A Recurrence Plot-Based Distance Measure

Stephan Spiegel, Johannes-Brijnesh Jain and Sahin Albayrak

Abstract Given a set of time series, our goal is to identify prototypes that cover the maximum possible amount of occurring subsequences regardless of their order. This scenario appears in the context of the automotive industry, where the goal is to determine operational profiles that comprise frequently recurring driving behavior patterns. This problem can be solved by clustering, however, standard distance measures such as the dynamic time warping distance might not be suitable for this task, because they aim at capturing the cost of aligning two time series rather than rewarding pairwise recurring patterns. In this contribution, we propose a novel time series distance measure, based on the notion of recurrence plots, which enables us to determine the (dis)similarity of multivariate time series that contain segments of similar trajectories at arbitrary positions. We use recurrence quantification analysis to measure the structures observed in recurrence plots and to investigate dynamical properties, such as determinism, which reflect the pairwise (dis)similarity of time series. In experiments on real-life test drives from Volkswagen, we demonstrate that clustering multivariate time series using the proposed recurrence plot-based distance measure results in prototypical test drives that cover significantly more recurring patterns than using the same clustering algorithm with dynamic time warping distance.

1 Introduction

Clustering of times series data is of pivotal importance in various applications [1] such as, for example, seasonality patterns in retail [2], electricity usage profiles [3], DNA microarrays [4], and fMRI brain activity mappings [5]. A crucial design decision

S. Spiegel (✉) · J.-B. Jain · S. Albayrak
DAI Laboratory, Berlin Institute of Technology, Ernst-Reuter-Platz 7,
10587 Berlin, Germany
e-mail: spiegel@dai-lab.de

J.-B. Jain
e-mail: jain@dai-lab.de

S. Albayrak
e-mail: albayrak@dai-lab.de

N. Marwan et al. (eds.), *Translational Recurrences*, Springer Proceedings
in Mathematics & Statistics 103, DOI 10.1007/978-3-319-09531-8_1

of any clustering algorithm is the choice of (dis)similarity function [6, 7]. In many clustering applications, the underlying (dis)similarity function measures the cost of aligning time series to one another. Typical examples of such functions include the DTW and the Euclidean distance [8–10].

Alignment-based (dis)similarity functions, however, seem not to be justified for applications, where two time series are considered to be similar, if they share common or similar subsequences of variable length at arbitrary positions [11–14]. A real-life example for such an application comes from the automotive industry, where test drives of vehicles are considered to be similar, if they share similar driving behavior patterns, i.e. engine behavior or drive maneuvers, which are described by the progression of multiple vehicle parameters over a certain period of time [15, 16]. In this scenario, the order of the driving behavior patterns does not matter [17], but the frequency with which the patterns occur in the contrasted time series.

Recent work [18] on time series distance measures suggests to neglect irrelevant and redundant time series segments, and to retrieve subsequences that best characterize the real-life data. Although subsequence clustering is a tricky endeavor [19], several studies [11–14, 20] have demonstrated that in certain circumstances ignoring sections of extraneous data and keeping intervals with high discriminative power contributes to cluster centers that preserve the characteristics of the data sequences. Related concepts that have been shown to improve clustering results include time series motifs [11, 12], shapelets [13, 14], and discords [20].

In this contribution, we propose to adopt recurrence plots (RPs) [21–23] and related recurrence quantification analysis (RQA) [24–26] to measure the similarity between multivariate time series that contain segments of similar trajectories at arbitrary positions in time [17]. We introduce the concept of joint cross recurrence plots (JCRPs), an extension of traditional RPs, to visualize and investigate multivariate patterns that (re)occur in pairwise compared time series. In dependence on JCRPs and known RQA measures, such as determinism, we define a **R**ecu**RR**ence plot-based (RRR) distance measure, which reflects the proportion of time series segments with similar trajectories or recurring patterns respectively.

In order to demonstrate the practicability of our proposed recurrence plot-based distance measure, we conduct experiments on both synthetic time series and real-life vehicular sensor data [15–17]. The results show that, unlike commonly used (dis)similarity functions, our proposed distance measure is able to (i) determine cluster centers that preserve the characteristics of the data sequences and, furthermore, (ii) identify prototypical time series that cover a high amount of recurring patterns. The rest of the paper is organized as follows. In Sect. 2 we state the general problem being investigated. Subsequently we introduce traditional recurrence plots as well as various extensions in Sect. 3. Recurrence quantification analysis and corresponding measures are discussed in Sect. 4. Our proposed recurrence plot-based distance measure and respective evaluation criteria are introduced in Sect. 5. The experiments results are presented and discussed in Sect. 6. Finally we conclude with future work in Sect. 7.

2 Problem Statement

Car manufacturers aim to optimize the performance of newly developed engines according to operational profiles that characterize recurring driving behavior. To obtain real-life operational profiles for exhaust simulations, Volkswagen (VW) collects data from test drives for various combinations of driver, vehicle and route.

