Clustering von Recurrence Plots

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science (M. Sc.)

eingereicht von: Carl Witt
geboren am: 28.05.1988
geboren in: Berlin

Gutachter: Prof. Dr. Ulf Leser,
Humboldt-Universität zu Berlin
Dr. Norbert Marwan,
Potsdam-Institut für Klimafolgenforschung

eingereicht am: ..............................
Contents

1. Introduction 5
2. Foundations 7
   2.1. Dynamical Systems 7
   2.1.1. Phase Space Trajectories and the Butterfly Effect 7
   2.1.2. Time Delay Embedding 8
   2.2. Recurrence Plots 10
   2.3. Recurrence Quantification Analysis 10
   2.4. Clustering 14
      2.4.1. Clustering Validation 14
      2.4.2. Clustering Algorithms 16
3. Related Work 18
   3.1. Similarity Measures for RPs 18
   3.2. Clustering RQA Measures 20
4. Similarity Measures for Recurrence Plots 21
   4.1. Evaluation Framework 22
   4.2. Preliminary experiments 27
   4.3. Geometric and Temporal Abstractions in RQA Measures 29
5. Clustering Recurrence Plots 31
   5.1. Artificial Data Set 31
      5.1.1. Computation of Trajectories 31
      5.1.2. Selection of Trajectory Lengths 33
   5.2. RQA Measures of the Artificial Data Set 34
      5.2.1. Computation of RPs and RQA Measures 34
      5.2.2. Qualitative Results 35
   5.3. Clustering RQA Measures 37
      5.3.1. Experimental Design 37
      5.3.2. Quantitative Results 40
6. Influence of Noise and Embedding 42
   6.1. Artificial Noise Model 42
      6.1.1. Alternative Models 43
      6.1.2. Equal Impairment of Classes 44
      6.1.3. Dependence on Dimensionality 45
      6.1.4. Clustering Quality under Observational Noise 47
   6.2. Influence of Embedding 49
      6.2.1. Estimation of Embedding Parameters 50
      6.2.2. Clustering Quality for Univariate Trajectories 52
7. Extending Recurrence Quantification Analysis
   7.1. Lines of Definite and Indefinite Length ...................... 53
   7.2. Line Types and Descriptors .................................. 55
   7.3. Conditional Recurrence Time .................................. 56

8. Conclusions and Future Work .................................. 60

Appendices ......................................................... 68
   A. Artificial Noise Model ........................................ 68
   B. Experimental Data ............................................. 68
   C. Data Figures ................................................... 69

List of Figures
   2.1. Chaos in the Logistic Map .................................... 7
   2.2. A trajectory of the Lorenz System .......................... 8
   2.3. A reconstructed trajectory of the Lorenz System .......... 9
   2.4. Recurrence Plots of the trajectories in Figure 2.1 ........ 11
   2.5. Recurrence Plots of the trajectories in Figures 2.2 and 2.3 . 11
   2.6. Diagonal, vertical and white vertical line structures in a Recurrence Plot 13
   4.1. Overview of the classes of Recurrence Plots in the artificial data set .... 23
   4.2. Expert assessments of proximity ................................ 25
   4.3. Hamming and Spatiogram Distance .......................... 28
   4.4. Euclidean distance of RQA vectors .......................... 29
   4.5. Geometric abstraction in Recurrence Plots ................. 30
   5.1. Convergence of $L_{\text{entr}}$, $TT$, $L_{\text{max}}$, and $W_{\text{max}}$ 34
   5.2. Scatter plots of $L_{\text{entr}}$, $V_{\text{entr}}$, and $W_{\text{entr}}$ 35
   5.3. Principal Components of the Convergent feature set ........ 36
   5.4. Full distance matrix for RQA vectors ....................... 37
   5.5. External clustering quality for the correct number of clusters .... 41
   5.6. Internal clustering quality ................................... 42
   6.1. Persistent clustering problem: RQA measures under extreme noise .... 46
   6.2. Impact of the noise model on Recurrence Rates ............ 47
   6.3. Distance matrix of RQA vectors under 80% noise ............ 48
   6.4. Principal components under 80% noise ...................... 48
   6.5. External clustering quality under noise ..................... 49
   6.6. Estimation of embedding parameters ........................ 51
   6.7. External clustering quality under noise and embedding .... 52
   7.1. Testcase for comparing RQA implementations ................ 54
   7.2. Simple example for Conditional Recurrence Time ............ 56
   7.3. Conditional Recurrence Time under noise .................... 57
7.4. Conditional Recurrence Time for chaotic systems ........................................ 58
C.1. Excerpts of selected Recurrence Plots of the artificial data set. .................. 72
C.2. Convergence of RQA measures ............................................................... 73
C.3. RQA measures of the artificial data set under 10% noise ......................... 74
C.4. Trajectories of the Lorenz System under noise .......................................... 75
C.5. RPs of the Lorenz System under noise ..................................................... 76
C.6. RPs for selected classes under a noise ratio of 80% ................................. 77
C.7. Impact of noise on RQA measures .......................................................... 78
C.8. Evolution of clusters under noise ............................................................. 79
C.9. External clustering quality under all conditions ......................................... 80
C.10. External clustering quality for short trajectories ....................................... 81
C.11. White vertical line length histograms of selected trajectories .................... 82
C.12. Selected CRT statistics of the artificial data set ....................................... 83
C.13. Distance matrix of all CRT statistics ..................................................... 84
C.14. Scatter plot of CRT Entropy and Local Maxima ....................................... 84
1. Introduction

Chaotic behavior is everywhere. It arises for instance in Medicine, Financial Markets, and Physics. Typically, it is a result of multiple interacting entities. These can be neurons in the human brain, people trading on a stock market, and celestial bodies. The result is complex behavior that can usually not be predicted in the long term. To study such systems, their properties can be observed over time. However, to analyze chaotic time series, special techniques are required. Recurrence Analysis uses patterns of recurrence to characterize time series and abstract from their chaotic elements. For instance, two EEG signals will never be identical, but their differences are not just observational noise. They are much more complex in nature. Still, to relate the data to states of the brain, such as sleep and wake, epileptic and healthy, an abstraction from these differences is necessary. In general, it has been shown that recurrent patterns can be used to capture the commonalities between chaotic time series. For instance, EEG signals during wake and sleep phases can be discerned by comparing their patterns of recurrence [FKC+14].

