Analysis and comparison of similarity measures for validation of generative algorithms in the context of probability density functions

Nico Schick[1] and Roberto Corlito[2]
[1]*N. Schick studied Applied Computer Sciences (M. Sc.) and Computer Engineering (B. Eng.) at the Esslingen University of Applied Sciences. e-mail: Nico.Schick@hs-esslingen.de*
[2]*R. Corlito studied Computer Engineering (B. Eng.) at the Esslingen University of Applied Sciences. e-mail: Roberto.Corlito@stud.hs-esslingen.de*

Increasingly, machine learning algorithms are being used in the field of autonomous driving. Here, generative algorithms can be used to provide further data corresponding to driving situations. This type of algorithm is based on probability distributions. As a consequence, appropriate similarity measures can be used to validate them quantitatively. This paper answers the following scientific question:

Which similarity measures are suitable for validating a generative algorithm in the context of safety-critical driving scenarios of autonomous driving?

Keywords: Autonomous driving, safety-critical driving scenario, generative algorithm, probability density function, similarity measure

I. Motivation

About 3700 people die in traffic accidents every day. Human error is the number one cause of accidents. Autonomous driving can greatly reduce the occurrence of traffic accidents. To release self-driving cars for road traffic, the system including software must be validated and tested efficiently. However, due to their criticality, the amount of data corresponding to safety-critical driving scenarios are limited. These driving scenes can be expressed as a time series. They represent the corresponding movement of the vehicle, including time vector, position coordinates, speed and acceleration. Such data can be provided on different ways. For example, in the form of a kinematic model. Alternatively, artificial intelligence or machine learning methods can be used. They have been widely used in the development of autonomous vehicles. For example, generative algorithms can be used to generate such safety-critical driving data. However, the validation of generative algorithms is a challenge in general. In most cases, their quality is assessed by means of expert knowledge (qualitative). In order to achieve a higher degree of automation, a quantitative validation approach is necessary. Generative algorithms are based on probability distributions or probability density functions. Accordingly, similarity measures can be used to evaluate generative algorithms. In this publication, such similarity measures are described and compared on the basis of defined evaluation criteria. With respect to the use case mentioned, a recommended similarity measure is implemented and validated for an example of a typical safety-critical driving scenario. [1] [2] [3] [4] [5] [6] [7]

II. Generative Algorithms

Generative algorithms belong to the field of machine learning and have gained significant attention in recent years. This type of algorithm can be understood as the counterpart to discriminative algorithms. They can, based on a probability distribution, generate new data. For this, information about the characteristics of the features are needed. In this regard, the characteristics of the different observations from the training data set are transformed into a probability model. In the course of a stochastic process, the generative algorithm approximates the probability distribution of the training data set. The training phase of the probability model can be considered complete as soon as new observations hardly differ from the original data set. Classical generative algorithms do not require labeled data sets (unsupervised learning). However, mixed forms also exist, which also depend on labels (unsupervised and supervised learning). Different taxonomies of generative algorithms can also be derived. [8]

III. Probability density functions

In everyday life, there are many processes whose results depend on probabilities. For those processes, it is not possible to predict which result will occur. Such random experiments are also referred to as non-deterministic processes. The individual elementary events of the random experiment or the realizations of a random variable can be assigned to a result set. A random variable is considered to be continuous if it can assume any real value at least in a certain interval. Since a finite or infinite interval of real numbers is not constituted by a countably infinite number of values, the probability that a continuous random variable takes one specific value is zero. Probability density functions (PDF) can be used to specify the probability of occurrence of a particular realization of the random variable. A PDF depends on the realization of the random variable X and is defined as follows:

$$P(a < X \leq b) = \int_{a}^{b} f(x)\,dx \;\wedge\; \int_{-\infty}^{+\infty} f(x)\,dx = 1, \quad f(x) \geq 0, \quad a, b \in \mathbb{R}, \quad a \leq b \tag{1}$$

The probability P corresponds to the product $f(x)\,dx$. This is the probability that the continuous random variable X takes a value in an arbitrarily small interval $[x, x + dx]$. In practice, the PDF of an underlying random variable is typically unknown. Accordingly, it is important, depending on the use case considered, to estimate these not only qualitatively, but also quantitatively. There are several methods to estimate appropriate PDFs. So-called kernel density estimators (KDE) can be used. An alternative is the so-called histogram-spline approximation. [7] [9] [10]