Given a set $\mathscr{X} = \{X_1, X_2, \ldots, X_t\}$ of t test drives, the challenge is to find a subset of k prototypical time series $\mathscr{Y} = \{Y_1, \ldots, Y_k\} \in \mathscr{X}$ that best comprehend the recurring (driving behavior) patterns found in set $\mathscr{X}$. Test drives are represented as multivariate time series $X = (x_1, \ldots, x_n)$ of varying length n, where $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector summarizing the observed measurements at time i. A *pattern* $S = (x_s, \ldots, x_{s+l-1})$ of $X = (x_1, \ldots, x_n)$ is a subsequence of l consecutive time points from X, where $l \leq n$ and $1 \leq s < s + l - 1 \leq n$. Assuming two time series $X = (x_1, ..., x_n)$ and $Y = (y_1, \ldots, y_m)$ with patterns $S = (x_s, \ldots, x_{s+l-1})$ and $P = (y_p, \ldots, y_{p+l-1})$ of length l, we say that S and P are *recurring patterns* of X and Y if $d(S, P) \leq \varepsilon$, where and $d : X \times X \to \mathbb{R}^+$ is a (dis)similarity function and ε is a certain similarity threshold. Note that recurring patterns of X and Y may occur at arbitrary positions and in different order.

Since we aim to identify k prototypical time series that (i) best represent the set $\mathscr{X}$ and (ii) are members of the set $\mathscr{X}$, one can employ the k-mediod clustering algorithm.

3 Recurrence Plots

Recurrence plots (RPs) are used to visualize and investigate recurrent states of dynamical systems or rather time series [26, 27]. Even though RPs give very vivid and impressive images of dynamical system trajectories, their implicit mathematical foundation is deceptively simple [21]:

$$R^x_{i,j}(\varepsilon) = \Theta(\varepsilon - ||x_i - x_j||) \quad x_i \in \mathbb{R}^d, i, j = 1 \ldots n \tag{1}$$

where x is a time series of length n, $||\cdot||$ a norm and Θ the Heaviside function. One of the most crucial parameters of RPs is the recurrence threshold ε, which influences the formation of line structures [22]. In general, the recurrence threshold should be chosen in a way that noise corrupted observations are filtered out, but at the same time a sufficient number of recurrence structures are preserved. As a rule of thumb, the recurrence rate should be approximately one percent with respect to the size of the plot. For quasi-periodic processes, it has been suggested to use the diagonal line structures to find the optimal recurrence threshold. However, changing the threshold does not preserve the important distribution of recurrence structures [26].

A general problem with standard thresholding methods is that an inappropriate threshold or laminar states cause thick diagonal lines, which basically corresponds to redundant information. Schultz et al. [27] have proposed a local minima-based thresholding approach, which can be performed without choosing any particular threshold and yields in clean RPs of minimized line thickness. But this approach comes with some side-effects, e.g., bowed lines instead of straight diagonal lines.

Furthermore, it is important discuss the definition of recurrences, because distances can be calculated using different norms [21]. Although the L_2-norm is used in most cases, the L_∞-norm is sometimes preferred for relatively large time series with high computational demand [26].

Although traditional RPs only regard one trajectory, we can extend the concept in a way that allows us to study the dynamics of two trajectories in parallel [23]. A cross recurrence plot (CRP) shows all those times at which a state in one dynamical system occurs in a second dynamical system. In other words, the CRP reveals all the times when the trajectories of the first and second time series, x and y, visits roughly the same area in the phase space. The data length, n and m, of both systems can differ, leading to a non-square CRP matrix [22, 24].

$$CR_{i,j}^{x,y}(\varepsilon) = \Theta(\varepsilon - ||x_i - y_j||) \quad x_i, y_j \in \mathbb{R}^d,\ i = 1 \ldots n,\ j = 1 \ldots m \tag{2}$$

For the creation of a CRP, both trajectories, x and y, have to present the same dynamical system with equal state variables because they are in the same phase space. The application of CRPs to absolutely different measurements, which are not observations of the same dynamical system, is rather problematic and requires some data preprocessing with utmost carefulness [22].

In order to test for simultaneously occurring recurrences in different systems, another multivariate extension of RPs was introduced [23]. A joint recurrence plot (JRP) shows all those times at which a recurrence in one dynamical system occurs simultaneously with a recurrence in a second dynamical system. With other words, the JRP is the Hadamard product of the RP of the first system and the RP of the second system. JRPs can be computed from more than two systems. The data length of the considered systems has to be the same [22, 24].

$$JR_{i,j}^{x,y}(\varepsilon^x, \varepsilon^y) = \Theta(\varepsilon^x - ||x_i - x_j||) \cdot \Theta(\varepsilon^y - ||y_i - y_j||) \tag{3}$$
$$x_i \in \mathbb{R}^{d1}, \quad y_j \in \mathbb{R}^{d2}, \quad i, j = 1 \ldots n$$

Such joint recurrence plots have the advantage, that the individual measurements can present different observables with different magnitudes or range. They are often used for the detection of phase synchronization [22, 24].