Recurrence is a basic concept that can be used for prediction and classification in various contexts. For instance, a look into the past suffices to make an educated guess about the amount of rainfall in Berlin during the next summer. Likewise, significant changes in these patterns might indicate changes in the underlying system – the climate system. The Recurrence Plot (RP) is a technique to visualize recurrence patterns, and Recurrence Quantification Analysis (RQA) provides quantitative measures of the patterns in an RP. In this thesis, the combination of Recurrence Analysis and clustering techniques was studied. The task is to cluster time series according to their RPs, that is, their recurrence patterns. The motivation is to exploit the power of Recurrence Analysis within explorative data analysis. A clustered data set allows for systematic and selective exploration by providing an overview and allowing to focus on interesting groups. In contrast, exploring a large, unordered list of thousands of RPs is not feasible.

The problem is a lack of systematic studies. In an effort to close this gap, an evaluation framework was created in cooperation with experts from the Potsdam Institute for Climate Impact Research (PIK). Within this framework, similarity measures, feature sets and clustering algorithms have been evaluated. It was found that vectors of RQA measures well describe a set of prototypical dynamics, even under observational noise and time delay embedding. Clustering qualities above 80% were achieved under a wide range of experimental conditions. Specifically, K-Means clustering was found to yield good results in combination with dimensionality reduction through Principal Component Analysis. Finally, Conditional Recurrence Time, an approach to quantify structures in RPs based on consecutive lines, was proposed. It was noted to be an alternative characterization of the patterns in RPs that might serve as an extension to traditional RQA measures.

A successful clustering of Recurrence Plots identifies typical behaviors in a time series data set. There are various applications for this task. Vlahogianni et al. [VKG08] have identified typical patterns in urban car traffic by means of clustering and Recurrence Analysis. Ishii et al. [IDM12] employ Recurrence Analysis for load prediction in dis-
tributed computing environments. Rodrigues et al. [RG09] propose a clustering method to predict the electricity load of a power plant. The Recurrence Quantification Analysis (RQA) method is not even limited to time series. Zbilut et al. [ZGWC98] have shown that it is useful to apply RQA measures to protein structures. An application is the discrimination of mutant protein structures [CPG+06]. In general, scenarios that produce large numbers of RPs are candidates for clustering Recurrence Plots. Three such scenarios were identified.

**Spatiotemporal Data** In [LB14], remote sensing data is used to investigate global photosynthetic activity over a period of 14 years. The data is available for a grid of positions on the surface of the earth. At each grid point, the photosynthetic activity is modeled as a time series. In general, the exploration of spatiotemporal data is a challenging task. The approach of Lange et al. is to eliminate the temporal dimension first, by summarizing the time series at each location by features of its Recurrence Plot. The subsequent investigation of the features on a world map revealed connected regions of coherent photosynthetic activity. The paper concludes with the statement that “The description of the visual impressions from the world map could be lifted up to a quantitative level through cluster analysis.” The use of clustering to develop climate classification systems is also supported by Zscheischler et al. [ZMH12].

**Time Dependent Recurrence Analysis** According to [MW14], “One of the most useful applications of recurrence quantifications is to examine long time series of data using a small moving window traversing the data.” Each of these windows gives an Recurrence Plot (RP), and comparing the RPs can be used to detect changes in a system’s behavior. This way, climate changes could be detected by analyzing recurrence patterns of ancient dust deposit data [MDZ+09]. A common method is to compute confidence intervals for RQA measures. If a measure changes significantly, a change in the system’s behavior is hypothesized. This method has been recently extended to multivariate confidence intervals [YC14]. Clustering the RPs could be an alternative approach to the problem. This approach lacks statistical significance, but could identify the system states instead of only detecting the transition times.

**Multiscale Recurrence Analysis** Another method that produces large numbers of RPs is Multiscale Recurrence Analysis. In [CY12], heart rate variability time series are recursively decomposed into high and low frequency parts, using Wavelet Packet Decomposition. The decomposed time series are summarized by their RQA measures. An advantage of the method is that mixed processes are separated, which results in clearer pictures in the individual RPs. Furthermore, processing more, but shorter time series reduces the impact of the quadratic running time of RQA. The resulting features are used to train a classifier to detect the conditions of Congestive Heart Failure. Yuan et al. [YYC14] use Wavelet decomposition in conjunction with RQA measures to detect anomalous traffic in a computer network.
2. Foundations

2.1. Dynamical Systems

2.1.1. Phase Space Trajectories and the Butterfly Effect

A dynamical system consists of a phase space and an evolution rule. Each possible state of the dynamical system corresponds to a point in phase space. A wide range of real-world processes can be modeled using dynamical systems, including epidemiology, fluid dynamics, market prices, and celestial mechanics. Thus, the state of the system can refer to many things, such as the size of a biological population, the price of some commodity, and the temperatures at different locations on the surface of the earth. The evolution rule is a deterministic law that dictates the future states of the system depending on the current state of the system. In this respect, recursion\(^1\) is an inherent property of dynamical systems. The evolution of the system’s state over time is called a phase space trajectory.

The following section exemplifies these concepts using the Logistic Map, a dynamical system from population biology. The Logistic Map describes the size of the next generation of some population as a function of the current population. It models reproduction with a growth rate \(\mu\) and starvation with a death rate dependent on the remaining available resources. The system’s current state is described by the size of the population relative to the maximum possible population size. The evolution rule is:

\[
x_{n+1} = \mu x_n (1 - x_n)
\]

Where \(\mu\) is the growth rate. The phase space trajectory is computed by recursively applying the evolution rule to an initial population \(x_0\). It turns out that the trajectory strongly depends on the choice of \(x_0\). In Figure 2.1, the red trajectory results from \(x_0 = 0.2\), and the black trajectory results from \(x_0 = 0.20000001\). After a few steps, the trajectories diverge strongly, which is often referred to as Butterfly Effect. Consequently, it is not possible to make long term predictions without exactly knowing the state of the system. As the figure shows, a tiny measurement error of the state of the system leads to strongly different predictions in the long term.

