IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2023, accepted 13 August 2023, date of publication 18 September 2023, date of current version 26 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316354

RESEARCH ARTICLE

Determining the Validity of Simulation Models for the Verification of Automated Driving Systems

BIRTE NEUROHR[1], TJARK KOOPMANN[1], EIKE MÖHLMANN[1], AND MARTIN FRÄNZLE[2]
[1]German Aerospace Center (DLR), 26121 Oldenburg, Germany
[2]Carl von Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany

Corresponding author: Birte Neurohr (birte.neurohr@dlr.de)

This work was supported in part by the Project SET Level 4 to 5 funded by the Federal Ministry of Economic Affairs and Climate Action (BMWK) Based on a Decision of the Deutsche Bundestag.

ABSTRACT As the verification of automated driving systems poses an immense challenge, recent approaches aim for a virtualization of such efforts using computer simulations. This goal, however, motivates a strong need for trustworthy simulation environments and models. As to assess the modeling quality, this work proposes a process to measure the difference between the behaviors of several models. To achieve this, we consider sets of discretized simulation runs to be modeled by time-homogenous Markov chains and under this assumption derive a computable distance measure between sets of simulation traces. If it can be assured that all relevant variables may be observed and no crucial hidden factors are left out, the method can be extended to compare real-world traces with their simulated counterparts.

INDEX TERMS Markov chains, maximum mean discrepancy, simulation, two-sample tests, validity.

I. INTRODUCTION

Automated driving systems (ADSs) are a way to make traveling more comfortable, even more important they have the potential to improve traffic safety by reducing the number and the severity of accidents [1]. A successful introduction of automated driving systems and their societal acceptance requires guarantees regarding their safe operation. Thus, prior to their release to the market one needs to develop them in a safe way and is further required to test them extensively. This is a hard task as both the driving systems themselves and even more the ever-changing environment that these systems have to understand and act in are of very high complexity [2]. A naive statistical verification of these systems would require several hundreds of millions of driven kilometers [3], [4]. To put this in perspective: all streets in the USA only amount to 6.69 million kilometers as of 2016 [5]. Even worse, without further arguments such tests would need to be performed with every newly developed or modified automated driving system.

An approach that is addressing this challenge is to structure the testing space according to scenarios. This so called scenario-based approach [6], where a scenario is a description of the temporal evolution of physical objects as well as environmental conditions [7], is followed by research projects such as PEGASUS[1] VVMethods[2] or SET Level[3] Thus, they allow for different approaches of assuring safe operation of an automated transportation system that do not rely solely on the number of kilometers driven but rather on an identification, understanding, and capturing which principles are essential for the safety of automated transportation systems. Menzel et al. introduced different abstraction levels for scenario description [8]. Such an abstraction allows to group possibly infinite similar *concrete scenarios* [8] into a finite amount of *abstract scenarios* [9]. In turn, such abstract scenarios can be approximated with a finite amount of parameters. This is called a *logical scenario* where parameters could, for example, constrain the actors' velocity or actions. Valid valuations of parameters are thereby

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaogang Jin.

[1]https://www.pegasusprojekt.de/en/home
[2]https://www.vvm-projekt.de/en/
[3]https://setlevel.de/en

defined by finite sets or discrete or continuous distributions, respectively [10].

For the quantification in scenario-based approaches it is required to be able to control the input parameters and observe scenarios with rich variability. This strongly encourages the usage of simulation based methods. However, using synthetically generated quantitative and qualitative evidences in a safety argumentation for the release of these kind of system – as it is currently planned [11] – imposes another challenge: *It has to be certain that the simulation is actually valid, i.e. imitating reality in the necessary aspects with a given quality.* This challenge is called *model validation* and has generated great research interest across several disciplines. Still, in context of automated driving, applicable approaches are needed that allow to support the safety argumentation on results obtained by simulation. Thus, there is an urgent need for the development and analysis of methods to validate the employed simulation models, including the environment model.

One viewpoint of (operational) validity is that the simulation acts as a representative of the real world [12]. Ideally, for ensuring safety of automated driving systems, a simulation would be entirely interchangeable with the real world, i.e. all aspects (e.g. velocity of actors) – in all possible sequences – are equal. While this is an unrealistic (and probably unnecessary) expectation, we assume that there is a finite subset of very important aspects where this should actually be true. These aspects can be obtained from the tests executed during the homologation of automated vehicles. Moreover, if the validation method does not produce a binary outcome – i.e. whether the simulation is indistinguishable from the real world – but instead produces an estimate on how similar these two are, then we can still gain insights from testing, as this would, for example, help to decide which model is suited better for a given task.

Our proposed method takes relevant aspects of logical scenarios and shows that for these aspects the simulation acts as a representative of the real world. Without access to real world probabilities one would need to estimate them, which we prefer to omit. Instead, we would like to be able to do a test on indistinguishability on the basis of samples. Furthermore, one of our goals is to only impose rather weak assumptions, e.g. we do not want to assume the data to follow a normal distribution, about the involved multi-dimensional distributions. Thus, a non-parametric multivariate test is needed. For these reasons, we propose the usage of the *maximum mean discrepancy* (MMD) for a similarity estimation to tackle the validation challenge [13].

A. CONTRIBUTION

The main contributions of this work are:

- Modeling of the logical scenarios as Markov chains. These scenarios are widely used in the application context *virtual assessment of automated driving systems*.
- Definition of a method to apply the maximum mean discrepancy to compare two Markov chains with each other. This extends the approach of simply comparing the distribution of the output.
- Application of the defined method as defined in the aforementioned contributions in the context of automated driving.

In Section II we provide general remarks concerning model validity. We then discuss related concepts and our contributions in Section III and introduce the relevant terminology that we require in this work in Section IV. Once the foundations have been established, we present the proposed method in Section V, including details about the limitations of the presented methodology and possible solutions. In order to demonstrate the feasibility of the method, we evaluated the approach in two experiments in Section VI. Finally, Section VII concludes this paper and outlines future research.

II. SCENARIO-BASED MODEL VALIDATION

In the scenario-based approach, relevant scenarios are elicited by various means, most often distinguished in data-based [14] or expert-based approaches [15]. A combination of data-based and expert-based strategies appears to be the most promising and leads to an identification of relevant (abstract) scenarios that are approximated to logical scenarios by selective parametrization. Fixing these parameters leads to concrete scenarios that can be tested in a simulation. Thus, two research questions may be posed:

1. Is the distribution of abstract scenarios the same as in the real mileage?
2. Are both the virtual model and the real world behaving equally in the logical scenarios?

We address the second question in this work. However, not all aspects, e.g. position of actors, traffic light state, color of surrounding buildings, of these scenarios will need to be close to indistinguishable in a simulation. Thus, when identifying the aspects of a scenario that need to be very close to reality, e.g. the relative positioning of the actors, we need to bear in mind that the proposed method only shows representativeness given these relevant aspects. Hence, the model may differ in other aspects, e.g. whether the shutters of surrounding buildings are opened or closed.

One way to get a first impression of the usefulness of a proposed validation method is the Method of the Manufactured Universe (MMU) introduced by Stripling et al. [16], where validity refers to operational validity in this work, Schlesinger et al. [17]. The idea is to replace reality by a manufactured universe in which all ''true'' values are known and samples can be drawn easily. This means that instead of comparing a reality with a simulation, two simulations are compared where one acts as the reality and one as the simulation to be tested. The MMU gives a good first impression of the usefulness of a proposed validation method, even more so when the manufactured universe is chosen in a realistic way, e.g. including typical measurement errors.

Thus, we evaluate our method in Section VI by comparing two simulations. However, in reality data may only be partially observable. In how far this may affect the proposed method will be discussed in Section V-B.