Since this work aims at clustering test drives, which involves pairwise (dis)similarity comparisons of multivariate time series, we propose a combination of joint and cross recurrence plot, namely (JCRP) joint cross recurrence plot. A JCRP shows all those times at which a multivariate state in one dynamical system occurs

simultaneously in a second dynamical system.

$$JCR_{i,j}^{x,y}(\varepsilon^1, \ldots, \varepsilon^k) = \Theta(\varepsilon^1 - ||x_i^1 - y_j^1||) \cdot \ldots \cdot \Theta(\varepsilon^k - ||x_i^k - y_j^k||) \quad (4)$$
$$x_i, y_j \in \mathbb{R}^d, \quad i = 1 \ldots n, \quad j = 1 \ldots m$$

For the creation of a JRCP both trajectories, x and y, need to have the same dimensionality or number of parameters d, but can have different length, n and m. We shall see that JCRPs are very useful, because they enable us to compare two multivariate systems with the same set of observables that can have different magnitudes. In other words, the introduced *JCR* notation allows us to determine an ε-threshold for each individual parameter, which is advantageous for observables with different variance. A toy example for JCRPs is given in the following:

$$x = \begin{cases} \texttt{dfcghGATHERSPEEDlmknhDECELERATEghfkd} \\ \texttt{rsqtpACCELERATORxywzvBRAKEPEDALtvswr} \end{cases}$$
$$y = \begin{cases} \texttt{kdhfSLOWDOWNglbkchdgfGATHERSPEEDnkml} \\ \texttt{tpsBRAKEPEDALzrysxtwvACCELERATORxtwv} \end{cases}$$

Assume two multivariate time series x and y which comprise the speed and accelerator signal recorded during different car drives. Both time series contain multivariate states or rather string sequences that occur in both systems. The corresponding JCRP of x and y, as illustrated in Fig. 1, shows the times at which a multivariate state occurs simultaneously in both systems. Furthermore, the diagonal line structure in Fig. 1 reveals that both trajectories run through a similar region in phase space for a certain time interval. With other words, both systems contain the same multivariate pattern, which represents that the driver hits the 'ACCELERATOR' pedal and the vehicle simultaneously 'GATHERSPEED'. In Sect. 4 we discuss how to interpret single recurrence points and diagonal line structures, and explain how to use them to define a distance measure for time series with certain distortions or invariance.

4 Recurrence Quantification

Recurrence quantification analysis (RQA) is used to quantify the structures observed in recurrence plots [22]. RQA is grounded in theory, but possesses statistical utility in dissecting and diagnosing nonlinear dynamic systems across multiple fields of science [28]. The explicit mathematical definition to distinct features in recurrence plots enables us to analyze signals that are multivariate, nonlinear, non-stationary and noisy.

The global (large-scale) appearance of a RP can give hints on stationarity and regularity, whereas local (small-scale) patterns are related to dynamical properties, such as determinism [28]. Recent studies have shown that determinism, the percentage

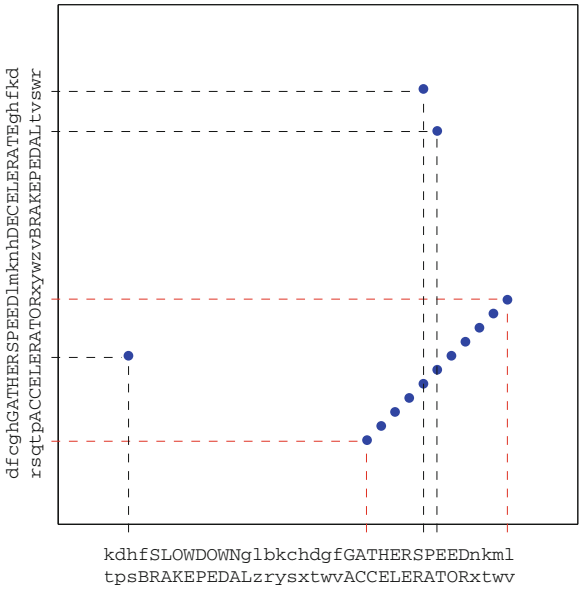

Fig. 1 Joint cross recurrence plot (JCRP) of sample drive x and y from our toy example, with $\varepsilon = 0$

of recurrence points that form lines parallel to the main diagonal, reflects the predictability of a dynamical system [22].

Given a recurrence matrix R with $N \times N$ entries generated by any of the introduced recurrence plot variations, such as our proposed JCRP, we can compute the determinism $DET(\varepsilon, l_{min})$ for a predefined ε-threshold and a minimum diagonal line length l_{min} as followed [22, 24]:

$$DET(\varepsilon, l_{min}) = \frac{\sum_{l=l_{min}}^{N} l \cdot P(\varepsilon, l)}{\sum_{i,j=1}^{N} R_{i,j}(\varepsilon)}$$

$$P(\varepsilon, l) = \sum_{i,j=1}^{N} \left\{ \left(1 - R_{i-1,j-1}(\varepsilon)\right) \cdot \left(1 - R_{i+l,j+l}(\varepsilon)\right) \cdot \prod_{k=0}^{l-1} R_{i+k,j+k}(\varepsilon) \right\} \quad (5)$$

where $P(\varepsilon, l)$ is the histogram of diagonal lines of length l with respect to a certain ε neighborhood.