\(^1\)An inspiring excursus that relates recursion to elementary concepts like meaning, communication and intelligence, is Hofstadter’s famous monograph Gödel, Escher, Bach: an Eternal Golden Braid [Hof00].
Another interesting aspect of the Logistic Map is that the system’s behavior also depends on the growth rate parameter $\mu$. In certain parameter ranges, the system has a simple mathematical description, while in others it exposes chaotic behavior similar to that in Figure 2.1. Note that although the behavior is complex, it exhibits recurrent patterns, such as a drastic decrease in population followed by slow regeneration phases, or short periods of near constant population.

Another prototypical dynamical system is the Lorenz System, which was developed by Edward Lorenz as a simple model for atmospheric convection. It was later found to have applications in laser physics and chemistry as well. While the basic principles are the same as for the Logistic Map, this system evolves in continuous time and three-dimensional space. This means that recursion proceeds in infinitesimal time steps. The evolution rule for the standard parameters $\sigma = 10, \beta = 8/3, \rho = 28$ is:

\[
\begin{align*}
\frac{dx}{dt} &= 10(y - x) \\
\frac{dy}{dt} &= x(28 - z) - y \\
\frac{dz}{dt} &= xy - \frac{8}{3}z
\end{align*}
\]

Recursively applying the evolution rule to some initial condition $(x_0, y_0, z_0)$ is not as straightforward as for the Logistic Map. The infinitesimal time steps $dt$ need to be approximated with small, finite time steps $\Delta t$, i.e. applying some numerical integration scheme. Figure 2.2 shows a phase space trajectory of the Lorenz System. Although the trajectory looks well structured, it is also sensitive to the initial state. Depending on the chosen $(x_0, y_0, z_0)$, the traversal order of the two scrolls will change strongly. Thus, long term predictions are also not possible without knowing exactly the current state of the system. Note that the Lorenz System exposes even stronger recurrent patterns, since there are basically only two patterns that alternate in chaotic order. These two patterns recur infinitely and independently of the chosen numerical integration routine [Spa82].

### 2.1.2. Time Delay Embedding

In real world scenarios, often only a single dimension of a phase space trajectory can be measured. Under certain conditions, a phase space trajectory that has equivalent properties to the original phase space trajectory can be reconstructed. A simple method
for this task is Time Delay Embedding. The method requires the choice of an embedding dimension $m$ and a time delay $\tau$. The $i$-th observation $x'_i$ of the reconstructed trajectory is formed by $m$ consecutive observations of the original time series $x$, spaced by $\tau$ time steps: $x'_i = \sum_{j=0}^{m-1} x_{i+j\tau} e_j$, where $e_j$ is the $j$-th basis vector of the reconstruction space, usually $\mathbb{R}^m$, cf. [MCRTK07, p. 246]. The basic intuition behind this approach is that interrelation of equations causes a variable to contain some information about the other variables. In the Lorenz System, the evolution of the $y$ component depends on the $x$ and $z$ components. Time Delay Embedding is able to restore this information to a certain degree. However, the theoretical foundations, Takens’ delay embedding theorem [Tak81], are beyond the scope of this thesis. For practical application, it suffices to know that this is a standard method and both original and reconstructed trajectory have similar properties from a Recurrence Analysis point of view. Figure 2.3 shows a three-dimensional phase space trajectory of the Lorenz System, reconstructed from the $x$-component of the trajectory shown in Figure 2.2. The embedding dimension and delay have been set to 3.

In practice, embedding dimension and delay are not known a priori and several methods have been developed to estimate them. In this thesis, the first\(^2\) local minimum of the Mutual Information is used to estimate $\tau$, and the False Nearest Neighbor algorithm is used to estimate $m$. The basic idea of the Mutual Information algorithm is to compare a time series $x_i$ to a lagged version $x_{i+\tau}$ of itself. A small Mutual Information of the time series and its lagged version indicates a certain type of independence, which is related to a good choice of the embedding delay $\tau$ [MW14]. The basic idea of the False Nearest Neighbor algorithm is to compare the trajectory embedded in $k$ dimensions to the trajectory embedded in $k + 1$ dimensions. A too low embedding generation can bring far away points close together. Two points are considered false nearest neighbors if they are neighbors in $k$ dimensions but not in $k + 1$ dimensions. Usually the smallest $k$ is selected as embedding dimension for which no more false nearest neighbors are found. See [MW14, HKS99] and Section 6.2.1 for further details.

\(^2\)In [MW14], the smallest local minimum is proposed, the choice of the first local minimum is justified in Section 6.2.1.
2.2. Recurrence Plots

Recurrence Plots visualize patterns of recurrence in a phase space trajectory. Recurrence is a fundamental property of dynamical systems [MW14], and visualizing patterns of recurrence provides insights into a system’s behavior. Furthermore, RPs perform a dimensionality reduction, since they visualize phase space trajectories in two dimensions, regardless of the dimensionality of the phase space.

Basically, recurrence is defined as spatial closeness in phase space, here measured as $L_2$ (euclidean) distance, as in the original definition of RPs [MW14]. By measuring the pairwise distances between the points on the trajectory, a distance matrix is obtained. Applying a threshold $\varepsilon$ to the distance matrix gives the binary recurrence matrix $R_{i,j}$. If a metric is used to compare phase space points, the recurrence matrix is symmetric and includes all points on the main diagonal.

$$R_{i,j} = \begin{cases} 
1 & \text{if } ||x_i - x_j|| \leq \varepsilon \\
0 & \text{otherwise}
\end{cases} \quad (1)$$

Where $x_i$ and $x_j$ are the $i$-th and $j$-th point on the phase space trajectory. A pair of phase space points $(x_i, x_j)$ such that $R_{i,j} = 1$ is called a recurrence point. The magnitude of the threshold defines how precise the self-similarity must be, where $\varepsilon = 0$ defines recurrence as an exact return to a previous location. This is too restrictive in practice, hence, the parameter is usually chosen according to some heuristic.

