alvarez et al. 2017; Geyer et al. 2014; Stellet et al. 2015), especially when deducing safety claims based on collected data, e.g., through testing scenarios based on collected data, we require knowledge about the degree of completeness of the dataset used. Hence, questions like "do we have enough data?" are highly relevant when our work and conclusions depend on the data. Furthermore, since the collection of data is time-consuming and requires high investments and resources, we should ask ourselves "how much more data do we need?" or "how much more information can we gain when obtaining more data?"

The aforementioned questions are already explored in other fields (Blair et al. 2004; Guest et al. 2006; Marks et al. 2018; Wang et al. 2017; Yang et al. 2012), but the question of how much data are enough regarding traffic-related applications is less frequently answered. Wang et al. (2017) appear to be the first in literature to point out and discuss issues concerning the amount of data needed to understand and model driver behaviors. They propose a statistical approach to determine how much naturalistic driving data are enough for understanding driving behaviors. For scenario-based assessments (Alvarez et al. 2017; Elrofai et al. 2018; Geyer et al. 2014; Ploeg et al. 2018; Stellet et al. 2015), however, the approach of Wang et al. (2017) might not be applicable, because they only consider the individual measurements at consecutive time instants instead of taking into account the whole driving scenario. Hence, there is a need for a quantitative measure for the completeness of a dataset that takes into account the different scenarios a vehicle encounters in real-world traffic.

We describe a method for quantifying the completeness of a data set. The data are interpreted as a sequence of different scenarios that can be grouped into a finite set of scenario classes. Activities, such as "braking" and "lane change," form the building blocks of the scenarios (Elrofai et al. 2018). For every activity, we create a parametrization that encodes the information in the data of this activity. For each type of activity, we estimate a probability density function (pdf) of the associated parameters. Our proposed method approximates the degree of completeness of a data set using the expected error of the estimated pdf. The smaller this error, the higher the degree of completeness.

To illustrate the proposed method, two different case studies are presented. The first case study involves artificial data of which the underlying distributions are known. Be-

[a] TNO, P.O. Box 756, 5700 AT Helmond, The Netherlands
[b] Corresponding author. Email: erwin.degelder@tno.nl
[c] Delft University of Technology, Delft Center for Systems and Control

cause the underlying distributions are known, we can show that the proposed method correctly quantifies the degree of completeness. Next, a case study with real-world data is performed to quantify the degree of completeness of the acquired data for which the underlying distributions are unknown. Additionally, we show how we can estimate the required amount of data to meet a certain requirement.

The article is structured as follows. In Section II, we describe in more detail what the problem for which a solution is proposed in Section III. The two case studies are presented in Section IV. After a discussion in Section V, this paper is concluded in Section VI.

II. Problem definition

The required amount of data depends on the use of the data (Wang et al. 2017). For example, when investigating (near)-accident scenarios from naturalistic driving data, more data might be required compared to studying nominal driving behavior, because of the low probability of having a (near)-accident scenario in naturalistic driving data. Therefore, in this paper, the goal is to define a quantitative measure for the completeness of the data that can be used to determine whether the data are enough.

To define the problem of quantifying the completeness of the data, few assumptions are made:

1) The data are interpreted as an endless sequence of scenarios, where scenarios might overlap in time (Elrofai et al. 2018). Several definitions of the term scenario in the context of traffic data have been proposed in literature, e.g., by Elrofai et al. (2018, 2016), Geyer et al. (2014), and Ulbrich et al. (2015). Because we want to differentiate between quantitative and qualitative descriptions, the definition of the term scenario is adopted from Elrofai et al. (2018) as it explicitly defines a scenario as a quantitative description: "A scenario is a quantitative description of the ego vehicle, its activities and/or goals, its dynamic environment (consisting of traffic environment and conditions) and its static environment. From the perspective of the ego vehicle, a scenario contains all relevant events." Extracting scenarios from data received significant attention and the applied methods are very diverse. For example, Elrofai et al. (2016) use a model-based approach to detect scenarios in which the ego vehicle is changing lane, whereas Kasper et al. (2012) use Bayesian networks to detect scenarios with lane changes of other vehicles around the ego vehicle. Xie et al. (2018) use a random forest classifier for extracting various scenarios and Paardekooper et al. (2019) employ rule-based algorithms for scenario extraction.
2) Just as Elrofai et al. (2018), we assume that a scenario consists of activities: "An activity is considered [to be] the smallest building block of the dynamic part of the scenario (maneuver of the ego vehicle and the dynamic environment)." An activity describes the time evolution of state variables. For example, an activity can be "braking", where the activity describes the evolution of the speed over time. Furthermore, "the end of an activity marks the start of the next activity" (Elrofai et al. 2018).
3) Though a scenario refers to a quantitative description, these scenarios can be abstracted by means of a qualitative description, referred to as scenario class; see also Elrofai et al. (2018) and Ploeg et al. (2018). An example of a scenario class could have the name "ego vehicle braking"; that is, this scenario qualitatively describes a scenario in which the ego vehicle brakes. An actual (real-world) scenario in which the ego vehicle is braking would fall into this scenario class. It is assumed that all scenarios can be categorized into these scenario classes. This assumption does not limit the applicability of this paper, though it might require a large number of scenario classes to describe all scenarios that are in the data.
4) It is assumed that all scenarios that fall into a specific scenario class can be parametrized similarly. As a result, the specific activities that constitute the scenario are also parametrized similarly. As with the previous assumption, this does not limit the applicability of this article. However, it might constrain the variety of scenarios that fall into a scenario class.

Using these assumptions, we can describe the problem of quantifying the completeness of a dataset into three subproblems:

1) How to quantify the completeness regarding the scenario classes?
2) How to quantify the completeness regarding all scenarios that fall into a specific scenario class?
3) How to quantify the completeness regarding the activities?

The first step towards quantification of the completeness of the data is to assess the completeness of the activities. The next step is to quantify the completeness of the scenarios, i.e., the combinations of activities. The final step is to quantify the completeness of the scenario classes. In this article, the first step, i.e., the third subproblem, is addressed. Because of the different approach required for answering the first and second subproblem, those will be addressed in a forthcoming paper.