To measure the distance between real and simulated data, validation metrics are used. According to Ferson et al. [18] desirable properties of such a validation metric are:

- D1 It should be objective. That means, given a collection of observations and predictions, a validation metric produces the same assessment.
- D2 It should generalize deterministic comparisons between scalar values that have no uncertainty (backward compatibility).
- D3 It should reflect all differences in the two distributions, not just the lower moments, e.g. mean or standard deviation, of these distributions.
- D4 For ease of understanding, the unit of the metric should be the same as the unit of the variables, if possible.
- D5 The range of the metric should be unbounded.
- D6 The metric should be a true metric, i.e. positive, symmetric and fulfill the triangle equality.

III. RELATED WORK

Model validation is usually defined to mean ''substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model'' [17]. According to this definition, model validation always takes into account the intended applicability, hence the challenge of model validation becomes harder, the broader this applicability context is. Thus, if the application context is too broad, one needs to resort to testing. If model validity is (only) shown via experiments, then there also needs to be made an argument why the validity should also be true for other parts of the application context. Otherwise, one cannot guarantee the transferability of results from the virtual into the physical world.

The problem of model validity is studied across several disciplines. *Computer Simulation Validation* is a whole book on the state of the art of computer simulation validation, that provides an interdisciplinary perspective on the challenge and includes authors from multiple disciplines [19]. Sargent presents four different approaches for model validation as well as several validation techniques [20]. They reach from the model developers deciding themselves if the model is valid to statistical arguments over confidence intervals obtained from the difference between means, variances and distributions of different simulation model and system output variables. While the first notion is a very subjective view on validity, the latter approach tries to generate some sort of objectivity. This objectified notion of validity poses several benefits as this leads to a better comparability of models.

Allemang et al. [21] present metrics to validate a digital twin of a virtual aircraft model with regards to its trajectory. In their paper, the authors demonstrate how they form the metrics with the help of image decomposition methods and quantify the margin between the simulation and the test data as well as the associated uncertainty.

Another approach to model validation is letting one model attempt to generate realistic samples, and having a discriminator, which attempts to tell these apart from data samples. This is performed by Sutherland et al. to be able to indicate how the model and data distributions differ [22]. The application context here is handwriting and the authors also use the maximum mean discrepancy to achieve their goal. In the context of automated driving, there is only few work on model validation and mainly for the simulation of vehicle dynamics and sensors.

Viehof et al. [23] gives an overview of the state of the art in vehicle dynamics simulation. Further, Viehof et al. [24] is a dissertation that proposes a research methodology for a new validation concept in vehicle dynamics. This method was applied to an idealized radar sensor model by Rosenberger et al. [25]. However, they reported several challenges (e.g. find more appropriate metrics). Rosenberger et al. [26] present a suggestion for a more appropriate metric for the validation of sensor models.

Riedmaier et al. [27] developed a unified framework and survey for model verification, validation and uncertainty quantification. They applied their framework in [28] to a Lane Keeping Functional Test and assessed the validity comparing the minimal lateral distance to the lane markings. This framework was tailored for statistical validation with a focus on uncertainty arising e.g. from the input data in Danquah et al. [29].

The general idea to compare distributions to show model validity is not new. There exist several methods to estimate distances in distributions (e.g. total variation [30] or Kullback-Leibler divergence [31]). This was done for example by Kuefler et al. [32], where speed, acceleration, turn-rate, jerk, and inverse time-to-collision over simulated trajectories were compared with real world trajectories to imitate driver behavior. However, these techniques rely on estimating distributions from the real world. Gretton et al. [13] present a method that does not need to estimate distributions before comparing them. Here, a statistical test is proposed with null hypothesis that the distributions are equal and alternative hypothesis that they differ. This is solely carried out on the basis of samples. As a test statistic the maximum mean discrepancy (MMD), i.e. the difference between mean function values on the two samples, is used.

IV. PRELIMINARIES

As outlined in Section II, we want to compare sets of concrete realizations of given logical scenarios. For this, we will first formally introduce Markov chains as a way to model logical scenarios and afterwards discuss theoretical foundations how to compare said models. Finally, we present how the requirements arising from these theoretical foundations can be checked in our application.

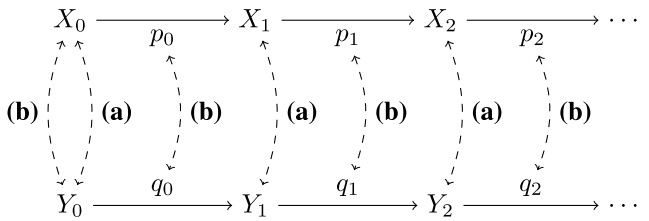


FIGURE 1. Visualization of two Markov chains and two conjectured strategies to compare them (a) by compairing the distribution at each step 0, 1, 2, ... and (b) by compairing the initial distribution and all transition matrices thereafter.

A. MARKOV CHAINS

Behavioral aspects in decision-making components of automated driving systems have been explicitly modeled as Markov decision processes before [33]. Thus, taking into account that the route planning is defined by the logical scenario mentioned in Section I, it would be possible to model logical scenarios as Markov decision processes. However, since we analyze the simulation traces, we assume that the simulation run data can be described using Markov chains, i.e. Markov decision processes with a single action for each state and all rewards being zero. Note that both Markov decision processes as well as Markov chains considered in this work are assumed to have an at most countably infinite state space. Starting from Section V, we will further only consider markov chains with finite state space, as infinite state spaces would lead to unbounded computational effort for the proposed method.

Definition 1 (Markov Chain): A sequence of discrete random variables $(X_t)_{t\in\mathbb{N}_0}$, with identical state space $\mathcal{S}$, is called a **Markov chain** if it fulfills the so called Markov property, i.e.

$$\begin{aligned} P(X_{t+1} = x_{t+1} \mid X_t = x_t, \dots, X_0 = x_0) \\ = P(X_{t+1} = x_{t+1} \mid X_t = x_t) \end{aligned}$$

when $P(X_0 = x_0, \dots, X_t = x_t) > 0$ holds. Further, we call Markov chains **time-homogenous** if

$$P(X_{t+1} = x | X_t = y) = P(X_t = x | X_{t-1} = y)$$

holds for all $x, y \in \mathcal{S}$ and $t \in \mathbb{N}$.
Indexing the state space $\mathcal{S} = \{s_0, s_1, s_2, \dots\}$, we denote the so-called transition probabilities as

$$p_t^{ji} = P(X_{t+1} = s_i \mid X_t = s_j) \text{ for all } s_i, s_j \tag{1}$$

for the transition from X_t to X_{t+1}. We further define $p_t = (p_t^{ji})_{ji}$ as the transition matrix from state X_t to X_{t+1}.

For our analysis we need to specify which requirements we need to impose for a Markov chain to be uniquely defined. Since each step of a Markov chain is represented by a random variable, one might naively assume that ensuring equal distributions of states for each step, i.e. $X_j \overset{d}{=} Y_j$ for all $j \in \mathbb{N}_0$, should be a sufficient condition to uniquely define a Markov chain. However, assuming we have that

i) the state space contains 3 elements s_0, s_1 and s_2,
ii) X_i is uniformly distributed for all i, i.e. $P(X_i = s_j) = \frac{1}{3}$ for all $j = 0, 1, 2$, and
iii) the Markov chain is time-homogenous,

we have the following equations for the computation of the distribution of X_{t+1} from X_t:

$$P(X_{t+1} = s_j) = \sum_{k=0}^{2} p^{jk} P(X_t = s_k) \text{ for } j = 0, 1, 2\,. \tag{2}$$

Assumption ii) then implies

$$\frac{1}{3} = P(X_{t+1} = s_j) = \sum_{k=0}^{2} p^{kj} P(X_t = s_k) = \frac{1}{3} \sum_{k=0}^{2} p^{kj}$$

for all $j = 0, 1, 2$ implying that for the equations to be fulfilled, the transition matrix has to be left stochastic, i.e. columns summing to 1, and thus combined with the general requirement of transition matrices being right stochastic, every doubly stochastic matrix suffices this equation. To give an example, possible candidates to fulfill this requirement are the identity matrix as well as the matrices

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix}, \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \text{ and } \begin{pmatrix} \frac{1}{8} & \frac{7}{8} & 0 \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix},$$

indicating that the Markov chain is not uniquely-defined based on knowledge about the distribution in each time step alone. However, based on Definition 1 we are able to derive that two time-homogenous Markov chains $(X_j)_{j\in\mathbb{N}_0}$ and $(Y_j)_{j\in\mathbb{N}_0}$ with transition matrices p and q, respectively, are equal iff the following two requirements are fulfilled:

R1 The start distributions are equal, i.e. $X_0 \overset{d}{=} Y_0$.
R2 The transition matrices p and q are equal.