In general, processes with chaotic behavior cause none or short diagonals, whereas deterministic processes cause relatively long diagonals and less single, isolated recurrence points [22, 29]. In respect to JCRPs, diagonal lines usually occur when the trajectory of two multivariate time series segments is similar according to a certain threshold. Since we aim to measure the similarity between time series that contain segments of similar trajectories at arbitrary positions, which in turn cause diagonal

line structures, we propose to use determinism as a similarity measure. According to the introduced JCRP approach, a high *DET* value indicates high similarity or rather a high percentage of multivariate segments with similar trajectory, whereas a relatively low *DET* value suggests dissimilarity or rather the absence of similar multivariate patterns.

However, data preprocessing like smoothing can introduce spurious line structures in a recurrence plot that cause high determinism value. In this case, further criteria like the directionality of the trajectory should be considered to determine the determinism of a dynamic system, e.g. by using iso-directional and perpendicular RPs [22, 24, 26]. In contrast to traditional recurrence plots, perpendicular recurrence plots (PRPs) consider the dynamical evolution of only the neighborhoods in the perpendicular direction to each phase flow, resulting in plots with lines of the similar width without spreading out in various directions. Removing spurious widths makes it more reasonable to define line-based quantification measures, such as divergence and determinism [30]. Another solution is to estimate the entropy by looking at the distribution of the diagonal lines [26]. The entropy is based on the probability $p(\varepsilon, l)$ that diagonal lines structures with certain length l and similarity ε occur in the recurrence matrix [22, 24].

Recurrence plots (RPs) and corresponding recurrence quantification analysis (RQA) measures have been used to detect transitions and temporal deviations in the dynamics of time series. Since detected variations in RQA measures can easily be misinterpreted, Marwan et al. [25] have proposed to calculate a confidence level to study significant changes. They formulated the hypothesis that the dynamics of a system do not change over times, and therefore the RQA measures obtained by the sliding window technique will be normally distributed. Consequently, if the RQA measures are out of a predefined interquantile range, an observation can be considered significantly. Detecting changes in dynamics by means of RQA measures obtained from a sliding window have been proven to be useful in real-life applications such as comparing traffic flow time series under fine and adverse weather conditions [29].

Since recurrence plot based techniques are still a rather young field in nonlinear time series analysis, systematic research is necessary to define reliable criteria for the selection of parameters, and the estimation of RQA measures [26].

5 Recurrence Plot-Based Distance

According to our formalization of joint cross recurrence (JCR) in Eq. 4 and the denotation of the determinism (DET) in Eq. 5, we can define our RecuRRence Plot-based (RRR) distance measure as follows:

$$RRR(\varepsilon, l_{min}) = 1 - DET(\varepsilon, l_{min}) \tag{6}$$

Since the *DET* value ranges from 0 to 1, depending on the proportion of diagonal line structures found in a *JCR* plot, the *RRR* distance is 0 if the trajectory of both dynamical systems is identical and 1 if there are **no** similar patterns at any position in time.

Although our proposed *RRR* distance measure can be used as a subroutine for various time series mining tasks, this work primarily focuses on clustering. Our aim is to group a set of t unlabeled time series T into k clusters C with centroids Z. In order to evaluate the performance of the time series clustering with respect to our *RRR* distance, we suggest to quantify the number of similar patterns that recur within the established clusters. Therefore, we define the following cluster validation index:

$$E(k) = \frac{1}{t-k} \sum_{z \in \{Z\}} \sum_{c \in \{C_z \setminus z\}} RRR(z, c) \tag{7}$$

According to our problem setting, the more patterns occur jointly when comparing each centroid $z \in \{Z\}$ with all objects $c \in \{C_z \setminus z\}$ of the corresponding cluster, the lower E, the better our clustering, and the more characteristic are the corresponding prototypes.

Furthermore we are going to evaluate the clustering of time series according to the index I [31], whose value is maximized for the optimal number of clusters:

$$I(k) = \left(\frac{1}{k} \cdot \frac{E(1)}{E(k)} \cdot D_k \right)^p \tag{8}$$

The index I is a composition of three factors [31], namely $1/k$, $E(1)/E(k)$, and D_k. The first factor will try to reduce index I as the number of clusters k increases. The second factor consists of the ratio of $E(1)$, which is constant for a given dataset, and $E(k)$, which decreases with increase in k. Consequently, index I increases as $E(k)$ decreases, encouraging more clusters that are compact in nature. Finally, the third factor, D_k (which measures the maximum separation between two clusters over all possible pairs of clusters), will increase with the value of k, but is bounded by the maximum separation between two points in the dataset.

$$D_k = \max_{i,j=1}^{k} ||z_i - z_j|| \tag{9}$$

Thus, the three factors are found to compete with and balance each other critically. The power p is used to control the contrast between the different cluster configurations. Previous work [31] suggests to choose $p = 2$.

The index I has been found to be consistent and reliable, irrespective of the underlying clustering technique and data dimensionality, and furthermore has been shown to outperform the Dunn and David-Bouldin index [31].

6 Evaluation

The goal of our evaluation is to assess how well the RRR distance is suited for: (i) clustering time series that contain similar trajectories at arbitrary positions (in Sect. 6.1), and (ii) identifying prototypical time series that cover as much as possible recurring patterns (in Sect. 6.2).