III. Method

In this section, we present how to quantify the completeness regarding the activities. As explained in Section II, all scenarios that fall into a specific scenario class are parametrized similarly. Therefore, all similar types of activities are also parametrized similarly. For example, all activities labeled "braking" are parametrized similarly. In the remainder of this section, we assume that all activities that we are dealing with are a similar type of activities, such that they are parametrized similarly.

Let n denote the number of activities such that we have n parameter vectors that describe these activities, denoted by $X_i \in \mathbb{R}^d$ with $i \in \{1, \ldots, n\}$ and d denoting the number of parameters for one activity. We will estimate the underlying

distribution of X_i. Let $f(\cdot)$ denote the true probability density function (pdf) and let $f(x)$ denote the probability density evaluated at x. Similarly, let $\hat{f}(\cdot;n)$ denote the estimated pdf using n parameter vectors.

To quantify the completeness of the collection of the n activities, we use the estimated pdf $\hat{f}(\cdot;n)$. For example, suppose that $\hat{f}(x;n)$ equals $f(x)$ for all $x \in \mathbb{R}^d$. In this case, it would be reasonable to say that the n activities give a complete view of the variety and the distribution of the different activities that are labeled similarly. On the other hand, when $\hat{f}(x;n)$ is very different from $f(x)$, it would be reasonable to say that the opposite is the case, i.e., the n scenarios do not give a complete view. One common measure for comparing the estimated pdf with the true pdf is the Mean Integrated Squared Error (MISE):

$$\mathrm{MISE}_f(n) = \mathrm{E}\left[\int_{\mathbb{R}^d} \left(f(x) - \hat{f}(x;n)\right)^2 \mathrm{d}x\right]. \tag{1}$$

The index f indicates that the MISE is calculated with respect to the pdf $f(\cdot)$.

A low MISE indicates a high degree of completeness whereas a high MISE indicates a low degree of completeness, because the expected integrated squared error is high. Therefore, the MISE can be used to quantify the completeness of set of activities that are of a similar type. The problem is, however, that Eq. (1) depends on the true pdf $f(\cdot)$ which is unknown. So the MISE of Eq. (1) cannot be evaluated.

In the remainder of this section, we will explain how the MISE of Eq. (1) can be estimated when Kernel Density Estimation (KDE) is employed. First, KDE will be explained. Next, in Section III-B, a method is presented for estimating the MISE when assuming that the d parameters are correlated. Section III-C shows how the MISE can be approximated when some of the d parameters are independent from each other.

A. Estimating the distribution using Kernel Density Estimation

The shape of the probability densities is unknown beforehand. Furthermore, the shape of the estimated pdf might change as more data are acquired. Assuming a functional form of the pdf and fitting the parameters of the pdf to the data may therefore lead to inaccurate fits unless a lot of hand-tuning is applied. We employ a non-parametric approach using Kernel Density Estimation (KDE) (Parzen 1962; Rosenblatt 1956) because the shape of the pdf is automatically computed and KDE is highly flexible regarding the shape of the pdf.

In KDE, the estimated pdf is given by

$$\hat{f}(x;n) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2}$$

Here, $K(\cdot)$ is an appropriate kernel function and h denotes the bandwidth. The choice of the kernel $K(\cdot)$ is not as important as the choice of the bandwidth h (Turlach 1993). We use a Gaussian kernel because it will simplify some of our calculations. The Gaussian kernel is given by

$$K(u) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}\|u\|^2\right\}, \tag{3}$$

where $\|u\|^2$ denotes the squared 2-norm of u, i.e., $u^T u$.

The bandwidth h controls the amount of smoothing. For the kernel of Eq. (3), the same amount of smoothing is applied in every direction, although our method can easily be extended to a multi-dimensional bandwidth, see, e.g., Chen (2017) and Scott and Sain (2005). There are many different ways of estimating the bandwidth, ranging from simple reference rules like, e.g., Scott's rule of thumb (Scott 2015) or Silverman's rule of thumb (Silverman 1986) to more elaborate methods; see Bashtannyk and Hyndman (2001), Chiu (1996), Jones et al. (1996), and Turlach (1993) for reviews of different bandwidth selection methods.

B. Estimating the Mean Integrated Squared Error for dependent parameters

As an approximation of the MISE of Eq. (1), the asymptotic mean integrated squared error (AMISE) is often used. With the KDE of Eq. (2) employed, the AMISE is defined as follows (Marron and Wand 1992):

$$\mathrm{AMISE}_f(n) = \frac{h^4}{4}\sigma_K^4 \int_{\mathbb{R}^d} \left(\nabla^2 f(x)\right)^2 \mathrm{d}x + \frac{\mu_K}{nh^d}. \tag{4}$$

Here, σ_K and μ_K are constants that depend on the choice of the kernel $K(\cdot)$:

$$\sigma_K = \int_{\mathbb{R}^d} \|u\|^2 K(u)\,\mathrm{d}u, \tag{5}$$

$$\mu_K = \int_{\mathbb{R}^d} K(u)^2\,\mathrm{d}u. \tag{6}$$

Because we use the Gaussian kernel of Eq. (3), we have $\sigma_K = 1$ and $\mu_K = (2\sqrt{\pi})^{-d}$. In Eq. (4), $\nabla^2 f(x)$ denotes the Laplacian of $f(x)$, i.e.,

$$\nabla^2 f(x) = \sum_{l=1}^{d} \frac{\partial^2 f(x)}{\partial x_l^2}. \tag{7}$$

Note that the Laplacian equals the trace of the Hessian. Assuming that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, the AMISE only differs from the MISE by higher-order terms under some mild conditions[1] (Silverman 1986).

The influence of the bandwidth h is demonstrated in an illustrative way by the AMISE of Eq. (4). The first term of the AMISE of Eq. (4) corresponds to the asymptotic bias introduced by smoothing the pdf. Therefore, this term approaches zero when $h \to 0$. However, when $h \to 0$, the variance goes to infinity, as can be seen by the second term of the AMISE, which corresponds to the asymptotic variance.