B. COMPARISON OF DISTRIBUTIONS

From the requirements above, we can observe that when comparing Markov chains, we are essentially comparing several distributions. This is obvious for **R1** but is also true for **R2** as the rows of the transition matrices can be considered as (multivariate) distributions, where the multivariate part depends on the discretization to obtain a finite number of states and the data itself.

In theory, two distributions are equal iff all their moments, e.g. mean and variance, are the same. As to assess this in practice, several tests for the comparison of distributions have previously been discussed in the literature, with the likely most commonly known being the Kolmogorov-Smirnov test and the Wald-Wolfowitz runs test [34]. However, we propose to use the maximum mean discrepancy (MMD) which is a statistical kernel-based test proposed in Gretton et al. [13]. It has the advantage of being a non-parametric test and thus we do not need to estimate the distributions before comparing them. Further, it is also easy to implement and well-suited for

TABLE 1. Consideration whether the MMD fulfills the desirable properties as listed in Section II.

Property		Comment
D1	✓	It is objective.
D2	✓	The MMD generalizes from deterministic behavior.
D3	✓	The MMD compares all moments of the distributions if the chosen kernels are characteristic (see Subsection IV-C).
D4	✗	The MMD does not consider units, thus this property is not fulfilled.
D5	✓	The MMD is unbounded.
D6	✓	If the chosen kernels are characteristic (e.g. Gaussian or Laplace on $\mathbb{R}^n$) the associated MMD is a metric on the space of probability distributions for this domain according to [13].

multivariate problems. While the Kolmogorov-Smirnov test is also non-parametric, extensions to higher dimensions than the univeriate case are non-trivial [35]. To further motivate the usage of MMD as a validation metric Table 1 considers the desirable properties from Section II. However, even though it might appear in some examples in the literature as if the maximum mean discrepancy does not suffer from the *curse of dimensionality*, i.e. worsened test power with increasing dimensionality of the problem space, it has been shown that the maximum mean discrepancy suffers from this issue as well [36].

Definition 2 (Maximum Mean Discrepancy (MMD), Based on [13, Definition 2]): Let $(\mathcal{X}, d)$ be a metric space, P and Q be Borel probability measures on $\mathcal{X}$, $\mathcal{F}$ a class of functions $f : \mathcal{X} \to \mathbb{R}$, $x \sim P$ and $y \sim Q$. The **maximum mean discrepancy** is then defined as

$$\mathrm{MMD}(\mathcal{F}, P, Q) := \sup_{f \in \mathcal{F}} (\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]) .$$

Definition 2 requires a class of functions $\mathcal{F}$, that based on the provided formula has a significant impact on the ability of the MMD to distinguish between probability measures. A common choice for $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space, where the latter is defined as follows:

Definition 3 (Reproducing Kernel Hilbert Space [37]): Let N be a set. A subset $\mathcal{H} \subset \{f : N \to \mathbb{R}\}$ together with the usual addition and scalar multiplication is called **reproducing kernel Hilbert space (RKHS)** if it satisfies the following properties:

(i) $\mathcal{H}$ is a vector subspace of $(\{f : N \to \mathbb{R}\}, +, \cdot)$,
(ii) $\mathcal{H}$ is endowed with an inner product $\langle \cdot, \cdot \rangle$, with respect to which $\mathcal{H}$ is a Hilbert space, and
(iii) for every $x \in N$, the linear evaluation functional $E_N : \mathcal{H} \to \mathbb{R}$ defined by $E_x(f) = f(x)$ is bounded.

This definition provides a key ingredient for the method presented in this work, namely the function class $\mathcal{F}$. However, another main property, that will be especially relevant for the method at hand, of these Hilbert spaces is not yet visible in the definition and only reflected in its name. For this, we introduce the following definition.

Definition 4 (Reproducing Kernel [37]): Let $\mathcal{H}$ be a RKHS on a set N. Then by Riesz representation theorem, for each $x \in N$, there exists $k_x \in H$, such that for every $f \in \mathcal{H}$ we have $f(x) = E_x(f) = \langle f, k_x \rangle$. The function $K(\cdot, y) = k_y(\cdot)$ is then called the **kernel function**, in short **kernel**, for $\mathcal{H}$.

Note that RKHS are uniquely defined based on a given kernel function K (Moore-Aronszajin theorem) and thus we will from now on mostly discuss kernels K instead of classes of functions $\mathcal{F}$, but be aware that these concepts are strongly connected [38]. Based on this and the definition of the MMD, Gretton et al. [13] then derive the biased estimator for the squared MMD for two datasets X and Y, sampled from P and Q respectively, as

$$\begin{aligned}\mathrm{MMD_b}^2(\mathcal{F}, X, Y) = &\frac{1}{m^2} \sum_{i,j=1}^{m} K(x_i, x_j) \\ &- \frac{2}{mn} \sum_{i,j=1}^{m,n} K(x_i, y_j) \\ &+ \frac{1}{n^2} \sum_{i,j=1}^{n} K(y_i, y_j)\end{aligned}$$

with a kernel function K [13], $\mathcal{F}$ defined as the unit ball of the Hilbert space associated with K, $x_i, x_j \in X$, $y_i, y_j \in Y$, $m = |X|$ and $n = |Y|$.

In order to define a hypothesis test regarding the similarity of distributions, we consider the following commonly used hypothesis

$$\begin{aligned}\mathcal{H}_0 &: P = Q \\ \mathcal{H}_1 &: P \neq Q\end{aligned}$$

together with a selected significance level α. As Gretton et al. demonstrate [13, Theorem 7], it is possible to show a convergence bound based on $K = \sup_{x,y} K(x, y)$ and the sizes m, n of the employed datasets X and Y, respectively, that directly leads to a theorem about the acceptance region in the case $P = Q$ and $m = n$, namely

$$\mathrm{MMD_b}(\mathcal{F}, X, Y) < \sqrt{\frac{2K}{m}} \left(1 + \sqrt{2 \log(\alpha^{-1})}\right) .$$

However, since our use case is usually concerned with the case $m \neq n$, a slight generalization of the steps presented in the proof of the theorem [13, Theorem 7] was necessary, leading to the following threshold for the acceptance region of the hypothesis test:

$$\begin{aligned}\mathrm{MMD_b}(\mathcal{F}, X, Y) < 2 &\left(\sqrt{\frac{K}{m}} + \sqrt{\frac{K}{n}}\right) \\ &+ \sqrt{\frac{2K(m+n)\log(\alpha^{-1})}{mn}} .\end{aligned}$$

Thus, in the case of obtaining the result that $\mathrm{MMD_b}$ is larger than this threshold we reject the null-hypothesis $\mathcal{H}_0$ and otherwise accept it.