6.1 Synthetic Data

This controlled experiment aims at visualizing the clustering results of the proposed RRR distance measure compared to the DTW distance.

We generated a labeled dataset, which consists of nine time series from three different categories, called Wave, YoYo and Peak. Each category comprises three time series characterized by multiple occurrence of the same artificial patterns at arbitrary positions. The dataset consists of univariate time series of equal length, as shown in Fig. 2. To visualize the clustering results of the RRR and DTW distance, we applied agglomerative hierarchical clustering with complete linkage on the synthetic dataset.

Figure 3 illustrates the generated hierarchical cluster trees for both examined distance measures on the synthetic time series. The first observation to be made is that RRR perfectly recovers the cluster structure provided by the ground truth, given our knowledge that there are three categories. In contrast, the DTW distance fails and

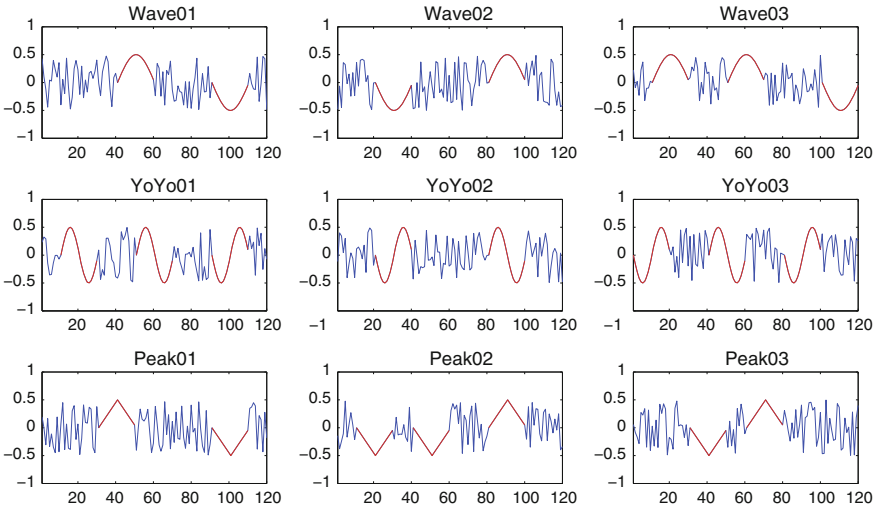

Fig. 2 Univariate synthetic time series with artificially implanted patterns (*red color*) at arbitrary positions, where each time series belongs to one of three groups (Wave, YoYo, and Peak)

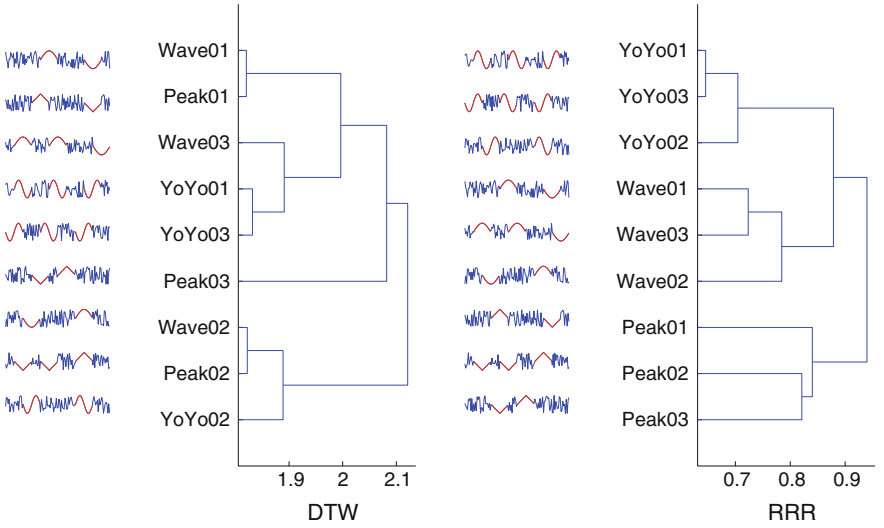

Fig. 3 Cluster tree (dendrogram) of multivariate synthetic time series (introduced in Fig. 2) according to the DTW and RRR distance. The x-axis reveals the distance between the time series being merged and the y-axis illustrates the corresponding name and shape of the time series

assigns time series of different categories to the same cluster at an early stage. The second observation to be made is that RRR is able to recover the ground truth even if a large portion of the time series is noisy. The DTW distance, however, groups time series into the same clusters, if they have globally a similar shape. Therefore, the noisy parts of the time series supersede or superimpose the relevant recurring patterns.

6.2 Real-Life Data

This experiment aims at assessing the time series prototypes identified by the proposed RRR distance measure compared to the DTW distance.

For our evaluation we consider the VW DRIVE dataset, which consists of 124 real-life test drives recorded by one vehicle operated by seven different individuals. Test drives are represented as multivariate time series of varying length and comprise vehicular sensor data of the same observed measurements. Since we aim to identify operations profiles that characterize recurring driving behavior, we exclusively consider accelerator, speed, and revolution measurements, which are more or less directly influenced by the driver. The complete VW DRIVE dataset contains various other measurements, such as airflow and engine temperature, and can be obtained by mailing the first author of this paper.