As with the MISE, we cannot evaluate the AMISE because it depends on the true pdf $f(\cdot)$. As suggested by Calonico et al. (2018) and Chen (2017), we can estimate the quantity

[1]The pdf $f(\cdot)$ needs to comply with the regularity conditions, $K(u) \geq 0, \forall u$, $\int_{\mathbb{R}^d} K(u)\,\mathrm{d}u = 1$ and σ_K from Eq. (5) is not infinite.

$\nabla^2 f(x)$ by $\nabla^2 \hat{f}(x;n)$, with $\hat{f}(x;n)$ defined in Eq. (2). Substituting $f(x)$ in Eq. (4) with $\hat{f}(x;n)$ gives the measure that we will use to quantify the completeness:

$$J_f(n) = \frac{h^4}{4}\sigma_K^4 \int_{\mathbb{R}^d} \left(\nabla^2 \hat{f}(x;n)\right)^2 \mathrm{d}x + \frac{\mu_K}{nh^d}. \tag{8}$$

In summary, the measure Eq. (8) is an estimation of the MISE of Eq. (1) given that the pdf is estimated using the KDE of Eq. (2). Because the MISE cannot be directly evaluated, the asymptotic MISE is used with the estimated pdf substituted for the real pdf.

C. Estimating the Mean Integrated Squared Error for independent parameters

As explained in Section III-A, KDE is employed because the KDE is highly flexible regarding the shape of the pdf. However, when a large number of parameters are used, i.e., for large values of d, the KDE becomes unreliable due to the curse of dimensionality (Scott 2015). One way to overcome this, is to assume that certain parameters are independent. In that case, the joint distribution is not modeled using only one multivariate KDE, but using a combinations of KDEs.

Without loss of generality, consider the parameter vector x that can be decomposed into two parts:

$$x = \begin{bmatrix} y \\ z \end{bmatrix}, \tag{9}$$

such that $y \in \mathbb{R}^{d_y}$ and $z \in \mathbb{R}^{d_z}$ with $d_y + d_z = d$. If the parameter vectors y and z are independent, the probability density of x equals

$$f(x) = g(y)h(z), \tag{10}$$

where $g(\cdot)$ and $h(\cdot)$ are pdfs. Because y and z have a lower dimensionality than x, the estimated pdfs of $g(\cdot)$ and $h(\cdot)$ will be more reliable. However, we cannot use the measure of Eq. (8) to quantify the completeness anymore. Therefore, we will show in this section how $J_f(n)$ can be computed in case the real distribution is assumed to take the form Eq. (10).

The first step is to estimate $g(\cdot)$ and $h(\cdot)$ using $\hat{g}(\cdot;n)$ and $\hat{h}(\cdot;n)$, respectively, where $\hat{g}(\cdot;n)$ and $\hat{h}(\cdot;n)$ are also estimated using KDE, see Eq. (2). Note that the bandwidths of $\hat{g}(\cdot;n)$ and $\hat{h}(\cdot;n)$ are generally different. Now let the MISE of $g(\cdot)$ and $h(\cdot)$ be defined similar as the MISE of $f(\cdot)$ in Eq. (1). It can be shown[2] that if Eq. (10) holds, then the MISE of $f(x)$ approximately equals

$$\begin{aligned} \mathrm{MISE}_f(n) \approx \mathrm{MISE}_g(n) \int_{\mathbb{R}^{d_z}} h(z)^2 \,\mathrm{d}z \\ + \mathrm{MISE}_h(n) \int_{\mathbb{R}^{d_y}} g(y)^2 \,\mathrm{d}y \\ + \mathrm{MISE}_g(n) \cdot \mathrm{MISE}_h(n). \end{aligned} \tag{11}$$

We can estimate the MISE of $g(\cdot)$ and $h(\cdot)$ in a similar manner as we did for the MISE of $f(\cdot)$ in Section III-B, such that we obtain $J_g(n)$ and $J_h(n)$. Since we cannot evaluate the integrals of Eq. (11), we estimate them by substituting the estimated pdfs. As a result, we have

$$\begin{aligned} J_f(n) = J_g(n) \int_{\mathbb{R}^{d_z}} \hat{h}(z;n)^2 \,\mathrm{d}z \\ + J_h(n) \int_{\mathbb{R}^{d_y}} \hat{g}(y;n)^2 \,\mathrm{d}y + J_g(n) \cdot J_h(n). \end{aligned} \tag{12}$$

In this section, we assumed that the parameters x can be split into two partitions that are independent. It is straightforward to extend the result of Eq. (12) in case that the parameters x can be split into three of more partitions.

IV. Examples

In this section, the proposed method of Section III is illustrated by means of two examples. The first example applies the method with data generated from a known distribution. Because the distribution is known, the real MISE can be accurately approximated and compared with the results from Eqs. (8) and (12). Secondly, in Section IV-B, the proposed method is applied on a dataset containing naturalistic driving data.

A. Example with known underlying distribution

In this example, the data samples Y_i with $i \in \{1,\ldots,n\}$ are independently and identically distributed random variables that are distributed according to the pdf $g(\cdot)$. Each data sample Y_i corresponds to a scalar, i.e., $d_y = 1$. Similarly, the data samples Z_i with $i \in \{1,\ldots,n\}$ are independently and identically distributed random variables that are distributed according to the pdf $h(\cdot)$. The data samples are combined, similar to Eq. (9), such that the likelihood of X_i is $f(X_i) = g(Y_i)h(Z_i)$.

Figure 1 shows the distributions $g(\cdot)$ (black solid line) and $h(\cdot)$ (gray dashed line). Both distributions are Gaussian mixtures, i.e., both pdfs equal the sum of multiple weighted Gaussian distributions. The pdf $g(\cdot)$ corresponds to the average of two Gaussian distributions with means of -1 and 1 and standard deviations 0.5 and 0.3, respectively. The pdf $h(\cdot)$ corresponds to the average of three Gaussian distributions with means -0.5, 0.5, and 1.5, and standard deviations 0.3, 0.5, and 0.3, respectively.