Visualization of the recurrence matrix is straightforward. The convention is to visualize recurrence points in black and the others in white. On account of clarity, the terms recurrence point and non-recurrence point should be preferred. For the sake of brevity, the terms black point and white point are used occasionally. Figure 2.4 shows the RPs for the two diverging trajectories of the Logistic Map. Figure 2.5 shows the RPs of the original and the reconstructed trajectory of the Lorenz System. Note that the different shape of the reconstructed trajectory in Figure 2.3 is abstracted from in the RP.

In several application papers, e.g. [SDSB13, SSB14, JDLP11, KT12], the distance matrix is used without applying a threshold. It is claimed that thresholding causes a loss of information. In the context of Recurrence Analysis, there are two arguments in favor of thresholding. Firstly, it has been shown that the recurrence matrix contains enough information to estimate properties of the underlying system, and even reconstruct the phase space trajectory to a certain degree [TRK04]. Secondly, thresholding results in discrete and separated line structures which can be easily quantified and reordered in time, which is useful to describe the structures in an RPs.

2.3. Recurrence Quantification Analysis

Recurrence Quantification Analysis supplements visual inspection of RPs with measures that quantify the structures in an RP. Although the definitions of almost all measures are simple, they are related to complex concepts from theoretical physics and chaos theory. A gentle introduction to the topic was presented by Marwan and Webber [MW14]. Even
Figure 2.4: The RPs of the two Logistic Map trajectories with slightly different initial conditions, cf. Figure 2.1.

Figure 2.5: The RPs of the original and reconstructed trajectories of the Lorenz system, cf. Figures 2.2 and 2.3.
without delving into their theoretical foundations, RQA measures are instructive with regard to which structures in RPs are relevant in the context of Recurrence Analysis. Most RQA measures are derived from diagonal, vertical and white vertical line structures in the RP. These correspond to three basic properties of recurrence. Diagonal lines occur when the phase space trajectory evolves in a similar way as before. The more recurrence points are part of diagonal lines, the more predictable is the system. Vertical lines indicate periods of stagnation, where the trajectory remains in the same region of the phase space. White vertical lines correspond to the time spans that elapse between recurrences of the trajectory to a given region in phase space. These aspects of recurrence are mainly quantified by line length histograms. The number of lines of length $l$ in the recurrence matrix $R$ is referred to as $P(R, l)$. Where the recurrence matrix is clear from the context, the shorthand notation $P(l)$ will be used.

Since the main diagonal is part of every RP, it is removed for quantification of the diagonal lines. Otherwise, the average and maximum line length measures would be biased or even useless. Because larger thresholds can cause the line along the main diagonal to become thicker, usually an entire corridor around the main diagonal is excluded. The width $w$ of this corridor is termed Theiler corrector. When applying a Theiler Corrector, only recurrence points $(i, j)$ are considered which satisfy $|i − j| ≥ w$.

Formally, a line is a sequence of adjacent recurrence matrix cells of value $v$, which is preceded and succeeded by a cell of value $1 − v$. Refer to Figure 2.6 for an illustration. In practice, lines of recurrence points that are incident to the borders of the matrix are also counted, although they lack a non-recurrence point that terminates them. This issue is discussed in Section 7.1. In the style of [MCRTK07], the line length histograms of the diagonal, vertical, and white vertical lines, $P_L$, $P_V$, and $P_W$, are defined as follows.

$$P_L(l) = \sum_{i,j=1,|i−j|≥w}^N (1 − R_{i−1,j−1})(1 − R_{i+l,j+l}) \prod_{k=0}^{l−1} R_{i+k,j+k}$$

$$P_V(l) = \sum_{i,j=1}^N (1 − R_{i,j−1})(1 − R_{i,j+l}) \prod_{k=0}^{l−1} R_{i,j+k}$$

$$P_W(l) = \sum_{i,j=1}^N R_{i,j−1}R_{i,j+l} \prod_{k=0}^{l−1} (1 − R_{i,j+k})$$

The line length histograms are commonly summarized by five descriptors: percentage of points on lines longer than a threshold $l_{min}$, average line length, maximum line length, reciprocal maximum line length, and Shannon Entropy. Table 1 summarizes the names and equations for these descriptors, depending on whether they refer to the distribution of diagonal, vertical or white vertical lines. $N$ denotes the length of the trajectory. The normalized histogram $p(l)$ is defined as $p(l) = P(l)/\sum_{l=1}^{N} P(l)$. The names DET, LAM, TT, and DIV abbreviate the terms Determinism, Laminarity, Trapping Time, and Divergence. The minimum line length $l_{min}$ serves as a filter to eliminate spurious recurrence structures. Besides the measures in Table 1, some RQA measures are not directly

\[3\] Of course, the converse does not hold.
Figure 2.6: Diagonal, vertical and white vertical line structures in a Recurrence Plot. For diagonal and vertical lines, the lines of length 1 were omitted.

<table>
<thead>
<tr>
<th>Description</th>
<th>Formula</th>
<th>Measure Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter Ratio</td>
<td>$\frac{\sum_{l=l_{\text{min}}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)}$</td>
<td>DET</td>
</tr>
<tr>
<td></td>
<td>$\frac{\sum_{l=l_{\text{min}}}^{N} lP(l)}{\sum_{l=l_{\text{min}}}^{N} P(l)}$</td>
<td>L</td>
</tr>
<tr>
<td>Average length</td>
<td>$\frac{\sum_{l=l_{\text{min}}}^{N} lP(l)}{\sum_{l=l_{\text{min}}}^{N} P(l)}$</td>
<td>L</td>
</tr>
<tr>
<td>Maximum length</td>
<td>$\max({l \mid P(l) &gt; 0})$</td>
<td>$L_{\text{max}}$</td>
</tr>
<tr>
<td>1/Maximum length</td>
<td>see above</td>
<td>DIV</td>
</tr>
<tr>
<td>Entropy</td>
<td>$- \sum_{l=l_{\text{min}}}^{N} p(l) \ln p(l)$</td>
<td>$L_{\text{entr}}$</td>
</tr>
</tbody>
</table>

Table 1: Line length histogram based RQA measures, formulas according to [MW14].
related to a single line length histogram. The Recurrence Rate \( RR = \sum_{i=1}^{N} \sum_{j=1}^{N} R_{i,j}/N^2 \) is defined as the fraction of recurrence points compared to the size of the recurrence matrix. Two derived measures are the ratio between DET and RR (RATIO), and the ratio between LAM and DET (LAM/DET). Finally, there are also measures that are not based on the line length histograms, such as TREND, which is designed to quantify non-stationary behavior. It is excluded, since non-stationary behavior is not evaluated in this thesis. In summary, 16 RQA measures are considered in this thesis, namely, the 13 measures from Table 1, RR, RATIO and LAM/DET.