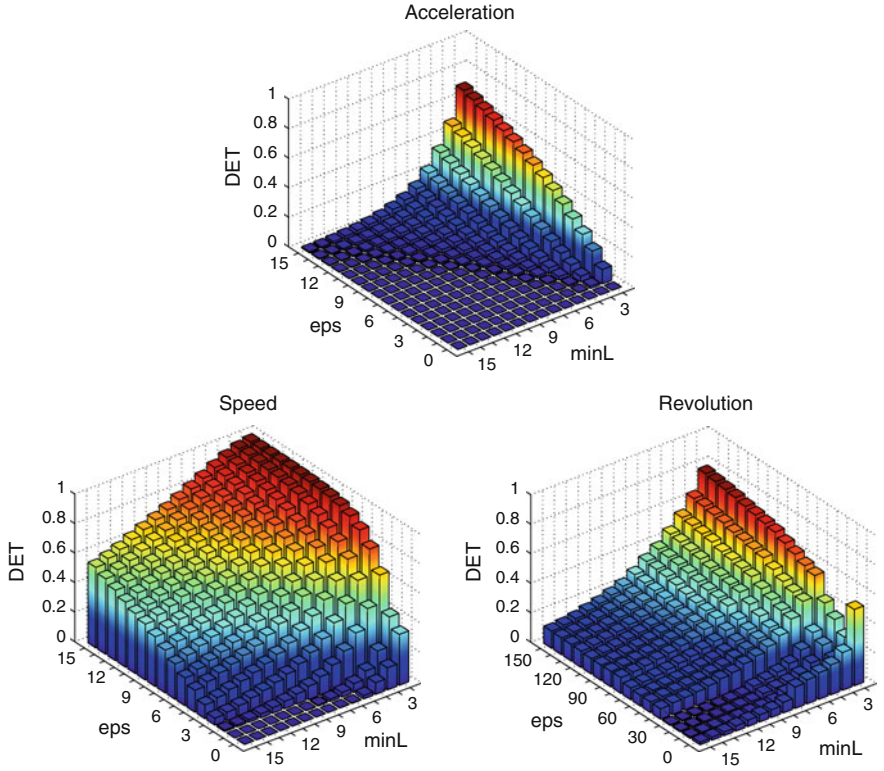

Fig. 4 Determinism (DET) value for changing similarity threshold ε and minimum diagonal *line* length l_{min} for accelerator, speed and revolution signal; based on the cross recurrence plots (CRPs) of 10 randomly selected pairs of tours from our DRIVE dataset. Note that the DET was averaged

To measure the (dis)similarity of the VW DRIVE time series using our proposed RRR distance, we first need to determine the optimal similarity threshold ε and pattern length l_{min} for each of the considered measurements, such that a considerable amount of the recurring patterns is preserved.

Figure 4 shows the determinism value for the accelerator, speed, and revolution signal, in regard to different parameters settings. We can observe that for all considered signals the *DET* value decreases with increasing pattern length l_{min} and decreasing similarity threshold ε. Furthermore, Fig. 4 reveals that the speed signal is highly deterministic, meaning that the same patterns occur frequently, whereas the acceleration and revolution signal are less predictable and show more chaotic behavior.

Since we aim to analyze all signals jointly by means of the proposed joint cross recurrence plot (JCRP) approach, we have to choose a pattern length or rather minimum diagonal line length l_{min} that is suitable for all signals. In general, we are looking for relatively long patterns with high similarity. In other words, we aim to

	(a) Speed		(a) Speed		(b) Acceleration, Speed, and Revolution		(b) Acceleration, Speed, and Revolution		
k	I RRR	E RRR	I DTW	E DTW	I RRR	E RRR	I DTW	E DTW	k
1	-	0.5441	-	0.7041	-	0.7959	-	0.8737	1
2	**1.0000**	**0.5168**	0.1162	0.6794	**1.0000**	**0.7393**	0.7775	0.8622	2
3	0.8778	0.5034	0.6904	0.6602	0.7820	0.7203	0.9088	0.8405	3
4	0.6431	0.4952	0.7548	0.6474	0.5558	0.7064	0.8585	0.8413	4
5	0.4647	0.4924	0.4438	0.6474	0.3883	0.6992	**1.0000**	**0.8407**	5
6	0.3479	0.4909	**1.0000**	**0.6480**	0.2821	0.6934	0.9746	0.8420	6
7	0.2687	0.4888	0.2993	0.6479	0.2141	0.6910	0.2529	0.8452	7
8	0.2151	0.4892	0.1894	0.6493	0.1679	0.6897	0.3100	0.8482	8
9	0.1751	0.4866	0.1189	0.6507	0.1362	0.6855	0.3955	0.8478	9
10	0.1469	0.4862	0.1271	0.6524	0.1131	0.6837	0.2119	0.8534	10
11	0.1254	0.4838	0.3730	0.6530	0.0960	0.6818	0.2624	0.8545	11
12	0.1078	0.4823	0.1184	0.6544	0.0825	0.6784	0.4089	0.8528	12
13	0.0947	0.4817	0.1616	0.6518	0.0717	0.6781	0.2517	0.8576	13
14	0.0838	0.4804	0.2449	0.6531	0.0635	0.6755	0.2453	0.8574	14
15	0.0745	0.4805	0.2988	0.6598	0.0565	0.6746	0.2941	0.8603	15
16	0.0672	0.4803	0.2365	0.6570	0.0508	0.6718	0.2753	0.8588	16
17	0.0609	0.4780	0.1862	0.6507	0.0462	0.6674	0.1106	0.8535	17
18	0.0557	0.4774	0.1761	0.6569	0.0422	0.6687	0.2091	0.8622	18
19	0.0514	0.4751	0.3307	0.6603	0.0387	0.6687	0.1336	0.8596	19
20	0.0473	0.4756	0.0899	0.6579	0.0358	0.6667	0.1036	0.8563	20