The expectation $\mathrm{E}[\cdot]$ of Eq. (1) is estimated by repeating the estimation of the pdf 200 times, such that the real MISE is approximated:

$$\mathrm{MISE}_f(n) \approx \frac{1}{m}\sum_{j=1}^{m} \int \left(f(x) - \hat{f}_j(x;n)\right)^2 \mathrm{d}x, \tag{13}$$

where $\hat{f}_j(x;n)$ is the j-th estimate and $m = 200$.

All three pdfs are estimated using Eq. (2). We use leave-one-out cross validation to compute the bandwidth h (see also Duin (1976)) because this minimizes the Kullback-Leibler divergence between the real pdf $f(\cdot)$ and the estimated pdf $\hat{f}(\cdot;n)$ (Turlach 1993; Zambom and Dias 2013). Note that although the estimation of the pdfs is repeated 200 times to accurately approximate the MISE using Eq. (13), the bandwidth is only determined once for a specific number of

[2] For the sake of brevity, the proof is omitted from this paper. The main idea is based on the variance of the product of two independent variables, see Goodman (1960), and the assumptions $\mathrm{E}[\hat{g}(y;n)] \approx g(y)$ for all y and $\mathrm{E}[\hat{h}(z;n)] \approx h(z)$ for all z.

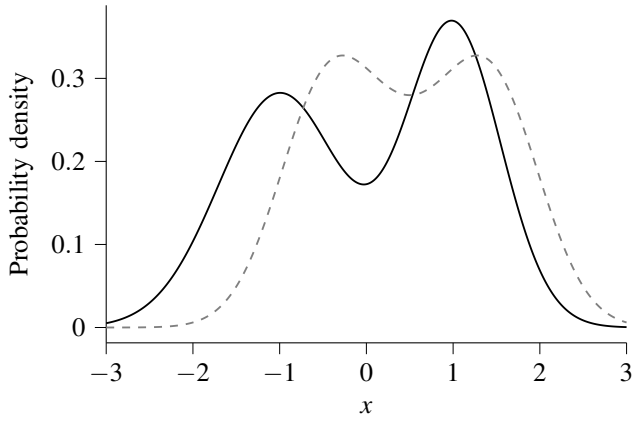

Fig. 1. The true probability density functions $g(\cdot)$ (black solid line) and $h(\cdot)$ (gray dashed line) that are used to illustrate the quantification of the completeness.

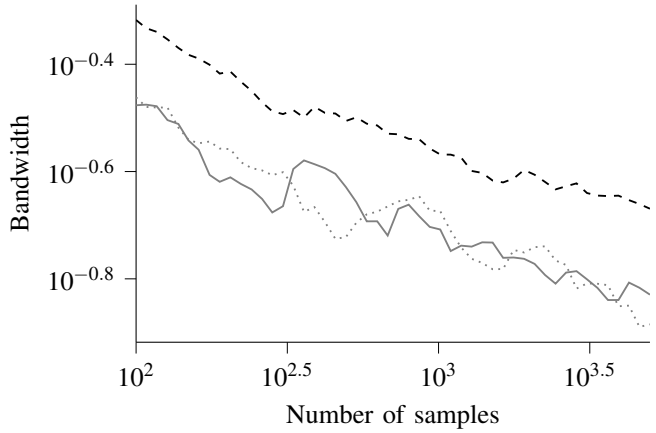

Fig. 2. The bandwidths of $\hat{f}(x;n)$ (black dashed line), $\hat{g}(y;n)$ (gray solid line), and $\hat{h}(z;n)$ (gray dotted line) for the example of Section IV-A. The bandwidths are computed using leave-one-out cross validation for different number of samples n.

samples. All the other 199 times, the same bandwidths are adopted. The resulting bandwidths are shown in Fig. 2. The bandwidth of $\hat{f}(\cdot;n)$ (black dashed line) is significantly larger than the bandwidths of $\hat{g}(\cdot;n)$ (gray solid line) and $\hat{h}(\cdot;n)$ (gray dotted line). This result is not surprising: because $\hat{f}(\cdot;n)$ represents a bivariate distribution, it requires more data to have a similar bandwidth compared with a univariate distribution (Scott and Sain 2005).

Figure 3 shows the results of this example. The black lines show the real MISEs, approximated using Eq. (13), where the black solid line represents the MISE when $f(\cdot)$ is directly estimated and the black dashed line represented the MISE when use is made of Eq. (10). The MISE is significantly lower when it is correctly assumed that the two parameters are independent. One way to look at this is that the degree of freedom of $f(\cdot)$ is reduced when assuming that the two parameters are independent and this lower degree in freedom leads to a more certain estimate. Hence, the MISE is lower.

The gray lines in Fig. 3 show the measures to quantify the completeness of the data. The gray solid line shows the result

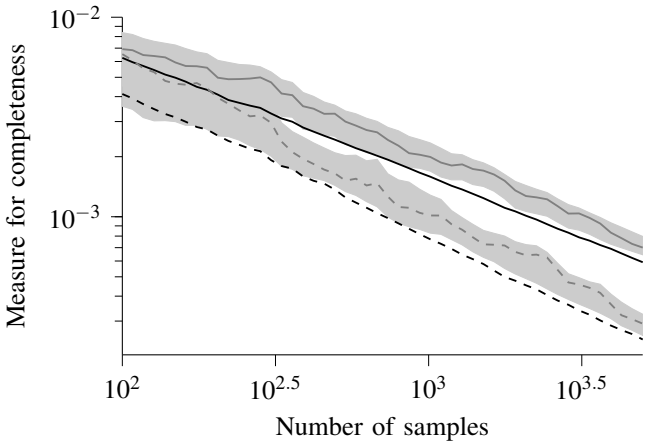

Fig. 3. The real MISEs (black lines) of the example of Section IV-A, approximated using Eq. (13), and the measures that are used to quantify the completeness (gray lines). The solid lines show the result of estimating a bivariate pdf, so here Eq. (8) is used to quantify the completeness. The dashed lines show the result of estimating two univariate pdfs and combining them according to Eq. (10) to create a bivariate pdf, so Eq. (12) is used to quantify the completeness. The gray areas show the interval $[\mu - 3\sigma, \mu + 3\sigma]$, where μ and σ denote the mean and standard deviation, respectively, of the measures of Eqs. (8) and (12) when repeating the experiment 200 times.