IV. Similarity measures of probability density functions

Generative algorithms generate new data sets based on an original data set. These data sets are subject to their corresponding PDFs. In addition to the qualitative consideration of both PDFs in the form of expert knowledge, similarity measures can be used as a quantitative approach. This includes the statistical distance between two PDFs. A distance (or dissimilarity) $d(x, y)$ is a function that maps a set X to the set of real numbers, $d : X \times X \to \mathbb{R}$. Here $x, y \in X$ are mapped by d onto the real numbers. Such a distance has to satisfy $d(x, y) \geq 0$ (non-negativity), $d(x, y) = d(y, x)$ (symmetry), and $d(x, x) = 0$ (reflexivity). Another way to determine the discrepancy of two density

functions is with respect to similarity. Similar to a distance, this maps a quantity to the set of real numbers, $s : X \times X \rightarrow \mathbb{R}$. Besides the properties of non-negativity and symmetry, a similarity has to satisfy $s(x, y) \leq s(x, x)$ for all $x, y \in X$ except for $y = x$. Similarities s and distances d can be transformed into each other by various transformations. For example, by $d = 1 - s$, $d = \frac{1-s}{s}$, $d = \sqrt{1-s}$ or $d = -\ln(s)$. The more similar x and y are, the higher the value of the similarity s. In contrast, the more similar x and y are, the lower is the distance d. In both cases, the corresponding ratios can be normalized. Then $0 \leq s(x, y) \leq 1$ or $0 \leq d(x, y) \leq 1$. The definition of a distance can be extended to that of a semi-metric. For this, in addition to non-negativity, symmetry and reflexivity, the triangle inequality $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$ must hold. A metric is a semi-metric that additionally satisfies the identity $d(x, y) = 0 \Leftrightarrow x = y$. A set X, including a metric d, is called a metric space (X, d). These evaluation techniques provide the framework to compare density functions, in the context of generative algorithms. Distances that are asymmetric can be transformed into symmetric ones. Some of these transformation possibilities are mentioned in Table 1. [7] [11] [12]

Method	**Description**
Addition	$d_{sym}(\mathbb{P}, \mathbb{Q}) = d_{asym}(\mathbb{P}, \mathbb{Q}) + d_{asym}(\mathbb{Q}, \mathbb{P})$
Maximum	$d_{max-sym}(\mathbb{P}, \mathbb{Q}) = \max\left(d_{asym}(\mathbb{P}, \mathbb{Q}), d_{asym}(\mathbb{Q}, \mathbb{P})\right)$
Minimum	$d_{min-sym}(\mathbb{P}, \mathbb{Q}) = \min\left(d_{asym}(\mathbb{P}, \mathbb{Q}), d_{asym}(\mathbb{Q}, \mathbb{P})\right)$
Average	$d_{avg-sym}(\mathbb{P}, \mathbb{Q}) = avg\left(d_{asym}(\mathbb{P}, \mathbb{Q}), d_{asym}(\mathbb{Q}, \mathbb{P})\right)$

Table 1 Transformation approaches of symmetric and asymmetric distances

For comparing two PDFs, there are various similarity measures in the literature. Based on the corresponding mathematical descriptions and the focus on generative algorithms, a corresponding taxonomy can be derived. It is presented in Figure 1. [7]

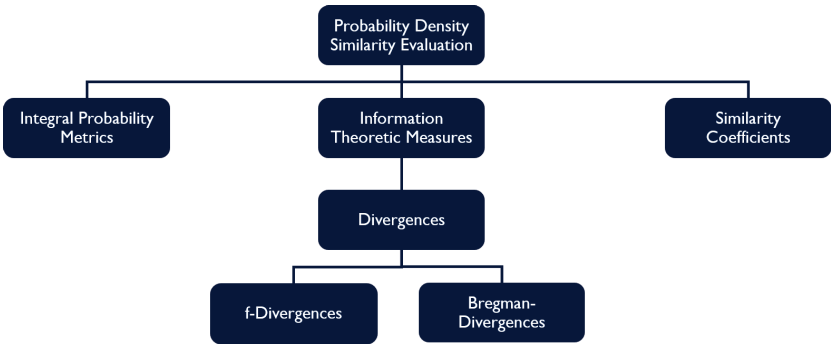

Fig. 1 Taxonomy of evaluation techniques related to PDF comparisons

In preparation for the next chapter, the following information is described in [7] in more detail. An essential group of similarity measures are given by the Integral Probability Metrics. The basic

structure of these metrics is the difference of two probability measures $\mathbb{P}$ and $\mathbb{Q}$. They are defined as follows: [13]

$$\gamma_F(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int f \, d\mathbb{P} - \int f \, d\mathbb{Q} \right| \tag{2}$$