Fig. 5 Evaluation of *RRR* and *DTW* distance for clustering **a** univariate and **b** multivariate time series of our DRIVE dataset. We compare the index E for the number of clusters k where the (normalized) index I reaches its maximum. The results are based on 1,000 runs of k-mediods clustering with random initialization

find a parameter setting with preferably large l_{min} and small ε which results in a *DET* value that is above a certain threshold. To preserve the underlying characteristics or rather recurring patterns contained in examined data, at least 20 % of the recurrence points should form diagonal line structures, which corresponds to $DET \geq 0.2$. Based on this criterion we choose $l_{min} = 5$ and $\varepsilon = 14/2/40$ for the accelerator, speed, and revolution signal respectively. Note that the individual signals were not normalized, wherefore the ε-threshold represents the accelerator pedal angle, kilometers per hour, and rotations per minute.

To identify prototypical time series using RRR and DTW distance respectively, we applied k-mediods clustering with random initialization. For evaluation purpose we computed index I and E for a varying number of k prototypes. The results of index I were normalized in a way that the highest value, which indicates the optimal number of clusters, equals one. Since index E is a sum of RRR values (see Eq. 7) and $RRR = 1 - DET$, the lower E, the higher the average *DET* value, and the more recurring (driving behavior) patterns are comprised of the prototypes identified by the respective distance measure.

Figure 5 shows the empirical results for clustering univariate and multivariate time series of the VW DRIVE dataset using RRR and DTW distance respectively. Since the VW DRIVE dataset consists of 'only' 124 test drives recorded by one and the same vehicle, the optimal number of clusters for both RRR and DTW distance is rather small. However, the proposed RRR distance is able to find cluster configurations with lower index E values or rather prototypes with higher amount of recurring patterns than the DTW distance. In case of univariate time series (a), in particular speed measurements, RRR and DTW achieved an index E value of around 0.52 and 0.65 for the optimal number of clusters, which corresponds to a determinism value of 0.48 and 0.35 respectively. In the multivariate case (b), RRR and DTW reached an index E value of around 0.74 and 0.84 for the optimal number of clusters,

which corresponds to determinism value of 0.26 and 0.16 respectively. As might be expected, the results for the univariate time series are better than for the multivariate case, because the search space expands and the probability of recurring patterns decreases with an increasing number of dimensions or measurements respectively. In both cases, however, our RRR distance performs about 10 % better than the compared DTW distance, meaning that the identified prototypes contain 10 % more recurring (driving behavior) patterns.

7 Conclusion

This work is a first attempt to solve time series clustering with nonlinear data analysis and modeling techniques commonly used by theoretical physicists. We adopted recurrence plots (RPs) and recurrence quantification analysis (RQA) to measure the (dis)similarity of multivariate time series that contain segments of similar trajectories at arbitrary positions and in different order.

Strictly speaking, we introduced the concept of joint cross recurrence plots (JCRPs), a multivariate extension of traditional RPs, to visualize and investigate recurring patterns in pairwise compared time series. Furthermore, we defined a recurrence plot-based (RRR) distance measure to cluster time series with order invariance.

The proposed RRR distance was evaluated on both synthetic and real-life time series, and compared with the DTW distance. Our evaluation on synthetic data demonstrates that the RRR distance is able to establish cluster centers that preserve the characteristics of the time series. The results on real-life vehicular data show that, in terms of our cost function, RRR performs about 10 % better than DTW, meaning that the determined prototypes contain 10 % more recurring driving behavior patterns.

Worthwhile future work includes (1) the investigation of RQA measures which quantify recurring patterns with uniform scaling, (2) the application of speed-up techniques for RP computations, and (3) the formalization/analysis of an RP-based distance metric.

Acknowledgments The proposed recurrence plot-based distance measure for clustering multivariate time series was developed in cooperation with the Volkswagen AG, Wolfsburg. Thanks to Bernd Werther and Matthias Pries for their contribution of expert knowledge and their help in recording vehicular sensor data.