C. KERNELS FOR THE MAXIMUM MEAN DISCREPANCY

Kernels are a very useful tool as they allow to operate in a high-dimensional, implicit feature space by computing the

inner products between the images of all pairs of data in the feature space [39]. Particularly relevant for the succesful application of the MMD are so-called characteristic kernels.

Definition 5 (Characteristic Kernel, [40]): Let $\mathcal{X}$ be a topological space and denote with $B_{\mathcal{X}}$ the set of Borel probability measures on $\mathcal{X}$. A measurable and bounded kernel K is said to be **characteristic** if

$$\iota : B_{\mathcal{X}} \to \mathcal{H}, R \mapsto \int_{\mathcal{X}} K(\cdot, x) dR(x)$$

is an injective embedding of $B_{\mathcal{X}}$ into $\mathcal{H}$.

In practice, this property implies that $\mathrm{MMD}(\mathcal{F}, P, Q) = 0$ iff $P = Q$ and thus one can prove that the MMD becomes a metric [13, Theorem 5] when using a characteristic kernel.

As the property of a kernel being characteristic is rather difficult to prove, this work will be limited to two kernels that were previously proven to be characteristic that both employ a heuristic for the respective kernel parameter as well as the non-characteristic linear kernel. The linear kernel shall thereby serve as an example of a kernel failing to differentiate between distributions that differ only in higher moments, e.g. variance.

1) GAUSSIAN KERNEL

The first kernel we will consider is the translation-invariant, strictly positive-definite, continuous and bounded Gaussian kernel, that most importantly has been shown to be characteristic [40], [41]. We employ the following common definition of the Gaussian kernel

$$K(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right) \tag{3}$$

where we choose the kernel parameter σ heuristically as the median of $\{\|x - y\|_2 | x \in X, y \in Y\}$.

2) LAPLACE KERNEL

Another widely-used kernel in machine learning is the Laplace kernel, that shares many of the properties of the Gaussian kernel, especially that it is translation-invariant, positive-definite and characteristic [40]. It is commonly defined as

$$K(x, y) = \exp\left(-\frac{\|x - y\|_1}{\sigma}\right) \tag{4}$$

where we choose the kernel parameter σ heuristically as the median of $\{\|x - y\|_1 | x \in X, y \in Y\}$.

3) LINEAR KERNEL

An example of a kernel that is not characteristic is the homogenous polynomial kernel of degree one, where the range is scaled to [0, 1] by dividing by the maximum value of the employed discretization, which can assumed to be finite due to the finite sample space. The linear kernel can then be written as

$$K(x, y) = \frac{\|x - y\|_1}{\max\{\|x - y\|_1 \mid x \in X, y \in Y\}}. \tag{5}$$

We have now introduced the kernels that will be applied in the remainder of this work. When considering the simulation of automated driving systems, we are receiving time series as an output, in which large correlations between data points inside each time series are possible and stochastic independence can thus not be assumed. For this reason, we introduce a method based on Markov chains and the MMD for our application domain in Section V.

V. MMD-BASED COMPARISON PROCESS

In the following, we describe our proposed method for the comparison of Markov chains using the MMD. Corresponding to the definitions in Section IV, we make the following assumptions:

- **A1** The time series contained in the datasets can be interpreted as the realization of a Markov chain.
- **A2** The Markov chains are time-homogenous.

First, we describe the process under the assumption of full observability of relevant features, e.g. if both datasets stem from simulation models, and then outline the differences if partial observability of one of the datasets must be assumed.

A. FULLY OBSERVABLE FEATURES

The general process for the comparison of Markov chains is depicted in Figure 2. Starting with two datasets and the features they contain, one must select a subset of relevant features for the comparison. While this is required due to the aforementioned curse of dimensionality, cf. Subsection IV-B, that is a yet unsolved issue when comparing (multi-dimensional) distributions, it is also reasonable to assume that not all observable data is relevant for an automated driving system. Once the set of relevant features has been identified, the samples contained in both datasets are projected with respect to said relevant features and subsequently discretized. Note that choosing a well-suited discretization method is both highly relevant and very difficult, as it determines the definition of the states in both Markov chains. Thus, it inherently defines a notion of distance between states in the state space of both Markov chains and which real world states may be identified with each other. As the employed biased MMD estimator is of quadratic complexity and reachability of the states is a direct concern, we decided to apply a simple equal-frequency binning strategy, cf. Figure 3, that also induces a notion of distance on the state space. However, since this is a highly relevant step in the process of this comparison, further research in this direction is needed and shall be done in future work.

Following the discretization, if real world data shall be used in one of the datasets one shall continue with path **b** further described in the next subsection, otherwise path **a** shall be used directly. Then, the start distributions are compared between the discretized datasets and if they are inequal, the process stops as the Markov chains may not be equal in the sense expressed in Subsection IV-A. However, if the comparison indicates that the start distributions are in fact equal, and thus requirement **R1** is fulfilled, we may

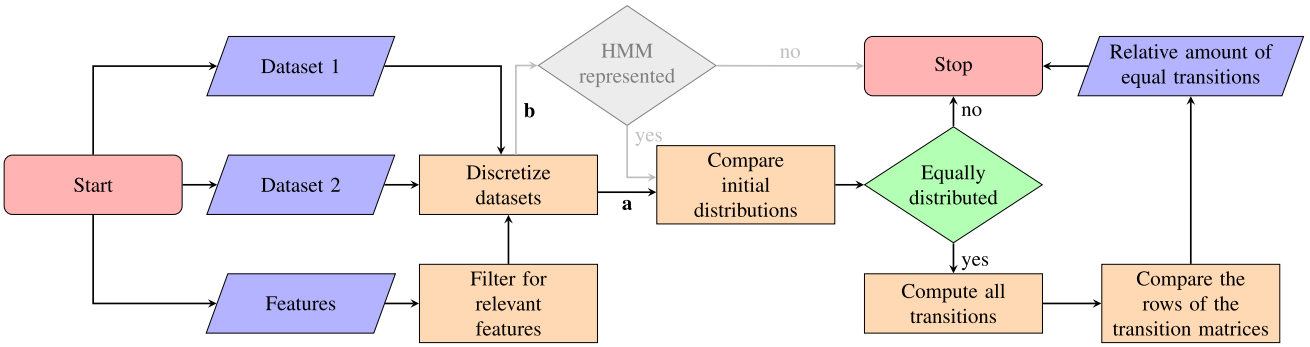


FIGURE 2. Flow chart depicting the workflow of the Markov chain comparison method. The implemented functionality comparing fully-observed simulation data employs the arrow marked with a while a comparison with reality, modeled as hidden Markov model (HMM) and described in Subsection V-B, would require the intermediate step reached by b.

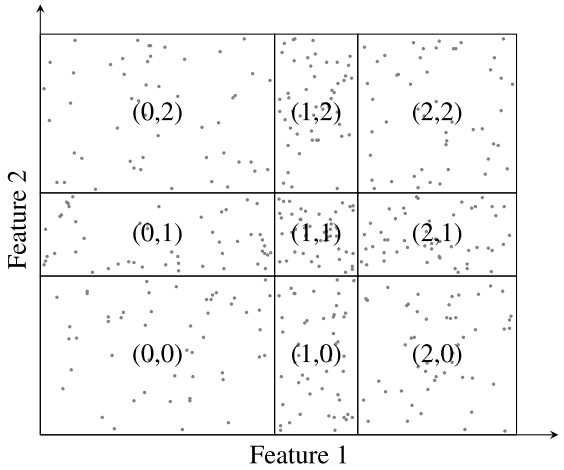


FIGURE 3. Visualization of the discretization strategy employed in the experiments for two features with vectorial state indices, where each bin contains an equal amount of data points. When considering higher dimensions, the strategy is analogously extended.

move forward to the comparison of transition matrices. As mentioned in Section IV, the rows of the transition matrices can be understood as a probability distribution, describing the probability of a transition to state s_j when currently in state s_i. Due to this idea, we may again employ the MMD-based hypothesis test derived in Subsection IV-B to compare the distribution of end states conditioned on a start state between the two models, iterating over the possible start states of the transitions. Further, we aggregate the results of the hypothesis tests to obtain the relative amount of distinguishable transitions in the Markov chains as the end result of the comparison, i.e.

$$R = \frac{|\{s \in \mathcal{S} \mid \mathcal{H}_0 \textit{ rejected}\}|}{|\mathcal{S}|} \tag{6}$$

where $\mathcal{S}$ refers to the common state space of the markov chains and $\mathcal{H}_0$ *rejected* denotes that $\mathcal{H}_0$ has been rejected for the comparison of end state distributions between datasets when conditioned on the start state $s \in \mathcal{S}$.