Using a naïve algorithm, computing RQA measures requires \( O(N^2) \) time and space, where \( N \) denotes the length of the phase space trajectory. While it is easy to come up with a linear-space algorithm for computing the line length distributions, the quadratic time complexity is hard to overcome. This poses a challenge for processing long time series. In theory, spatial indexing can be used to reduce the runtime to \( O(n \log n) \), but these solutions degrade in practical scenarios\(^4\). Rawald et al. have massively parallelized the computation of RQA measures using graphical processing units [RSMD14, RSML15]. Another approach is to approximate the RQA measures based on phase space discretization [SSMA15]. In some scenarios, a time series can be decomposed into shorter time series using Wavelet Packet Decomposition [CY12], which reduces the impact of the quadratic running time. In this thesis, a custom Java implementation was used, which showed good performance, compared to the Command Line Recurrence Plots implementation [Marb].

### 2.4. Clustering

Clusters in data are commonly defined as groups of objects that are more similar to each other than to other objects. In the style of [ES13], let a clustering \( f \) be defined as a function \( f : O \rightarrow C \) where \( O = \{o_1, \ldots, o_N\} \) is the set of \( N \) objects and \( C = \{C_1, \ldots, C_k\} \) is the set of \( k \) clusters, where \( \bigcup_i C_i = O \) and \( C_i \cap C_j = \emptyset, i \neq j \). When writing \( C_i \) or \( C_j \), sensible indices \( 1 \leq i, j \leq k \) will be assumed.

#### 2.4.1. Clustering Validation

Two common approaches to assessing the quality of a clustering are external and internal validation measures. The former are based on a reference assignment of objects to clusters, which is often assumed to be the correct assignment. Some difficulties of using class labels in clustering validation have been pointed out in [FGK+10]. A more comprehensive introduction to clustering validation can be found in [Mil96].

**External Validation** The Adjusted Rand Index [HA85] is a measure of agreement between two clusterings \( f \) and \( g \) of the same set of objects with \( k \) and \( k' \) clusters, respectively. It is based on assessing all pairs \((o_i, o_j)\) of objects. Let \( a = |\{(o_i, o_j) \mid f(o_i) = f(o_j) \land g(o_i) = g(o_j)\}| \) be the number of pairs of objects that belong to the same cluster

\(^4\)Private conversation with Tobias Rawald, GeoForschungsZentrum Potsdam
in both clusterings. Let \( b = |\{(o_i, o_j) \mid f(o_i) \neq f(o_j) \land g(o_i) \neq g(o_j)\}| \) be the number of pairs of objects that do not belong to the same cluster in both clusterings. The Rand Index RI is defined as \((a + b)/(n(n-1)/2)\), that is, the fraction of agreements relative to all pairs. Since even a random clustering is expected to cluster some objects together correctly, a correction for chance is often used. The Adjusted Rand Index ARI is defined as \((RI - E[RI])/(MaxRI - E[RI])\). The expressions \( E[RI] \) and \( MaxRI \) are based on the so-called contingency table \( n_{ij} = C_i \cap C_j \), where \( a_i = \sum_j n_{ij} \) and \( b_j = \sum_i n_{ij} \) are the row and column sums, respectively.

\[
\text{ARI}(f, g) = \frac{\sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} / \binom{n}{2}}{\frac{1}{2} \left[ \sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} / \binom{n}{2}}
\]

The Adjusted Mutual Information measures the agreement between two clusterings by the amount of information that one clustering contains about the other. Using the contingency table notation as in the definition of the ARI, the Mutual Information is defined as

\[
\text{MI}(f, g) = \sum_i \sum_j \frac{n_{ij}}{N} \log \frac{n_{ij} N^2}{a_i b_j N}
\]

The Mutual Information can be chance corrected using the same method as for the RI, as \( \text{AMI} = (\text{MI} - E[\text{MI}]) / (\max(H(f), H(g)) - E[\text{MI}]) \), where \( H(f) \) is the Shannon entropy of the clustering \( f \), i.e. \( -\sum (|C_i|/N) \log(|C_i|/N) \). The expression for the expected Mutual Information is pretty verbose and can be looked up in [VEB09, VEB10].

**Internal Validation** Let \( d(A, B) = \{\|a - b\|_2 \mid (a, b) \in A \times B\} \) be the pairwise euclidean distances between the objects in sets \( A \) and \( B \). For convenience, if an argument \( a \) of \( d \) is an object, replace it with the one-element set \( \{a\} \). Let \( \text{avg}(a) \) be the arithmetic mean of the elements in the set \( A \). Let \( \overline{d}(A, B) = \text{avg}(d(A, B)) \) be the average pairwise distance. Let \( \overline{C}_i = \text{avg}(C_i) \) denote the centroid of the cluster \( C_i \).