References

1. Keogh, E.J., Zhu, Q., Hu, B., Hao, Y., Xi, X., Wei, L., Ratanamahatana, C.A.: The (UCR) time series classification/clustering homepage, www.cs.ucr.edu/eamonn/time_series_data/ (2011)
2. Kumar, M., Patel, N.R., Woo, J.: Clustering seasonality patterns in the presence of errors. In: KDD (2002)

3. Lines, J., Bagnall, A., Caiger-Smith, P., Anderson, S.: Classification of household devices by electricity usage profiles. In: IDEAL, pp. 403–412 (2011)
4. Moeller-Levet, C.S., Klawonn, F., Cho, K.-H., Wolkenhauer, O.: Fuzzy clustering of short time-series and unevenly distributed sampling points. In: IDA, pp. 28–30 (2003)
5. Axel, W., Oliver, L., Dersch, D.R., Leinsinger, G.L., Klaus, H., Benno, P., Dorothee, A.: Cluster analysis of biomedical image time-series. Int. J. Comput. Vision **46**(2), 103–128 (2002)
6. Gustavo, E.A., Batista, P.A., Wang, X., Keogh, E.J.: A complexity-invariant distance measure for time series. In: SDM, pp. 699–710 (2011)
7. Liao, T.W.: Clustering of time series data—a survey. J. Pattern Recognit. 38(11), 1857–1874 (2005)
8. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.J.: Querying and mining of time series data: experimental comparison of representations and distance measures. PVLDB **1**(2), 1542–1552 (2008)
9. Keogh, E.J., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min. Knowl. Discov. **7**(4), 349–371 (2003)
10. Rakthanmanon, T., Campana, B.J.L., Mueen, A., Batista, G., Westover, M.B., Zhu, Q., Zakaria, J., Keogh, E.J.: Searching and mining trillions of time series subsequences under dynamic time warping. In: KDD, pp. 262–270 (2012)
11. Chiu, B.Y.-c., Keogh, E.J., Lonardi, S.: Probabilistic discovery of time series motifs. In: KDD, pp. 493–498 (2003)
12. Lin, J., Keogh, E.J., Lonardi, S., Patel, P.: Finding motifs in time series. In: KDD (2002)
13. Rakthanmanon, T., Keogh, E.J.: Fast-shapelets: a scalable algorithm for discovering time series shapelets. In: SDM (2013)
14. Zakaria, J., Mueen, A., Keogh, E.J.: Clustering time series using unsupervised-shapelets. In: ICDM, pp. 785–794 (2012)
15. Stephan, S., Johannes, J.B., William, De L.E., Sahin, A.: Pattern recognition in multivariate time series: dissertation proposal. In: PIKM, pp. 27–34 (2011)
16. Stephan, S., Julia, G., Andreas, L., Ernesto, De L., Sahin, A.: Pattern recognition and classification for multivariate time series. In: SensorKDD, pp. 34–42 (2011)
17. Spiegel, S., Albayrak, S.: An order-invariant time series distance measure—Position on recent developments in time series analysis. In: KDIR, pp. 264–268 (2012)
18. Bing, H., Chen, Y., Keogh, E.J.: Time series classification under more realistic assumptions. In: SDM (2013)
19. Keogh, E.J., Lin, J.: Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowl. Inf. Syst. **8**(2), 154–177 (2005)
20. Keogh, E.J., Lin, J., Fu, A.W.-C.: HOT SAX: efficiently finding the most unusual time series subsequence. In: ICDM, pp. 226–233 (2005)
21. Marwan, N.: Encounters with neighbours: current developments of concepts based on recurrence plots and their applications. University of Potsdam (2003)
22. Marwan, N., Romano, M., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. Phys. Rep. **438**(5–6), 237–329 (2007)
23. Marwan, N.: A historical review of recurrence plots. Eur. Phys. J. Special Topics **164**(1), 3–12 (2008)
24. Marwan, N., Romano, M., Thiel, M.: Recurrence plots and cross recurrence plots. www.recurrence-plot.tk
25. Marwan, N., Schinkel, S., Kurths, J.: Recurrence plots 25 years later—gaining confidence in dynamical transitions. Europhys. Lett., **101**(2) (2013)
26. Marwan, N.: How to avoid potential pitfalls in recurrence plot based data analysis. I. J. Bifurcat. Chaos **21**(4), 1003–1017 (2011)
27. Schultz, A.P., Zou, Y., Marwan, N., Turvey, M.T.: Local minima-based recurrence plots for continuous dynamical systems. I. J. Bifurcat. Chaos **21**(4), 1065–1075 (2011)
28. Webber, C.L., Marwan, N., Facchini, A., Giuliani, A.: Simpler methods do it better: success of recurrence quantification analysis as a general purpose data analysis tool. Phys. Lett. A **373**(41), 3753–3756 (2009)

29. Vlahogianni, E.I., Karlaftis, M.G.: Comparing traffic flow time-series under fine and adverse weather conditions using recurrence-based complexity measures. Nonlinear Dyn. **69**(4), 1949–1963 (2012)
30. Choi, J.M., Bae, B.H., Kim, S.Y.: Divergence in perpendicular recurrence plot; quantification of dynamical divergence from short chaotic time series. Phys. Lett. A **263**(4–6), 299–306 (1999)
31. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. 24(12), 1650–1654 (2002)