of applying Eq. (8) and the gray dashed line shows the result of applying Eq. (12). Both lines follow the same trend as the black solid line and the black dashed line, respectively. This illustrates that the measures Eqs. (8) and (12) are applicable for estimating the real MISE of Eq. (1). To show that this is not a mere coincidence, the gray areas in Fig. 3 show the interval $[\mu - 3\sigma, \mu + 3\sigma]$, where μ and σ denote the mean and standard deviation, respectively, of the measures of Eqs. (8) and (12) when repeating the experiment 200 times. Note that the measures of completeness are consistently higher than the real MISE. This can be explained from the fact that the measures of completeness are approximations of the AMISE and the AMISE itself is always higher than the real MISE under some mild conditions, see Theorem 4.2 of Marron and Wand (1992).

B. Example with real data

In this example, 60 hours of naturalistic driving data from 20 different drivers (see also Gelder and Paardekooper (2017)) is used to extract approximately 2800 braking activities. Three parameters are used to describe each braking activity: the average deceleration, the total speed difference, and the end speed. A histogram of each of these parameters is shown in Fig. 4. Note that these braking activities do not include full stops, i.e., activities where the end speed is zero, because the distribution of the end speed will have a large peak at zero. The AMISE of Eq. (4) deviates more from the real MISE of Eq. (1), especially for larger bandwidths, when such peaks are present in the underlying distribution (Marron and Wand 1992). Because the measure Eq. (8) we use for quantification of completeness is based on the AMISE of Eq. (4), we want to avoid these peaks as much as possible.

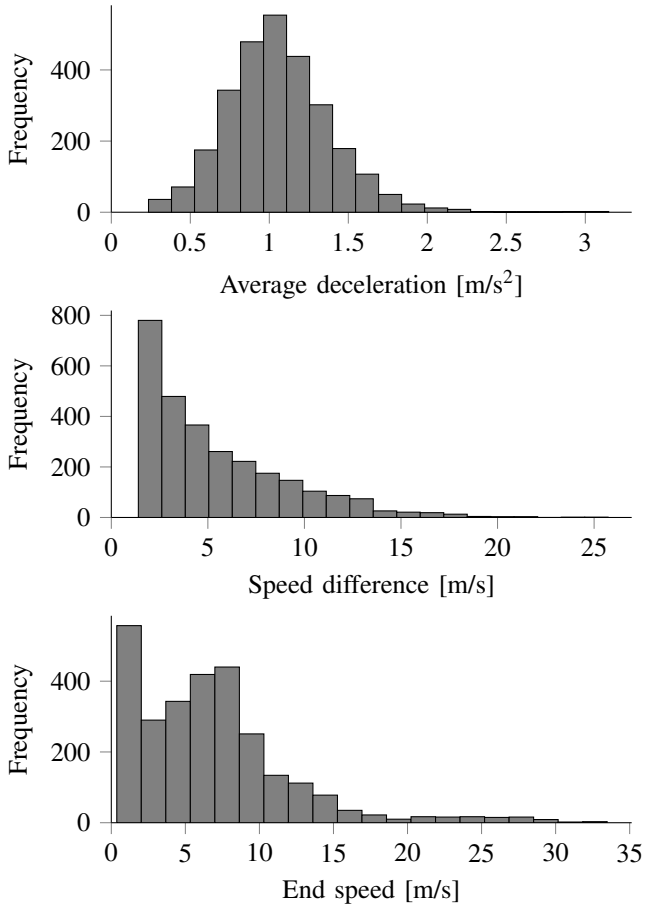

Fig. 4. Histogram of the data that is used for the example with the real data.

Therefore, the full stops are excluded. Note, however, that the method can be applied separately for the full stops. In fact, the analysis for full stops will be simpler, because a full stop activity can be parametrized using only two parameters because the end speed always equals zero.

The three parameters are correlated so this advocates the use of a multivariate KDE. However, as we have seen in the first example, the higher the dimension, the higher the measure for completeness will generally be. So there is a trade-off: Assuming that certain parameters are independent results in an error of the estimated pdf but the resulting MISE, and hence the measure of completeness, will be lower. To illustrate this, we estimate the pdf while assuming all parameters to be dependent and we estimate the pdf while assuming that the average deceleration is independent from the other two parameters. Note that the correlation between the average deceleration and the other parameters is fairly low, so this justifies this choice. The speed difference and end speed are highly correlated, so we will not assume that these two parameters are independent. Before estimating the pdfs, the parameters are translated and rescaled such that each parameter has a sample mean of zero and a sample variance of one. In this example, $\hat{f}(\cdot;n)$ denotes the estimated 3-dimensional pdf using all three parameters, $\hat{g}(\cdot;n)$ denotes

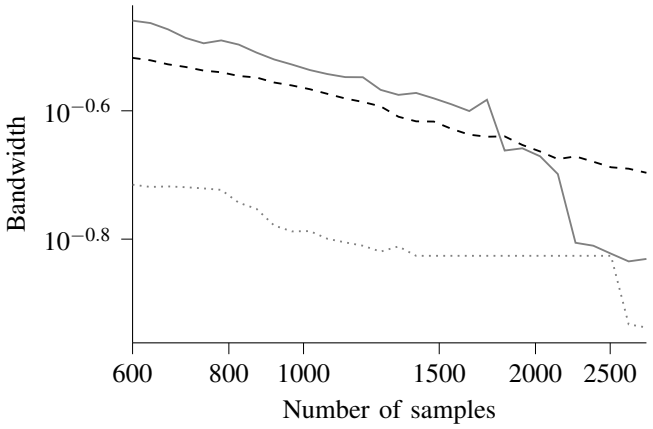

Fig. 5. The bandwidths of $\hat{f}(\cdot;n)$ (black dashed line), $\hat{g}(\cdot;n)$ (gray solid line), and $\hat{h}(\cdot;n)$ (gray dotted line) for the example of Section IV-B. The bandwidths are computed using leave-one-out cross validation for different number of samples n.

the estimated univariate pdf of the average deceleration, and $\hat{h}(\cdot;n)$ denotes the estimated bivariate pdf of the speed difference and the end speed.