Here the function f belongs to a class of real and bounded functions $\mathcal{F}$. Examples of Integral Probability Metrics are given by the Kolmogorov–Smirnov Distance [13] [14], Wasserstein Distance [15] [16], Cramér Distance [17], Minkowski Metric [11] [17], Maximum Mean Discrepancy [13] and Total Variation Metric [13]. Another important group of similarity measures is constituted by the information theoretic measures. Here, the measure of the entropy should be mentioned as well. This is a measure of the information content or uncertainty of a random variable X and is defined as follows: [18]

$$H(X) = -\sum_{i=1}^{N} x_i \cdot \log_2(x_i), \; i = 1, \ldots, N \tag{3}$$

Many statistical distances are based on entropy. Divergences are also significant for comparing density functions. It should be mentioned that many similarity measures (e.g. divergences) aren't metrics. However, this does not rule them out as suitable evaluation measures in general. Their wide use in the literature shows the significance of these evaluation techniques for the underlying use case also. One of the best known families of divergence measures is given by the f-divergences. They include a large number of the known statistical distances and are defined as follows:

$$d_f(p, q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) \tag{4}$$

Here, f is any function that is convex over the domain of definition $(0, \infty)$ and for which $f(1) = 0$. Furthermore, f-divergences are always non-negative, and zero only if the two PDFs $p(x)$ and $q(x)$ are equal. [19] Examples of f-divergences are given by the Jensen–Shannon Divergence [11] [13] [17], Kullback–Leibler Divergence [11], χ^2 Distance [11] and Hellinger Distance [13]. Another important group of similarity measures are the Bregman divergences. These measure the discrepancy between two values of PDFs $p(x)$ and $q(x)$: [19]

$$d_\varphi(p, q) = \varphi(p) - \varphi(q) - (p - q)\varphi'(q) \tag{5}$$

The total discrepancy between the PDFs $p(x)$ and $q(x)$, can be described, using Bregman divergence, as follows:

$$d_\varphi(\mathbb{P}, \mathbb{Q}) = \int \left[\varphi(p(x)) - \varphi(q(x)) - (p(x) - q(x))\varphi'(q(x))\right] \, dx \tag{6}$$

This calculation rule can also be discretized:

$$d_\varphi(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{N} \left[\varphi(p_i) - \varphi(q_i) - (p_i - q_i)\varphi'(q_i)\right] \tag{7}$$

Here, the function $\varphi(t)$ must be strictly convex and real. The term $\varphi'(\mathbb{Q})$ describes the derivative with respect to $\mathbb{Q}$. Bregman divergences also possess the property of non-negativity. Their symmetry, however, depends on $\varphi(t)$. Examples of Bregman divergences are given by the Mahalanobis Distance [20] and the Itakura–Saito Distance [21].

Similarity coefficients can also be used to compare two PDFs [11]. Here, the two PDFs P and Q are compared by means of different operations. It should be mentioned that these do not include exclusively measures based on a similarity. These also include measures based on the inner product, such as the cosine distance. Other distances are based, for example, on the sum of the geometric mean, such as the Bhattacharyya distance. Another similarity measure, which is based on the absolute distance, is the Canberra distance. An alternative approach is given by similarity measures from other use cases. These include the Dynamic Time Warp algorithm, used in the context of time series [22], and the Structural Similarity Index, used in the context of image processing [23]. These similarity measures can likewise be applied to vector data sets, such as two PDFs P and Q. [24]

A. Comparison based on evaluation criteria

In order to find a suitable similarity measure for the intended use case, these similarity measures will be compared with each other. For this purpose, the following evaluation criteria are defined and categorized in Table 2.

Effectiveness (E) The effectiveness of an evaluation technique is representative of the quality or goodness of the underlying similarity measure. In the present context, this can be considered the most important criterion. The quality depends, somewhat strongly, on the corresponding data. Another aspect of great relevance is the development or the course of the similarity values. These should deliver plausible values over the entire range of values.

Complexity (K) The complexity of an evaluation technique includes its time complexity. This can be determined on the basis of the respective implementation. Here the 'Big O notation' – also referred to as the Landau symbol $\mathcal{O}$ – is used. Especially, the complexity depends on the number of data points. It also influences the runtime and efficiency of the algorithm applied.