As Section II introduced several desirable properties of validation metrics, it was checked whether the defined method fulfills them. As shown in Table 2, the proposed method to compare Markov chains inherits the properties D1, D2, and D3 from the MMD. Further, due to the MMD, the result has no unit. Moreover, it is not possible to unambiguously assign a unit to the results, as the input may be a vector. Even though the MMD is unbounded, as the proposed method's output is a relative amount of equal transitions, cf. Equation 6, its upper bound is 100%. Since a transformation by $\sigma(x) = \frac{1}{1-x}$ would reward us with an unbounded result, the relevance of this requirement appears questionable to the authors. Lastly, regarding D6 the proposed method's result is likely positive definite and symmetric (up to estimation errors) as inherited from the baseline MMD. However, as can be seen in the example Table 3 the triangle inequality is probably not fulfilled – it would technically be possible that it only appears this way due to the employment of estimators in the example. Altogether, the proposed method has the most relevant properties for our use case out of the ones listed in Section II, yet there still remains room of improvement in particular concerning analytical results regarding D6.

TABLE 2. Assessment whether the method presented in Section V fulfils the desirable properties as listed in Section II.

Property		Assessment
D1	✓	It is objective.
D2	✓	The comparison does generalize from deterministic behavior, yet it requires a large enough dataset of time series to work as a baseline requirement and thus cannot argue about a single trajectory alone.
D3	✓	Due to the properties of the Gaussian and Laplace kernel, namely being characteristic, the MMD compares all moments of the distributions and the usage of Markov chains as an underlying model enables us to consider time as well.
D4	✗	As the method is based on the MMD, the result does not contain a unit.
D5	✗	The method produces a bounded result and thus does not fulfill this requirement.
D6	?	The method may not fulfill the triangle inequality.

B. COMPARING REAL WORLD WITH SIMULATION DATA

So far we have discussed the comparison of models whose simulation traces can be understood to represent time-homogenous Markov chains. This, however, implies the assumption of perfect observability and does not take into

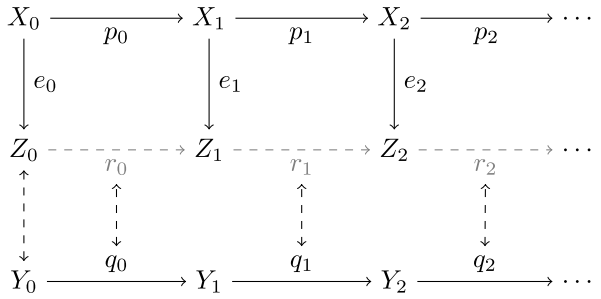


FIGURE 4. Visualization of the comparison of a hidden Markov model comprised of the underlying process $(X_t)_t$ and the observations $(Z_t)_t$ with a Markov chain $(Y_t)_t$. Greyed out transitions $(r_t)_t$ are inferred and it must be shown that they represent the actual transitions $(p_t)_t$ before comparing $(Z_t)_t$ and $(Y_t)_t$.

account that data from reality is generally incomplete as certain influencing factors may be missing. Incorporating *partial observability*, one may define *partially observable Markov chains*, also known as hidden Markov models (HMMs). The main concept behind partially observed Markov chains is to express the measurement process directly by considering the Markov chain as unobservable and to introduce new observable variables $\{Z_t\}_t$ for which Z_t only depends on X_t for all $t \geq 0$ and where the dependency of Z_t on X_t is encoded in an emission matrix e_t. As for the Markov chains considered in this work, it may be a valid assumption that the emissions are time-independent, i.e. $e_t = e_s$ for all $t \neq s$.

In the setting of Figure 4, $(X_t)_t$ represents the underlying idealization of the real world scenario and Z_t represents a partial observation of X_t for each time step t. Based on this understanding, the difference between the $(Z_t)_t$ and the underlying process $(X_t)_t$ is determined by the emissions $(e_t)_t$. Thus, it is necessary to formulate requirements on the emissions $(e_t)_t$ such that we are able to ensure the Markov property for $(Z_t)_t$ as required by the method proposed in this work. Further, it is not yet clear how the results derived for $(Z_t)_t$ can be transferred to the underlying process $(X_t)_t$.

One possible candidate to estimate the underlying hidden Markov model, and thus the emission matrices, from the observations is the Baum-Welch algorithm [42], that appears as a reasonable starting point for future considerations.

In the following, we restrict ourselves to the demonstration of the comparison between simulations and thus ignore the aforementioned issues when considering real world data. However, the issue of partial observability of real world data is crucial to the topic at hand and other frequentistic comparison approaches and shall be dealt with in future work.

VI. EXPERIMENTS

The process as described in Section V as well as the estimator and kernels as presented in Section IV have been implemented in `python` and applied in two use cases. First, we compare simple, easily scalable OpenModelica [43] models. This shall determine whether the method is able to detect quantifiable differences between the logical scenarios that result from the models execution with varied model input functions. Afterwards, we turn towards a comparison of more complicated CARLA [44] simulations without directly controlled differences, but deviations emerging from modifications in the scenario setup. In both cases the underlying data is supplied as sets of trajectories, in this case CSV files, which contain for each equidistant time step the relevant features for the comparison.

A. SIMPLECAR MODEL IN OPENMODELICA

For this experiment, the simulation models are created in OpenModelica using the SimpleCar model [45], sometimes also referred to as one track model, as a baseline.

This dynamic motion model can be written as

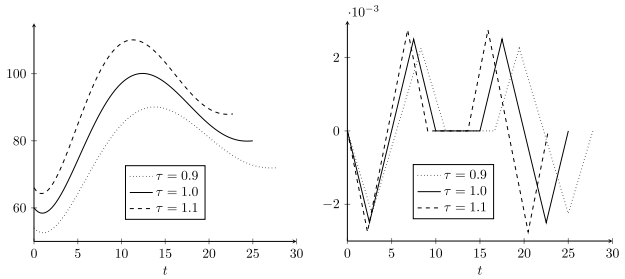
$$\begin{cases} \dot{x} = u_s \cdot \cos(\psi) \\ \dot{y} = u_s \cdot \sin(\psi) \\ \dot{\psi} = \dfrac{u_s}{L} \cdot \tan(u_\varphi) \end{cases} \tag{7}$$

where $u_s : \mathbb{R} \to \mathbb{R}$ and $u_\varphi : \mathbb{R} \to \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ are input functions describing the speed and steering angle of the SimpleCar and L denotes the wheelbase. For easier specification, we introduced a new variable u_ω representing the steering speed input and introduce another equation, $\dot{\varphi} = u_\omega$, to the model, thus resulting in

$$\begin{cases} \dot{x} = u_s \cdot \cos(\psi) \\ \dot{y} = u_s \cdot \sin(\psi) \\ \dot{\psi} = \dfrac{u_s}{L} \cdot \tan(\varphi) \\ \dot{\varphi} = u_\omega. \end{cases} \tag{8}$$

Notice that in the third equation u_φ was also replaced with the newly defined variable φ. The idea is to model a simple trajectory that will be traversed at different speeds, while keeping the path itself the same up to a shift by the initial values. Additionally, the starting coordinates x_0 and y_0 will be chosen as normally distributed around the origin $(0, 0)$ with a standard deviation of 10 in both dimensions, i.e. $x_0, y_0 \sim \mathcal{N}(0, 10)$, as well as $\psi_0 = 0$ and $\varphi_0 = 0$. In order to ensure that the differences in the trajectories remain controllable, a time scaling parameter τ is introduced in the input functions u_s and u_ω, see Figure 5. Further, to obtain multiple logical scenarios, a hyperparameter λ is introduced and for each λ, τ is sampled uniformly from the interval $[\lambda - 0.01, \lambda]$, thus obtaining a logical scenario for each $\lambda \in \{0.81, \ldots 1.2\}$.