A very internal validation measure is the sum of distances between all objects and their cluster centers. This is referred to as Inertia \( I(f) \) of the clustering \( f \).

\[
I(f) = \sum_{C_i \in C} \sum_{x \in d(C_i, \overline{C}_i)} x
\]

Since the distances are non-negative, \( I(f) \geq 0 \). Small Inertia indicates a high cohesion of the clusters, but a large Inertia does not necessarily indicate a bad clustering, because it does not account for possibly large separation. Furthermore, Inertia can be trivially optimized by choosing \( k = N \). Therefore, applying this measure is only sensible in conjunction with a fixed \( k \).
The Silhouette Score [Rou87] fixes this defect. The Silhouette \( s(o) \) of an object \( o \) is based on how well \( a(o) \) it fits its assigned cluster, compared to the best fit \( b(o) \) to any other cluster. The quantity \( b(o) \) incorporates the notion of separation into the validation.

\[
s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))}, \quad \text{where } a(o) = \bar{d}(o, f(o)), \quad \text{and } b(o) = \min_{C_i \neq f(o)} \bar{d}(o, C_i)
\]

The Silhouette Score of the clustering is defined as \( \text{avg}\{s(o) \mid o \in O\} \). Since the distance to the other cluster \( b(o) \) is ideally large and the distance to the own cluster \( a(o) \) is ideally small, a larger Silhouette Score is better. Since \( a(o) \) and \( b(o) \) are non-negative, the Silhouette Score is in range \([-1, 1]\).

The Dunn Index \( DI \) is based on the ratio of the smallest separation between two clusters and the largest diameter of any cluster. Let \( \delta(C_i, C_j) = \min d(C_i, C_j) \) denote the separation of the clusters \( C_i \) and \( C_j \). Let \( \Delta(C_i) \) denote the inverse cohesion of Cluster \( C_i \). Not that variations of the Dunn index exist, which differ in the definition of \( \Delta \). One variant is \( \Delta(C_i) = \max d(C_i, C_i) \). In this thesis, \( \Delta(C_i) = \bar{d}(C_i, \overline{C_i}) \), the mean distance between objects and their centroid. Note that both variants can be zero, leading to a possibly undefined index. However, from an internal validation point of view, the case that all objects have a distance of 0 to their centroid should indicate a perfect clustering.

\[
DI(f) = \frac{\min_{C_i \neq C_j} \delta(C_i, C_j)}{\max \Delta(C_i)}
\]

Since the separation should be large and the diameters small, a larger Dunn Index corresponds to a better clustering, but the index is not bounded.

The Davies-Bouldin-Index \( DB \) is based on the ratio between cohesion and separation. Cohesion of a cluster is measured as the mean distance between objects and their centroid. Separation is measured as distance between centroids. The DB Index of the clustering is the average of this ratio for each cluster \( C_i \) and the cluster which is separated worst from \( C_i \).

\[
DB(f) = \frac{1}{k} \sum_{C_i \in C} \max_{C_j \neq C_i} \frac{\bar{d}(C_i, \overline{C_j}) + \bar{d}(C_j, \overline{C_j})}{\|C_i - \overline{C_j}\|_2}
\]

Ideally, the ratio between cohesion and separation is small, thus a smaller value indicates better clusterings. Since the cohesion is non-negative, the index is also non-negative.

2.4.2. Clustering Algorithms

**Visual Assessment of Clustering Tendency (VAT)** To assess the degree to which the elements of a data set tend to form clusters, the VAT algorithm [BH02] can be applied. It reorders the matrix of pairwise distances to show possible clusters as blocks along the main diagonal. It is based on a greedy strategy that interprets the distances between objects as weights of a complete graph. The sought-after permutation is the visiting order of objects in building a minimum spanning tree according to Prim’s algorithm [HB12].
That is, one selects any object as leftmost column and incrementally picks as next column the object that is closest to the current object (and not yet in the matrix). In [HB12], the first object is chosen as any object that has maximal distance to any other object.

An improved version (iVAT) of this algorithm transforms the pairwise distances between objects beforehand. The transformed distance between two objects equals the largest weight of any edge on the shortest path between them. It was shown that this algorithm is capable of revealing a wide range of complicated clustering structures [HB12]. The role of iVAT in the qualitative analysis of clustering structure is discussed in Section 5.2.2.

**K-Means Clustering**  The K-Means problem consists in finding $k$ points in $\mathbb{R}^d$, called centroids, such that the distances of $N$ points to their nearest centroid are minimized. The problem is computationally hard and requires heuristics in practice. To find an approximate solution, the centroids can be iteratively refined. First, the centroids are initialized using some heuristic. In every iteration, each object is assigned to its nearest centroid. Then, the centroids are recomputed by averaging over all assigned objects. The procedure is repeated until the centroids converge or a maximum number of iterations is reached. The computational complexity of this method is $O(NDd)T$ where $d$ is the dimensionality of the feature space, $k$ is the number of centroids and $T$ is the number of iterations. The number of iterations is generally much less than $N$ [DHS12] and can be reduced by choosing good initial cluster centers. For this task, the k-means++ algorithm [AV07] was used. Since the heuristic is susceptible to getting stuck in local maxima, it is usually started several times, using different initial centroids. Among the results, the one with the lowest Inertia was chosen.

**Gaussian Mixture Models**  A Gaussian Mixture Model (GMM) models the distribution of the data in feature space as the weighted sum of $k$ normal distributions. The goal is to find optimal weights $\pi_j$, means $\mu_j$ and covariance matrices $\Sigma_j$ of the normals. Again, the problem is computationally hard and an iterative scheme is used as a heuristic. The algorithm starts with an initial guess of the parameters. In the every iteration, the probability $w_{ij}$ for each object $x_i$ to belong to cluster $j$ is computed.

$$w_{ij} = \frac{\pi_j p(x_i \mid \theta_j)}{\sum_k \pi_k p(x_i \mid \theta_k)}$$

Where $p(x_i \mid \theta_j)$ is the probability to observe $x_i$ under the multivariate normal $\mathcal{N}(\mu_j, \Sigma_j)$. In contrast to K-Means, there is no hard assignment to a cluster. Each object belongs
to every cluster with a certain probability. Then, the weights and parameters of the normals are recomputed.

\[
\pi_j = \frac{1}{N} \sum_{i=1}^{N} w_{ij}
\]

\[
\mu_j = \frac{1}{N\pi_j} \sum_{i=1}^{N} w_{ij} x_i
\]

\[
\Sigma_j = \frac{1}{N\pi_j} \sum_{i=1}^{N} w_{ij} (x_i - \mu_j)(x_i - \mu_j)^T
\]

This procedure is repeated until the parameters converge or a maximum number of iterations is reached. Finally, each object \( x \) is assigned to cluster \( j \) for which \( \pi_j p(x \mid \theta_j) \) is maximal. To initialize the means of the normals, the K-Means algorithm was applied. The weights of the normals were initialized as \( 1/k \) and the covariance matrices were initialized as the covariance matrix of the data set. The heuristic is susceptible to converging in a local optimum and is thus initialized several times. The result with the best log-likelihood is kept.