Applicability (A) The applicability of an evaluation technique depends on how easily it can be applied. It has to be considered whether the underlying implementation involves a high degree of effort, e.g. in the form of optimization processes.

Transparency (T) Transparency is a measure of the clarity and scope of the evaluation technique, in particular, whether there is enough in the literature describing this method in a well-founded and detailed manner.

Robustness (R) The robustness of an evaluation technique describes how well it can handle measurement inaccuracies, e.g. peaks, noise or outliers.

Parametrizability (P) The parametrizability indicates whether the evaluation technique contains parameters and, if so, how easily they can be set, according to quality aspects. For example, there are parameters that are typically determined via optimization processes or empirical methods.

Interpretability (I) Interpretability indicates how well the data can be interpreted from the results of the evaluation technique. In this regard, methods that are standardized, limited to a range of values, or converge to a certain value have a high degree of interpretability. If this is not the case, the results, based on expert knowledge or empirical values, can be evaluated.

Effectiveness	+	Evaluation technique has a high quality.
	o	Evaluation technique has a moderate quality.
	-	Evaluation technique has a low quality.
Complexity	+	Evaluation technique has low time complexity and runtime.
	o	Evaluation technique has moderate time complexity and runtime.
	-	Evaluation technique has high time complexity and runtime.
Applicability	+	Evaluation technique is basically easy to implement or reference implementations are available.
	o	Evaluation technique can only be implemented easily to a limited extent. No reference implementations are available.
	-	Evaluation technique is fundamentally difficult to implement and reference implementations are not available.
Transparency	+	Evaluation technique is transparently described and understandable.
	o	Evaluation technique is described rather moderately transparently (e.g. mathematical description not fully documented).
	-	Evaluation technique requires in-depth mathematical knowledge or its description is incomplete.
Robustness	+	Evaluation technique is robust against inaccuracies in the data (e.g. measurement errors, noise or outliers).
	o	Evaluation technique is not completely robust against inaccuracies in the data.
	-	Evaluation technique is not robust against inaccuracies in the data.
Parametrizability	+	Evaluation technique does not include parameters or includes only a few parameters. If it does include parameters, then they are easily adjustable.
	o	Evaluation technique includes several parameters. These are not all easily adjustable.
	-	Evaluation technique includes several or many parameters. Most of them are not easily adjustable.
Interpretability	+	The result of the evaluation is easy to interpret.
	o	The result of the evaluation is moderately interpretable.
	-	The result of the evaluation is difficult to interpret.

Table 2 Classification of the individual evaluation criteria

In general, the sign + stands for a positive, the sign o for a neutral, and the sign - for a negative characteristic. Based on the evaluation criteria and their classification, the similarity measures mentioned can be compared, as seen in Table 3.

Similarity measure	**E**	**K**	**A**	**T**	**R**	**P**	**I**
Kolmogorov–Smirnov Distance	+	$+\,O(n)$	o	+	-	-	+
Wasserstein Distance	+		o	o		-	+
Cramér Distance	o	$+\,O(n)$	o	+	o	+	+
Minkowski Metric	o	$+\,O(n)$	+	+	o	o	+
Maximum Mean Discrepancy	o		o	o	+	-	+
Total Variation Metric	+	$+\,O(n)$	+	+	o	+	+
Kullback–Leibler Divergence	+	$+\,O(n)$	+	+	+	+	+
Jensen–Shannon Divergence	+	$+\,O(n)$	+	+	+	+	+
χ^2 Distance	+	$+\,O(n)$	+	+	-	+	o
Hellinger Distance	+	$+\,O(n)$	+	+	o	+	+
Mahalanobis Distance	o	+	+	+	+	o	o
Itakura–Saito Distance	+	$+\,O(n)$	+	o	o	+	o

Table 3 Evaluation of the individual similarity measures according to evaluation criteria

A few of the cells in Table 3 are coloured grey. For these cells, the classification depends on the specific mathematical description of the underlying evaluation technique. Based on the assessment and comparison of the evaluation techniques carried out in Table 3, three very important methods emerge, namely, the Kullback–Leibler and Jensen–Shannon Divergence as well as the Hellinger Distance. Each of these evaluation techniques has, with respect to the evaluation criteria mentioned, positive classifications exclusively. In the case of such equality, it is best practice to make a final decision based on citations in the literature. This points to the Kullback–Leibler divergence. For this reason, the Kullback–Leibler divergence is recommended as an evaluation technique for the present use case. The Kullback–Leibler divergence results from using $f(t) = t\ln(t)$ in Equation 4. The discrete variant of the Kullback–Leibler divergence is given by [11]

$$d_{KL}(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{N} p_i \ln\left(\frac{p_i}{q_i}\right) \tag{8}$$