Since our proposed comparison method uses hypothesis tests, we have to choose an α-value, which we will set to $\alpha = 0.01$. As detailed in Section V we are comparing the models under the assumption that their trajectories can be interpreted as Markov chains, thus we can use MMD to compare the end point distribution of the transition for each discretized state. Since the optimal choice of a kernel still remains an open question in research, we execute the comparison using the kernels as listed in Subsection IV-C. As relevant features for this comparison we selected the x and y coordinate of the SimpleCar. Results regarding this



(a) Speed $u_{s,\tau}$ (b) Steering speed $u_{\omega,\tau}$

FIGURE 5. Graphs of the input functions for the steering speed and velocity of the simple car as used in Subsection VI-A.

comparison of a separately generated dataset ($\lambda = 1.0$) with the aforementioned datasets ($\lambda = 0.81, \ldots, 1.2$) can be seen in Figure 6, in particular that both characteristic kernels provide almost the same outcome and a quite clear ranking of the models that fits the non-similarity introduced to the datasets by the different λ values. Since the linear kernel completely failed to distinguish between the models and Gaussian and Laplace kernel performed almost equally well, we only consider one of the kernels for further experiments, namely the Gaussian kernel. Additionally, let us mention that the output of the MMD based comparison is, with a few outliers, monotonic in relation to the induced difference, by assigning different values of λ, between the datasets.

The same experiment was conducted taking the heading angle of the ego vehicle into account. However, as depicted in Figure 7 the differentiation between similar and non-similar logical scenarios is worse in this setting. There are multiple possible explanations for this. First, it is possible that this is partially due to the curse of dimensionality mentioned in Subsection IV-B, that reduces the power of the individual hypothesis tests. Second, it is also possible that this reduced differentiation stems from issues in the discretization when dealing with features on different orders of magnitude. Third, it may be that due to the dependencies in Equation 8, the data generation induces a history of length two, which does not fit the current modeling as a markov chain with history of length one. Regardless of the specific underlying reasons for this loss of accuracy, this example demonstrates the need for a selection method of features that should be considered in the comparison.

B. EXPERIMENTS USING A CARLA MODEL

In the last subsection we have demonstrated that the presented approach is able to differentiate between sets of simulation runs when all the inputs of the considered model are directly controlled for and the simulation is completely deterministic based on those inputs. However, it shall now be examined whether comparisons of data generated with different methods to introduce variation into the scenario between datasets confirms our previous results. Thus, in this subsection we use a scenario in which two actors are considered and an external influencing factor is given by a

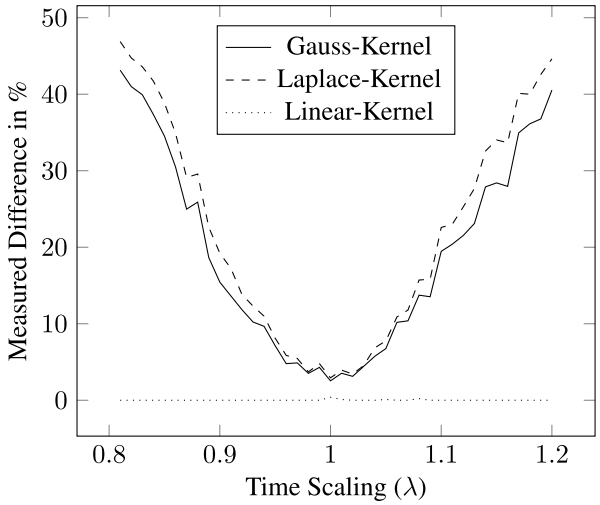


FIGURE 6. Performance of the kernels when considering datasets generated using the SimpleCar model varying the hyperparameter $\lambda \in \{0.81, 0.82, \ldots, 1.2\}$, using *x* and *y* coordinate of the vehicle as relevant features.

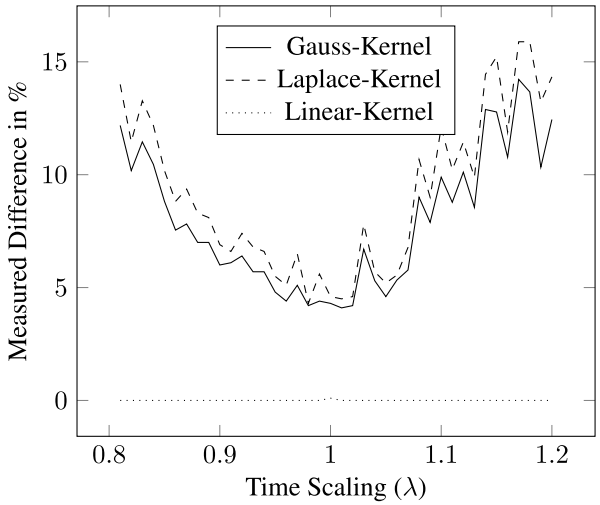


FIGURE 7. Performance of the kernels when considering datasets generated using the SimpleCar model with various hyperparameters $\lambda \in \{0.81, 0.82, \ldots, 1.2\}$, using *x* and *y* coordinate as well as the heading angle of the vehicle as relevant features.

static occlusion that leads to different behavior of the actors. Specifically, the road network of the scenario is a T-junction, cf. Figure 8, where an ego vehicle drives straight, from west to east, through the junction and a bicyclist coming from the southern side arm takes a turn into the western arm, i.e. the direction the ego vehicle came from. The static occlusion is represented by one or more vehicles on the right side of the western arm of the junction, blocking the ego's vision of the side road.

For the comparison, three scenario categories η_i were defined, namely η_1 without vehicles on the side of the road and thus no static occlusion present in the scenario, η_2 with vehicles at the side of the road and thus a high chance of a static occlusion occuring and finally η_3 with a 50% chance of vehicles at the side of the road being present in the scenario. For each of the scenario categories, three datasets

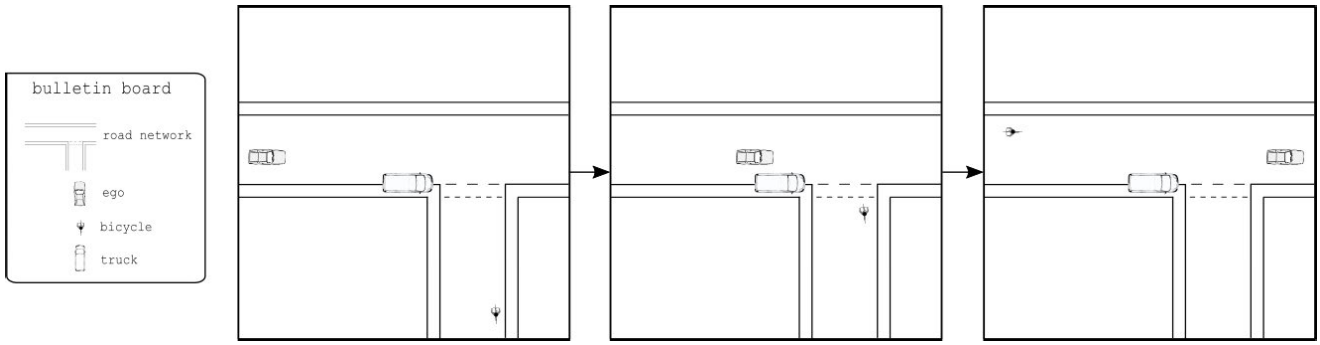


FIGURE 8. Traffic sequence chart (TSC) [46] depicting the scenario executed in CARLA for application in Subsection VI-B. The scenario has previously been considered in [9].