3. Related Work

To the best of the author’s knowledge, there are few publications related to clustering recurrence plots, and no systematic study of the topic. However, there is related work regarding similarity measures for RPs and clustering RQA measures.

3.1. Similarity Measures for RPs

Bello [Bel11] developed a similarity measure for music recordings, which is based on RPs. It was used to perform similarity search in an audio database of classical music recordings. According to Bello, self-similarity matrices are commonly used in music structure analysis, and their binarization is a common step in Music Information Retrieval. This naturally results in a Recurrence Plot. Time Delay Embedding is applied to the time series corresponding to the music recordings, which emphasizes the relationship to Recurrence Analysis. The embedding parameters were varied in a fixed range to optimize retrieval performance. To measure the similarity between RPs, each RP was represented as a bit string by concatenating its rows. The similarity of two RPs is proportional to the compression rate of the concatenation of their representations. The bzip2-algorithm was used for compression. A similar approach was chosen by Silva et al. [SDSB13] to classify time series. The approach differs in that no threshold was applied, that is, the distance matrices were compared instead of the RPs. Furthermore, it was noted that the bit string representation compromises the spatial neighborhoods in the similarity matrix. They proposed using the MPEG-1 video compression algorithm, which better suits an image interpretation of the matrix. This distance measure, the
CK-1 distance, has already been applied to discriminate textures in a wide variety of application domains [CK10]. Silva et al. referred to it as Recurrence Patterns Compression Distance. It was tested on 38 real world data sets taken from the UCR time series database [KZH+11]. Classification results were much better under CK-1 distance than using Bello’s bzip2 compression scheme. However, in the experiments described in Section 4.2, the compression distance performed poorly with regard to discerning different dynamics. Note that the data set used in this thesis is fundamentally different from the data sets in the UCR time series database. The latter is comprised of univariate time series with a few hundred observations. This thesis focuses on multivariate time series with several thousand observations. The number of observations matters, because dynamics are here assessed as patterns of recurrence. In a very short time series there is simply no space for a large number of recurrences. A visual inspection of several classes of the UCR time series database has not revealed any patterns of recurrence.

Souza et al. [SSB14] also approached the problem of time series classification by means of using similarity matrices as intermediate representation. In contrast to Silva et al., texture features were used to describe the matrices. Four kinds of texture descriptions were used to train a Support Vector Machine. The methods are: Local Binary Patterns, the Grey-Level Co-occurrence Matrix (GLCM), Gabor Filter Banks and Segmentation Based Fractal Texture Analysis. Note that the choice of a texture measures must take into account that Recurrence Plots have only two grey levels. For instance, the original version of the GLCM counts the occurrences of two neighboring pixels with given grey levels. [HSD73] The line structures that form the basis of the RQA measures are strongly expected to give a more differentiated description of the RP than the GLCM. In addition, the RQA measures are well known to be related to properties of the underlying system and its dynamics. The decision to work with Recurrence Plots and not with similarity matrices is justified in Section 4.3. Possible use of other texture measures is discussed in Section 8. In Section 7.3, an extended definition of line structures in an RP is given, which can be interpreted as a step towards texture based distance measures for RPs.

Recurrence Analysis provides methods to relate two Recurrence Plots, for instance for the purpose of analyzing synchronization effects. Two important techniques are the Cross Recurrence Plot (CRP) and the Joint Recurrence Plot (JRP). The CRP simply replaces the term \(||x_i - x_j||\) in Equation 1 by the term \(||x_i - y_j||\) where the \(y_i\) are the phase space points of a second trajectory. Since the points of the trajectories are directly compared, their phase spaces should be the same, especially of equal dimensionality. This requirement might be too restrictive for a general clustering scenario. It also leads to complications when different embedding dimensions are estimated for the time series in a data set. Comparing RPs instead of analyzing CRPs is thus considered the more flexible alternative, as it does not impose restrictions on the phase space trajectories. Joint Recurrence Plots are based on this principle. Given two Recurrence Matrices, the JRP is defined as the element-wise product of the matrices. This does not restrict the trajectories, but requires the RPs to be of the same size. This is again disadvantageous when working with estimated embedding parameters, because they lead to different length trajectories. Multiscale approaches also inherently produce RPs of different sizes.
Hamming Distance, which is related to the JRP, is discussed in Section 4.2, but did not perform well.

Spiegel et al. [SJA14] proposed a distance measure for time series that is based on their CRPs. The distance was defined as $1 - \text{DET}$ of the CRP of the trajectories. The task was to detect typical driving behaviors for an application in the automotive industry. Time series that share a lot of common short-term driving patterns should be clustered together. The K-Medoids clustering algorithm was applied, a variant of K-Means, where cluster representatives are selected among the cluster members. For the task of comparing dynamics, this distance measure does not appear suitable, for the reasons outlined in the discussion of the CRP. In general, more abstract aspects of the trajectories, like predictability, were considered more important than requiring the trajectories to share concrete subsequences.

3.2. Clustering RQA Measures

Rissanen et al. [RK09] have investigated the RQA measures of Electromyography time series. One set of time series was acquired from healthy subjects and another from patients suffering from Parkinson’s disease. The phase space trajectories were reconstructed using time delay embedding, using manually tuned embedding parameters. In addition to RQA measures, wavelets and other descriptors were used to describe the trajectories. To reduce the dimensionality of the feature space, Principal Component Analysis was applied. Using the K-Means algorithm, the trajectories were shown to fall into relatively distinct clusters related to the disease condition. In the context of this thesis, this was taken as an indicator that a simple clustering method might suffice to cluster RQA measures.