B. Implementation and validation of the recommended similarity measure

For validation, the recommended similarity measure, namely, the Kullback–Leibler divergence, was implemented. Equation 8 is used as reference. Overall, the function prototype is defined as follows:

$$[y_1, y_2, y_3] = f(P, Q), \quad y_1, y_2, y_3, P, Q \in \mathbb{R} \tag{9}$$

Here, two parameters P and Q are passed to the function. These correspond to the PDFs. The return value y_1 represents the actual Kullback–Leibler divergence. In addition to this, two further key values are also provided for the validation. For this purpose, the Kullback–Leibler divergence is calculated iteratively. The value y_2 describes the maximum single value of the Kullback–Leibler divergence between two points of the PDFs (maximum discrepancy). This is a measure of the highest discrepancy of two values of the corresponding PDFs. The value y_3, indicates the maximum gradient of the respective individual values. The values y_2 and y_3 check whether P and Q lie in plausible ranges of values.

The implementation has been validated using suitable data. For this purpose, an overtaking maneuver is used as a safety-critical driving scenario. Such an overtaking maneuver is described in [6], in the form of a kinematic model. Based on the underlying model, a data set with 100 different time series was created. Each time series has a length of 1602 data points. The corresponding variation parameters γ_1 and γ_2 have values lying in a range of ± 0.1 and a step size of 10^{-3}. These value ranges are plausible with respect to real drives and results in different motion curves of the overtaking vehicle. The variations are particularly visible when the vehicle is turning into and out of the lane. This simplifies the vehicle-related validation of the PDF in these driving periods. The corresponding time series are shown in Figure 2 for the longitudinal and transverse directions. For a transverse direction close to zero, the vehicle is in the original lane, and for a transverse direction close to three meters, it is in the opposite lane. Such an overtaking process can also be represented as a PDF. Here, the transverse direction is clearly more concise than the longitudinal direction of the vehicle movement. Accordingly, Figure 3 shows the relative probabilities of the underlying data across the transverse direction.