Figure 5 shows the bandwidths of the three estimated pdfs for different number of samples, starting from $n = 600$ samples to approximately 2800 samples. As opposed to the bandwidths of our previous example, see Fig. 2, the bandwidth of $\hat{f}(\cdot;n)$ (black dashed line) is not larger than the bandwidth of $\hat{g}(\cdot;n)$ (gray solid line) for low values of n. This is caused by some outliers of the average deceleration, because these outliers have a large influence on the bandwidth of $\hat{g}(\cdot;n)$ (Hall 1992). These outliers also influence the bandwidth of $\hat{f}(\cdot;n)$, but this influence is less because the bandwidth of $\hat{f}(\cdot;n)$ is also influences by the other parameters.

The measures of completeness of the data of the braking activities are shown in Fig. 6. The solid gray line results from the estimated 3-dimensional pdf, i.e., $\hat{f}(\cdot;n)$, where Eq. (8) is used to quantify the completeness. The dashed gray line results from the estimated univariate and bivariate pdfs $\hat{g}(\cdot;n)$ and $\hat{h}(\cdot;n)$, where Eq. (12) is used to quantify the completeness. The measure for the completeness is much lower for the latter case, indicating that the uncertainty of the pdf is much lower when it is assumed that the average deceleration is independent from the other two parameters.

Whether it is better to assume that all parameters are dependent or not depends on the threshold that defines the desired measure and the amount of data. If the threshold is not met, the result can be used to guess how much more data is required by extrapolating the result. To illustrate this, the straight black lines in Fig. 6 represent the least squares logarithmic fits of the corresponding gray lines that can be used for extrapolation. These straight solid and dashed black lines are described by the formulas

$$0.019 \cdot n^{-0.18}, \tag{14}$$

$$0.017 \cdot n^{-0.26}, \tag{15}$$

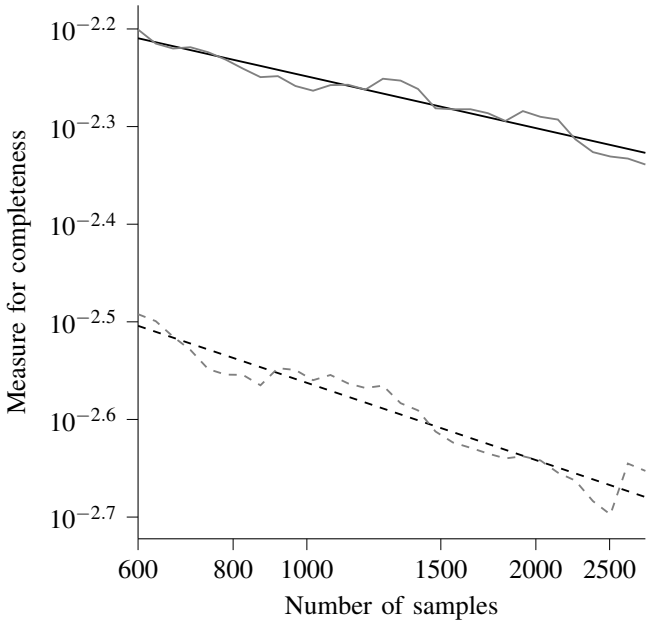

Fig. 6. The measures of completeness for the example of Section IV-B with the assumption that all three parameters depend on each other (gray solid line) and with the assumption that the first parameter, i.e., the average deceleration, does not depend on the other two parameters (gray dashed line). The corresponding black lines represent the least squares logarithmic fits given by Eqs. (14) and (15).

respectively. As an example, let us assume that the threshold equals 0.003. In that case, $n \approx 800$ would suffice if we assume that the average deceleration is independent from the speed difference and end speed, see the dashed lines in Fig. 6 and Eq. (15). This threshold, however, is not yet reached when assuming that all parameters are dependent, see the solid lines in Fig. 6. Extrapolating the result using Eq. (14) provides a rough estimate of the required number of samples: $n \approx 28000$, i.e., ten times as many samples as we used in this example.

V. Discussion

The measure for quantification of completeness of the set of activities that is presented in this work is based on the amount of data and the chosen parametrization. More data might be used to achieve a certain threshold. However, it might also be possible to adapt the parametrization to achieve a certain threshold if a parametrization exists that achieves a certain threshold. Hence, the presented method can be used to determine an appropriate parametrization of activities.

The method for quantifying the completeness of a set of activities presented in this work depends on a threshold that needs to be chosen. Only in case of an infinite set of data, the measure for completeness approaches zero, so this threshold needs to be larger than zero. This threshold might be different for different applications. For example, when the data are used for determining test scenarios (Elrofai et al. 2018; Ploeg et al. 2018), the desired threshold might be lower than when the data are used for determining driver models (Sadigh et al. 2014; Wang et al. 2017). Furthermore, the threshold depends on the number of parameters for one activity, denoted by d in Section III. Based on experience with the dataset used in Section IV-B, assuming that the dataset is normalized such that the standard deviation equals one, a threshold between 0.01 and 0.001 gives good results. When a threshold of 0.01 is reached, a reasonable reliable pdf can be constructed to analyze nominal driving behavior, whereas a threshold of around 0.001 is required to also accurately analyze the edge cases.

When using our measure for completeness, the following might be considered. As explained in Section III, the measure for completeness is based on the AMISE. It is also mentioned that the AMISE only differs from the MISE by higher-order terms under some mild conditions. This requires that the real pdf is smooth, i.e., without large spikes (Marron and Wand 1992). Marron and Wand (1992) also state that the AMISE is strictly higher than the MISE under some mild conditions[3]. As a result, it is likely that the measure for completeness, which is an approximation of the AMISE, is higher than the MISE. This could lead to an overestimation of the number of required samples.