TABLE 3. Comparison of datasets with occlusion, without occlusion and with the chance of an occlusion. The assigned value is the distance of the models as determined by Equation 6 based on the method detailed in Section V.

	No occlusion 1 (η_1^1)	No occlusion 2 (η_1^2)	No occlusion 3 (η_1^3)	Occlusion 1 (η_2^1)	Occlusion 2 (η_2^2)	Occlusion 3 (η_2^3)	Random occlusion 1 (η_3^1)	Random occlusion 2 (η_3^2)	Random occlusion 2 (η_3^3)
No Occlusion 1 (η_1^1)	0.0000	0.3953	0.3953	23.554	21.901	23.967	10.079	7.7075	9.2885
No Occlusion 2 (η_1^2)	0.3953	0.0000	0.3953	23.554	23.967	24.111	8.6957	8.1028	8.8933
No Occlusion 3 (η_1^3)	0.3953	0.3953	0.0000	21.739	22.530	21.739	9.6838	8.4980	8.4980
Occlusion 1 (η_2^1)	23.554	23.141	22.727	0.0000	1.6529	2.0661	10.277	10.474	10.277
Occlusion 2 (η_2^2)	21.488	22.934	22.134	1.8595	0.0000	1.4463	8.3004	9.0909	9.4862
Occlusion 3 (η_2^3)	23.141	20.356	21.146	2.2727	1.0331	0.0000	8.6957	9.4862	7.7075
Random Occlusion 1 (η_3^1)	10.870	9.0909	10.079	10.079	9.2885	9.2885	0.0000	1.1858	0.1976
Random Occlusion 2 (η_3^2)	9.2885	7.9051	8.8933	12.451	9.8814	11.660	1.3834	0.0000	1.5810
Random Occlusion 3 (η_3^3)	8.4980	7.5099	7.5099	10.474	9.8814	10.277	0.1976	1.3834	0.0000

with 1000 runs each were simulated (denoted with η_i^j for $j = 1, 2, 3$) and the resulting trajectories of all 9 datasets were compared with each of the other datasets using the proposed approach. Due to the negative impact on the results we have already observed when including the heading angle, cf. Figure 6 and Figure 7, we excluded the heading angle from the relevant features for now. Additionally, since the bicyclist is set to a strict trajectory following mode in the simulation, the behavior will be assumed to be the same for all datasets and thus, the considered features are chosen as the x and y coordinate of the ego vehicle. As Table 3 shows, comparisons of datasets generated using the same probability for a static occlusion in the scenario with an α-level of 0.01 result in a measured difference below 2%. One might assume at first that Table 3 should be symmetric, however this is not the case as the estimate of the kernel parameter, and thus also the computed MMD estimate, changes slightly when exchanging the position of the inputs with each other. Naturally, this difference on a rather microscopic level also shows itself on this more macroscopic scale presented in the table. Of course, since the comparison uses hypothesis tests, the α-value chosen for the tests will have an effect on the overall results. Furthermore, one can see from the results that larger differences in the probability of an occlusion between datasets lead to a higher measured difference between their interpretations as logical scenarios modeled as Markov chains. This implies that given a baseline dataset and the output datasets of two models, we can determine which model describes the baseline better with respect to the features deemed relevant.

VII. CONCLUSION

Solely using physical tests to statistically show safety of automated driving appears infeasible. Simulation based testing, on the other hand, seems to be a practical and cost efficient way to provide trustworthiness in automated driving systems before their release to the market. However, when relying on simulation, it needs to be ensured that the employed simulation models are a valid representation of the real world. This continues to be a great challenge and requires substantial effort during the simulation model validation. However, as simulation models are reusable, this validation effort appears worthwhile in the long run.

The presented method for model validation is one way to address the challenge of judging operational validity based on observed simulation runs. It is efficiently and effectively computable, yet suffers like many approaches with increasing dimensionality of the problem space. Thus, it is of high importance to find out which aspects need to be close to reality and should be compared by the presented method. Hence, methods are needed to determine the relevance of features. One promising way is to analyze the causal dependency structures, e.g. with the usage of causal models [47].

Future work includes experiments with hidden Markov models derived from real world data. This tests the approach presented in Subsection V-B. Even though the assumptions appear meaningful in the current scope, necessity and possibility of a relaxation of the requirements for a broader scope of application shall be investigated e.g. allowing the history of the Markov chains to be longer or not requiring the Markov chain to be time-homogeneous. Further, in this work we have employed a biased estimator for the maximum mean discrepancy together with a naive kernel heuristic. However, it could be beneficial to examine other estimators presented by Gretton et al. [13] as well as to investigate different indicators on whether the null hypothesis shall be rejected or not, e.g. Schrab et al. [48]. Moreover, it should be investigated whether our validation method could be calibrated using expert knowledge to influence the weight regarding the decision on acceptance or rejection, e.g. to allow for minor fluctuations in the data. Finally, the method shall be tested in more complex scenarios.

REFERENCES

[1] T. Winkle, *Safety Benefits of Automated Vehicles: Extended Findings from Accident Research for Development, Validation and Testing*. Berlin, Germany: Springer, 2016, pp. 335–364.

[2] A. Poddey, T. Brade, J. E. Stellet, and W. Branz, "On the validation of complex systems operating in open contexts," 2019, *arXiv:1902.10517*.

[3] W. Wachenfeld and H. Winner, *Autonomes Fahren*. Berlin, Germany: Springer, 2015, pp. 439–464.

[4] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" RAND Corp., Santa Monica, CA, USA, Tech. Rep., 2016.

[5] *Status of the Nation's Highways, Bridges, and Transit: Conditions and Performance Report to Congress*, U.S. Department of Transportation, Federal Highway Administration and Federal Transit Administration, Washington, DC, USA, 2021.

[6] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 8, pp. 87456–87477, 2020.

[7] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, "Defining and substantiating the terms scene, situation, and scenario for automated driving," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 982–988.

[8] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1821–1827.

[9] C. Neurohr, L. Westhofen, M. Butz, M. H. Bollmann, U. Eberle, and R. Galbas, "Criticality analysis for the verification and validation of automated vehicles," *IEEE Access*, vol. 9, pp. 18016–18041, 2021.

[10] *OpenSCENARIO V2.0.0*, ASAM, Rockville, MD, USA, 2022.

[11] *Safety of the Intended Functionality*, Standard ISO 21448:2022, Geneva, Switzerland, 2022.

[12] E. Böde, M. Büker, U. Eberle, M. Fränzle, S. Gerwinn, and B. Kramer, "Efficient splitting of test and simulation cases for the verification of highly automated driving functions," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.* Cham, Switzerland: Springer, 2018, pp. 139–153.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773 2012.

[14] R. Gruner, P. Henzler, G. Hinz, C. Eckstein, and A. Knoll, "Spatiotemporal representation of driving scenarios and classification using neural networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1782–1788.