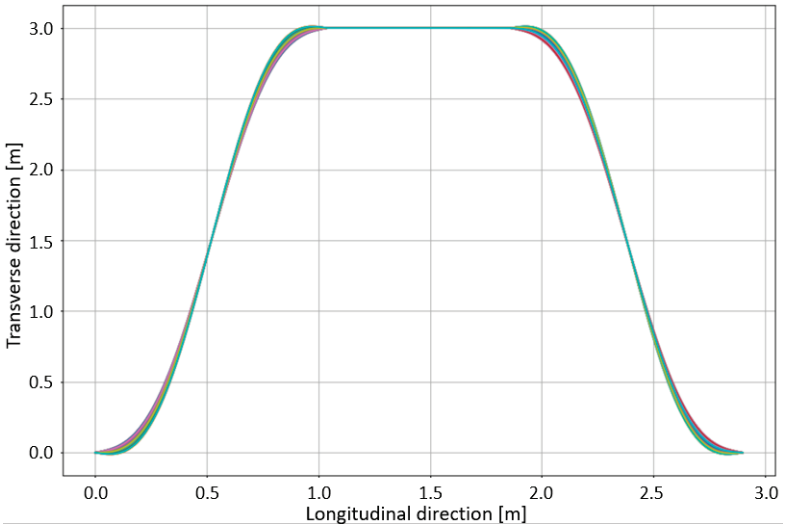

Fig. 2 Motion data of synthetically generated overtaking processes

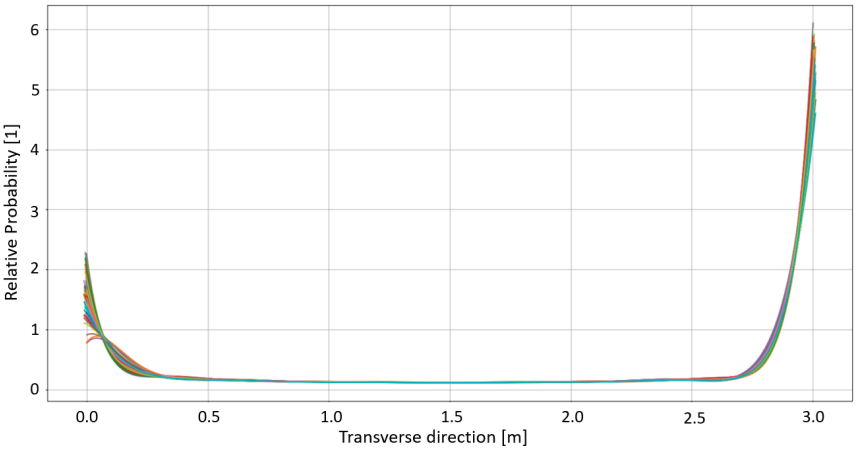

Fig. 3 PDFs of synthetically generated overtaking processes

The individual PDFs were determined using histogram spline approximations. Here, 13 bins are used for the underlying histogram and a spline of 11th order. Overall, the variations in the individual PDFs become apparent. These variations result from the different motion profiles as well as the approximation to the PDF. In particular, the relative probabilities are relatively high when the vehicle is turning into and out of the lane. In the opposite lane, however, the relative probabilities are relatively low.

In general, the corresponding Kullback–Leibler distances are of interest. Here, the PDF without variations is used as reference. Figure 4 shows the Kullback–Leibler distances of the remaining 99 samples. The result is always a value close to zero. Thus, these values indicate a high similarity of the density functions. The Kullback–Leibler distances for each data point of the remaining 99 samples are visualized in Figure 5. Here, all values are close to zero as well. These values also indicate a high similarity of the PDFs. The curves show a high variance when the vehicle is turning into and out of the lane in particular. In between, there is an approximately linear course.

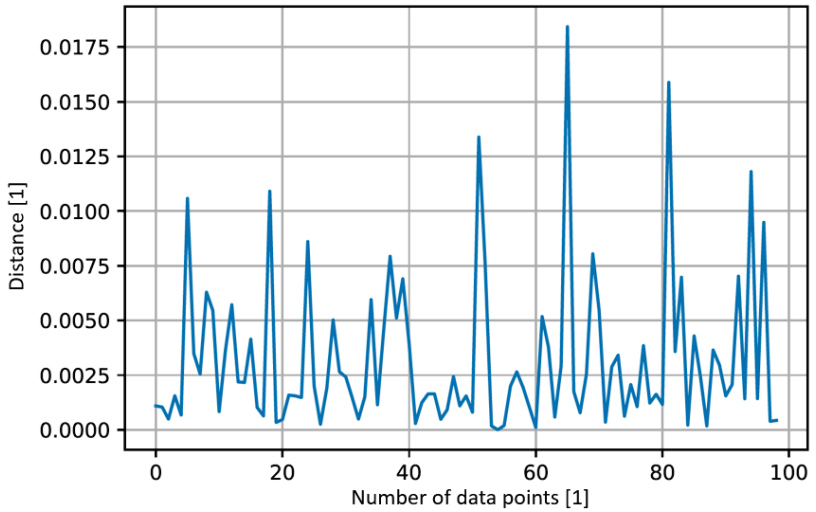

Fig. 4 **Kullback–Leibler distances for PDFs of synthetically generated overtaking processes**

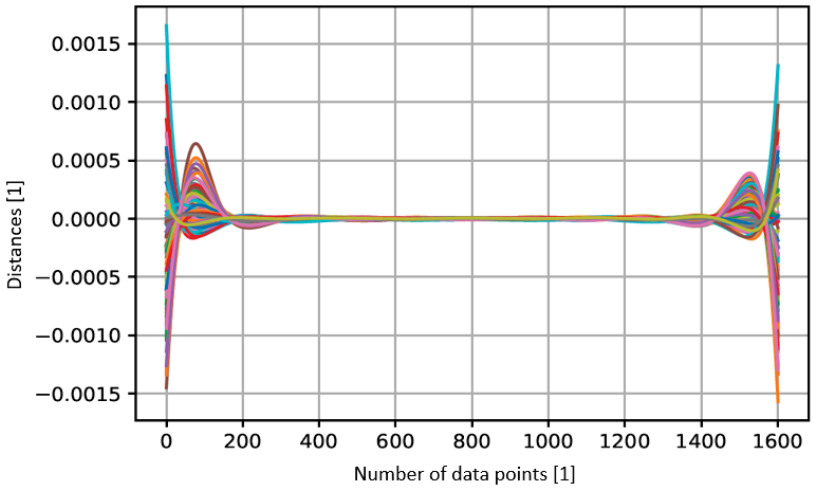

Fig. 5 **Single values of Kullback–Leibler distances for PDFs of synthetically generated overtaking processes**