The measure for completeness that is proposed in this paper can be regarded as a approximation of the MISE of Eq. (1). To minimize the MISE, the approximated pdf should be similar to the real pdf. It might be, however, that one is not interested in the exact likelihoods of certain values of the parameters, but in all possible values that the parameters can have. In this case, one might be interested in the support of the real pdf, because the support of the pdf defines all possible values for which the likelihood is larger than zero, see, e.g., Schölkopf et al. (2001).

As mentioned in Section II, our problem of quantifying the completeness of a dataset can be divided into three subproblems. The first subproblem, i.e., how to quantify the completeness regarding the scenario classes, can be regarded as the so-called unseen species problem (Bunge and Fitzpatrick 1993; Gandolfi and Sastri 2004) or species estimation problem (Yang et al. 2012). In case of the unseen species problem, the entire population is partitioned into C classes and the objective is to estimate C given only a part of the entire population. To continue the analogy, the second subproblem, i.e., how to quantify the completeness regarding all scenarios that fall into a specific scenario class, relates to quantifying whether we have a complete view on the variety among one species, given the number of individuals that we have seen. The third subproblem addresses a part of the scenarios, i.e., the activities. In line with the previous analogy, this can be seen as quantifying whether we have a complete view of the parts of the species, e.g., its limbs or organs.

Our proposed method answers the third subproblem, i.e., how to quantify the completeness regarding the activities. The advantage of using the activities for determining the completeness is that there is only a limited number of types of activities. As a result, for each type of activity, it is

[3]The Laplacian of $f(\cdot)$ needs to be continuous and square-integrable and $K(u) \geq 0, \forall u$.

expected that there is no need for an extremely large dataset to obtain a fair number of similar activities. On the other hand, however, it is not known how much data is required to obtain the desired threshold, because, e.g., this depends on the parametrization that is chosen. The next step is to quantify the completeness regarding all scenarios that fall into a specific scenario class. Here, the joint probability of the parameters of different activities in the same scenario class might be considered. Although the presented method can be applied, this might be impractical because the number of parameters will be higher than for the activities. The problems of quantifying the completeness regarding all scenarios that fall into a specific scenario class and quantifying the completeness regarding the scenario classes remain future work.

VI. Conclusions

More and more field data from (naturalistic) driving data become available. The data are used for all kinds of driving-related research, developments, assessments, and evaluations. When deducing claims based on the collected data, we require knowledge about the degree of completeness of the data. We considered the data as a sequence of scenarios, whereas activities are the building blocks of these scenarios. To obtain knowledge about the degree of completeness of the data, we propose a measure to quantify the completeness of the activities. This measure allows to partly answer questions like "have we collected enough field data?" We illustrated the method using an artificial dataset, for which the underlying distributions are known. These results show that the proposed method correctly quantifies the completeness of the activities. We also applied the method on a dataset with naturalistic driving to show that the method can be used to estimate the required number of samples. In future work, we will extend the method to whole traffic scenarios and scenario classes and we will investigate the appropriate thresholds for the measure to quantify completeness in different applications. Furthermore, the proposed method will be used to evaluate the level of completeness of the data collection aimed at defining relevant test cases for the assessment of automated vehicles.

References

Alvarez, S., Y. Page, U. Sander, F. Fahrenkrog, T. Helmer, O. Jung, T. Hermitte, M. Düering, S. Döering, and O. Op den Camp (2017). "Prospective Effectiveness Assessment of ADAS and Active Safety Systems via Virtual Simulation: A Review of the Current Practices". In: *25th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*.

Bashtannyk, D. M. and R. J. Hyndman (2001). "Bandwidth Selection for Kernel Conditional Density Estimation". In: *Computational Statistics & Data Analysis* 36.3, pp. 279–298.

Blair, S. N., M. J. LaMonte, and M. Z. Nichaman (2004). "The Evolution of Physical Activity Recommendations: How Much Is Enough?" In: *The American Journal of Clinical Nutrition* 79.5, 913S–920S.

Broggi, A., M. Buzzoni, S. Debattisti, P. Grisleri, M. C. Laghi, P. Medici, and P. Versari (2013). "Extensive Tests of Autonomous Driving Technologies". In: *IEEE Transactions on Intelligent Transportation Systems* 14.3, pp. 1403–1415.

Bunge, J. and M. Fitzpatrick (1993). "Estimating the Number of Species: A Review". In: *Journal of the American Statistical Association* 88.421, pp. 364–373.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018). "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference". In: *Journal of the American Statistical Association* 113, pp. 767–779.

Chen, Y.-C. (2017). "A Tutorial on Kernel Density Estimation and Recent Advances". In: *Biostatistics & Epidemiology* 1.1, pp. 161–187.

Chiu, S.-T. (1996). "A Comparative Review of Bandwidth Selection for Kernel Density Estimation". In: *Statistica Sinica* 6, pp. 129–145.

Dingus, T. A., F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey (2016). "Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data". In: *Proceedings of the National Academy of Sciences* 113.10, pp. 2636–2641.

Duin, R. P. W. (1976). "On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions". In: *IEEE Transactions on Computers* C-25.11, pp. 1175–1179.

Elrofai, H., J.-P. Paardekooper, E. de Gelder, S. Kalisvaart, and O. Op den Camp (2018). *Scenario-Based Safety Validation of Connected and Automated Driving*. Tech. rep. Netherlands Organization for Applied Scientific Research, TNO.

Elrofai, H., D. Worm, and O. Op den Camp (2016). "Scenario Identification for Validation of Automated Driving Functions". In: *Advanced Microsystems for Automotive Applications 2016*. Springer, pp. 153–163.

Gandolfi, A. and C. C. A. Sastri (2004). "Nonparametric Estimations about Species Not Observed in a Random Sample". In: *Milan Journal of Mathematics* 72, pp. 81–105.

Gelder, E. de and J.-P. Paardekooper (2017). "Assessment of Automated Driving Systems Using Real-Life Scenarios". In: *IEEE Intelligent Vehicles Symposium (IV)*, pp. 589–594.