[15] B. Kramer, C. Neurohr, M. Büker, E. Böde, M. Fränzle, and W. Damm, "Identification and quantification of hazardous scenarios for automated driving," in *Model-Based Safety and Assessment*, M. Z. Kai Höfig, Ed. Cham, Switzerland: Springer, 2020, pp. 163–178.

[16] H. F. Stripling, M. L. Adams, R. G. McClarren, and B. K. Mallick, "The method of manufactured universes for validating uncertainty quantification methods," *Rel. Eng. Syst. Saf.*, vol. 96, no. 9, pp. 1242–1256, Sep. 2011.

[17] S. Schlesinger, "Terminology for model credibility," *Simulation*, vol. 32, no. 3, pp. 103–104, 1979.

[18] S. Ferson, W. L. Oberkampf, and L. Ginzburg, "Model validation and predictive capability for the thermal challenge problem," *Comput. Methods Appl. Mech. Eng.*, vol. 197, nos. 29–32, pp. 2408–2430, May 2008.

[19] C. Beisbart and N. J. Saam, *Computer Simulation Validation*. Cham, Switzerland: Springer, 2019.

[20] R. G. Sargent, "Verification and validation of simulation models," in *Proc. Winter Simulation Conf. (WSC)*, Dec. 2011, pp. 183–198.

[21] R. Allemang, M. M. Kolluri, M. Spottswood, and T. Eason, "Decomposition-based calibration/validation metrics for use with full-field measurement situations," *J. Strain Anal. Eng. Des.*, vol. 51, no. 1, pp. 14–31, Jan. 2016.

[22] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton, "Generative models and model criticism via optimized maximum mean discrepancy," 2016, *arXiv:1611.04488*.

[23] M. Viehof and H. Winner, "Stand der technik und der wissenschaft: Modellvalidierung im anwendungsbereich der fahrdynamiksimulation," Tech. Univ. Darmstadt, Darmstadt, Germany, Tech. Rep., 2017.

[24] M. Viehof and H. Winner, "Research methodology for a new validation concept in vehicle dynamics," *Automot. Engine Technol.*, vol. 3, nos. 1–2, pp. 21–27, Aug. 2018.

[25] P. Rosenberger, J. T. Wendler, M. Holder, C. Linnhoff, H. Winner, and M. Maurer, "Towards a generally accepted validation methodology for sensor models—Challenges, metrics, and first results," in *Proc. Graz Symp. Virtual Vehicle (GSVF)*, 2019.

[26] P. Rosenberger, G. Schunk, F. Ikemeyer, and T. D. Quang, "Validation of test infrastructure—From cause trees to a validated system simulation," in *Proc. VVM Project, Mid-Term Presentation*, 2022.

[27] S. Riedmaier, B. Danquah, B. Schick, and F. Diermeyer, "Unified framework and survey for model verification, validation and uncertainty quantification," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 2655–2688, Jun. 2021.

[28] S. Riedmaier, J. Schneider, B. Danquah, B. Schick, and F. Diermeyer, "Non-deterministic model validation methodology for simulation-based safety assessment of automated vehicles," *Simul. Model. Pract. Theory*, vol. 109, May 2021, Art. no. 102274.

[29] B. Danquah, S. Riedmaier, J. Rühm, S. Kalt, and M. Lienkamp, "Statistical model verification and validation concept in automotive vehicle design," *Proc. CIRP*, vol. 91, pp. 261–270, Jan. 2020.
[30] S. Davies, A. Mazumdar, S. Pal, and C. Rashtchian, "Lower bounds on the total variation distance between mixtures of two Gaussians," in *Proc. 33rd Int. Conf. Algorithmic Learn. Theory*, vol. 167, S. Dasgupta and N. Haghtalab, Eds., Apr. 2022, pp. 319–341.
[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[32] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer, "Imitating driver behavior with generative adversarial networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 204–211.
[33] B. Paden, M. Cáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 33–55, Mar. 2016.
[34] R. C. Magel and S. H. Wibowo, "Comparing the powers of the Wald–Wolfowitz and Kolmogorov–Smirnov tests," *Biometrical J.*, vol. 39, no. 6, pp. 665–675, 1997.
[35] M. Naaman, "On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality," *Statist. Probab. Lett.*, vol. 173, Jun. 2021, Art. no. 109088.
[36] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman, "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3571–3577.
[37] V. I. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces* (Cambridge Studies in Advanced Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2016.
[38] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.
[39] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[40] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, "Universality, characteristic kernels and RKHS embedding of measures," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 2389–2410, 2010.
[41] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *J. Mach. Learn. Res.*, vol. 7, pp. 2651–2667, Dec. 2006.
[42] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[43] P. Fritzson et al., "The OpenModelica integrated environment for modeling, simulation, and model-based development," *Model., Identificat. Control, Norwegian Res. Bull.*, vol. 41, no. 4, pp. 241–295, 2020.
[44] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, 2017, pp. 1–16.
[45] S. M. LaValle, *Planning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[46] W. Damm, E. Möhlmann, and A. Rakow, *A Scenario Discovery Process Based on Traffic Sequence Charts*. Cham, Switzerland: Springer, 2020, pp. 61–73.
[47] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[48] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton, "MMD aggregated two-sample test," 2021, *arXiv:2110.15073*.

BIRTE NEUROHR received the B.Sc. and M.Sc. degrees in mathematics from the Carl von Ossietzky Universität Oldenburg, in 2015 and 2017, respectively. From 2017 to 2020, she was a Researcher and the Ph.D. Candidate with OFFIS e.V., where she was working in the area of scenario-based verification and validation of automated vehicles. Since 2021, she has been with the Institute of Systems Engineering for Future Mobility, German Aerospace Center, as a Researcher and the Ph.D. Candidate. Her main research interest includes validity and quality criteria for simulations.

TJARK KOOPMANN received the B.Sc. and M.Sc. degrees in mathematics from the University of Oldenburg, Germany, in 2018 and 2020, respectively. Since then, he has been with the Transportation Division, OFFIS e.V. (Research Institute), that became the Institute for Systems Engineering for Future Mobility, German Aerospace Center (DLR e.V.), in 2022. His main research area is in the scenario-based safeguarding of automated driving systems, especially in the domain of causal analysis of criticality increasing factors. Further, he is also concerned with research regarding methods for the validation of simulation models and simulation environments.

EIKE MÖHLMANN received the degree in computer science and the Ph.D. degree from the University of Oldenburg, Germany, in 2010 and 2018, respectively. From 2010 to 2016, he was a Research Assistant with the Transregional Collaborative Research Center Automatic Verification and Analysis of Complex Systems, University of Oldenburg. He continued working as a Research Assistant with OFFIS e.V., Oldenburg, and took over the lead of the group "Safety and Security-Oriented Analysis," in 2018. Since 2022, he has been leading the Evidence for Trustworthiness Group, Institute for Systems Engineering for Future Mobility, DLR. His research interests include systems engineering, formal modeling, formal verification, contract-based design, safety processes, simulation-based testing, and AI in safety-critical applications.

MARTIN FRÄNZLE received the Diploma degree in informatics from the Department of Informatics, Christian-Albrechts Universität Kiel, Germany, in 1991, and the Ph.D. degree in natural sciences from the Technical Faculty of Christian-Albrechts Universität Kiel, in 1997. He is currently a Full Professor of foundations and applications of cyber-physical systems with the Department of Computing Science, Carl von Ossietzky Universität Oldenburg, Germany. At the university and the international Ph.D. schools, he taught formal methods for embedded system design, the theory and application of hybrid discrete-continuous systems, design principles of autonomous systems, and related subjects. His research interests include modeling, verification, and synthesis of reactive, real-time, hybrid, and human-cyber-physical systems; applications in advanced driver assistance; highly automated cars and autonomous driving; and power networks.