Geyer, S., M. Baltzer, B. Franz, S. Hakuli, M. Kauer, M. Kienle, S. Meier, T. Weißgerber, K. Bengler, R. Bruder, F. Flemisch, and H. Winner (2014). "Concept and Development of a Unified Ontology for Generating Test and Use-Case Catalogues for Assisted and Automated Vehicle Guidance". In: *IET Intelligent Transport Systems* 8.3, pp. 183–189.

Goodman, L. A. (1960). "On the Exact Variance of Products". In: *Journal of the American Statistical Association* 55.292, pp. 708–713.

Guest, G., A. Bunce, and L. Johnson (2006). "How Many Interviews Are Enough? An Experiment with Data Saturation and Variability". In: *Field Methods* 18.1, pp. 59–82.

Hall, P. (1992). “On Global Properties of Variable Bandwidth Density Estimators”. In: *The Annals of Statistics* 20.2, pp. 762–778.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). “A Brief Survey of Bandwidth Selection for Density Estimation”. In: *Journal of the American Statistical Association* 91.433, pp. 401–407.

Kasper, D., G. Weidl, T. Dang, G. Breuel, A. Tamke, A. Wedel, and W. Rosenstiel (2012). “Object-Oriented Bayesian Networks for Detection of Lane Change Maneuvers”. In: *IEEE Intelligent Transportation Systems Magazine* 4.3, pp. 19–31.

Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey (2006). *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. Tech. rep. DOT HS 810 594. Virginia Tech Transportation Institute.

Krajewski, R., J. Bock, L. Kloeker, and L. Eckstein (2018). “The HighD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems”. In: *IEEE 21st International Conference on Intelligent Transportations Systems (ITSC)*, pp. 2118–2125.

Marks, I. H., Z. V. Fong, S. M. Stapleton, Y.-C. Hung, Y. J. Bababekov, and D. C. Chang (2018). “How Much Data are Good Enough? Using Simulation to Determine the Reliability of Estimating POMR for Resource-Constrained Settings”. In: *World Journal of Surgery* 42.8, pp. 2344–2347.

Marron, J. S. and M. P. Wand (1992). “Exact Mean Integrated Squared Error”. In: *The Annals of Statistics* 20.2, pp. 712–736.

Paardekooper, J.-P., S. Montfort, J. Manders, J. Goos, E. de Gelder, O. Op den Camp, A. Bracquemond, and G. Thiolon (2019). “Automatic Identification of Critical Scenarios in a Public Dataset of 6000 km of Public-Road Driving”. In: *26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*.

Parzen, E. (1962). “On Estimation of a Probability Density Function and Mode”. In: *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076.

Ploeg, J., E. de Gelder, M. Slavík, E. Querner, T. Webster, and N. de Boer (2018). “Scenario-Based Safety Assessment Framework for Automated Vehicles”. In: *16th ITS Asia-Pacific Forum*, pp. 713–726.

Pütz, A., A. Zlocki, J. Bock, and L. Eckstein (2017). “System Validation of Highly Automated Vehicles with a Database of Relevant Traffic Scenarios”. In: *12th ITS European Congress*, pp. 1–8.

Rosenblatt, M. (1956). “Remarks on Some Nonparametric Estimates of a Density Function”. In: *The Annals of Mathematical Statistics* 27.3, pp. 832–837.

Sadigh, D., K. Driggs-Campbell, A. Puggelli, W. Li, V. Shia, R. Bajcsy, A. L. Sangiovanni-Vincentelli, S. S. Sastry, and S. A. Seshia (2014). “Data-Driven Probabilistic Modeling and Verification of Human Driver Behavior”. In: *AAAI Spring Symposium Series*.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson (2001). “Estimating the Support of a High-Dimensional Distribution”. In: *Neural Computation* 13.7, pp. 1443–1471.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Scott, D. W. and S. R. Sain (2005). “Multi-dimensional Density Estimation”. In: *Handbook of Statistics* 24, pp. 229–261.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC press.

Stellet, J. E., M. R. Zofka, J. Schumacher, T. Schamm, F. Niewels, and J. M. Zöllner (2015). “Testing of Advanced Driver Assistance Towards Automated Driving: A Survey and Taxonomy on Existing Approaches and Open Questions”. In: *IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 1455–1462.

Turlach, B. A. (1993). *Bandwidth Selection in Kernel Density Estimation: A Review*. Tech. rep. Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Ulbrich, S., T. Menzel, A. Reschka, F. Schuldt, and M. Maurer (2015). “Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving”. In: *IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 982–988.

Wang, W., C. Liu, and D. Zhao (2017). “How Much Data Are Enough? A Statistical Approach with Case Study on Longitudinal Driving Behavior”. In: *IEEE Transactions on Intelligent Vehicles* 2.2, pp. 85–98.

Williamson, A., D. A. Lombardi, S. Folkard, J. Stutts, T. K. Courtney, and J. L. Connor (2011). “The Link between Fatigue and Safety”. In: *Accident Analysis & Prevention* 43.2, pp. 498–515.

Xie, J., A. R. Hilal, and D. Kulić (2018). “Driving Maneuver Classification: A Comparison of Feature Extraction Methods”. In: *IEEE Sensors Journal* 18.12, pp. 4777–4784.

Yang, H., B. F. Van Dongen, A. H. M. Ter Hofstede, M. T. Wynn, and J. Wang (2012). *Estimating Completeness of Event Logs*. Tech. rep. BPM center.

Zambom, A. Z. and R. Dias (2013). “A Review of Kernel Density Estimation with Applications to Econometrics”. In: *International Econometric Review (IER)* 5.1, pp. 20–42.

Zofka, M. R., F. Kuhnt, R. Kohlhaas, C. Rist, T. Schamm, and J. M. Zöllner (2015). “Data-driven Simulation and Parametrization of Traffic Scenarios for the Development of Advanced Driver Assistance Systems”. In: *18th International Conference on Information Fusion*, pp. 1422–1428.