Vlahogianni et al. [VKG08] applied clustering to urban traffic time series to identify typical patterns of traffic. The method of Time Delay Embedding was applied, and the parameters were estimated using the Mutual Information Method and the False Nearest Neighbors method. DET, $L_{\text{max}}$, and two non-rqa measures were used to describe the trajectories. A Kohonen Self-Organizing Map was used to reduce the dimensionality of the feature space from four to two. The goal of this preprocessing is to reduce the impact of outliers on the clustering. Finally, K-Means clustering was used to detect prototypical traffic patterns.

Yuan et al. [YYC14] stated that the “network traffic system is a complex dynamic system” and used RQA measures to describe network traffic. The task was to detect anomalous traffic. The RPs of the traffic time series were computed using a threshold of 10% of the phase space diameter. The embedding parameters were chosen based on a visual assessment of the outputs of the Mutual Information and False Nearest Neighbor methods. After applying several transformations to the time series, DET, RR, and $L_{\text{entr}}$ were used to describe them. A training set was clustered using the K-Means algorithm. The clustering is then used to detect anomalous traffic. A given traffic time series is classified anomalous if its closest centroid represents a cluster that contains a majority of anomalous traffic time series.
Pereira et al. [PDM13] developed a method to cluster data streams according to the behavior of the generating system. The algorithm was presented as a general approach to clustering data streams and demonstrated using simulated financial data. In principle, this is similar to comparing trajectories according to their dynamics, but their definition of behavior does not incorporate recurrence structures. Accordingly, RQA measures are not among the measures chosen to describe the underlying systems. A monothetic divisive clustering algorithm is proposed, which has several desirable properties with regard to using clustering as a knowledge discovery method. Monothetic divisive clustering [KR09] can be described as building a decision tree for unlabelled objects. The splits can for instance be chosen such that they reduce the variance in the resulting partitions, which is a costly operation. However, defining clusters as leafs of a decision tree is an attractive approach, since it allows to explore the data set at different levels of detail and explains why the objects in a cluster are grouped together. On the downside, it might not be sufficient to investigate one dimension at a time to reveal the clusters in a data set. Another disadvantage is that the algorithm exhibits quadratic time complexity with respect to the number of streams. In the proposed scenario, this was no problem, since the number of streams was assumed to be in the order of tenths. In contrast, this thesis focuses on larger input sets and a recurrence-based definition of behavior.

In summary, none of the reviewed publications compares different notions of similarity between RPs. Furthermore, the author is not aware of a systematic study of a similarity measure for clustering RPs.

4. Similarity Measures for Recurrence Plots

When developing a similarity measure, both domain knowledge and consideration of application specific requirements are necessary. To describe for instance the similarity between strings, knowledge in linguistics or genetics is useful, depending on the type of the data. In Recurrence Analysis, basic knowledge about the meaning of the structures in an RP is required, as provided in Chapter 2. Furthermore, the design of the similarity measure depends on the task at hand. Returning to the string similarity example, Edit Distance is a reasonable choice to find similar words for spell correction, but for detecting near-duplicate documents, other similarity measures are more appropriate. In Recurrence Analysis, a common task is to compare systems based on their dynamics. The term dynamics refers to different aspects of the underlying system, such as predictability. RQA measures quantify many important such aspects. In this thesis, they are used as a black box specification of key aspects of dynamics. To a certain degree, such an approach is mandatory, since delving into the underlying concepts, like Lyapunov Exponents or Rényi Entropy, is not within the scope of this thesis.

In summary, using RQA measures as a starting point for assessing the similarity between RPs solves both problems. Firstly, for domain experts, the similarities will
be interpretable\textsuperscript{5} and refer to relevant aspects of the RPs. Secondly, they provide a quantitative approach to assessing system dynamics [GG12, p. 232], and comparing system dynamics is a common application scenario.

### 4.1. Evaluation Framework

Similarity measures were assessed within the following evaluation framework. An artificial data set was created, which consists of prototypical systems with varying degrees of predictability. Additionally, measures of success were derived, the central one being adherence to a reference clustering. In general, decisions like the selection of systems, which are strongly related to Recurrence Analysis, were discussed with Dr. Marwan from the Potsdam Institute for Climate Impact Research (PIK). The artificial data set comprises nine classes of RPs, as summarized in Table 2 and illustrated in Figure 4.1.

**Description of Classes** The Oscillator systems follow a simple circular motion, which is completely predictable. Oscillation is a general behavior that occurs in many different contexts, for instance as seasonal behavior in weather data or as periodic processes in the human heart and brain. The Lorenz System is supposedly the best known chaotic system, and was developed as a simple model for atmospheric convection. It has various applications in describing the dynamics of processes related to lasers, electric circuits, and chemical reactions. Under the Noisy Oscillation parameters, the Lorenz System exposes semi-periodic behavior [Spa82], i.e. it exposes a predictable traversal order of the two scrolls, but with the sizes of the scroll instances vary. The Rössler System was initially devised as a simpler to analyze version of the Lorenz System, and has become another commonly used example for chaotic systems. Its dynamics are for instance

\textsuperscript{5}Under observational noise, the interpretability of RQA measures might suffer. Nonetheless, they were still found to be distinctive for different types of systems.
Figure 4.1: Overview of the classes of Recurrence Plots in the artificial data set.
related to processes in chemistry [Rös76]. The Funnel variation of the Rössler System exposes less structured oscillations [OPRK97]. Normally distributed noise is a common type of time series, and a typical use case for grouping all noise trajectories would be to remove them from a data set. Gamma noise is a skewed, i.e. bursty variant of noise. Finally, the Autoregressive processes are a simple stochastic time series model that adds a linear correlation to a noise signal. For an Autoregressive process of order one, each observation depends linearly on the previous observation and a noise component. Autoregressive processes are e.g. related to meteorological observations [Wil11, p. 144] and financial time series.

For each class, 100 RPs were created, mostly by varying the initial conditions of the systems. In the case of the oscillators, varying the initial conditions would not affect the RPs, since the phase of the signal has no effect on the RP. A natural parameter of the oscillators is their wavelength, or the traversal speed of the circle. The oscillator trajectories are the only systems that vary their time scale, which was found to have a substantial effect on clustering. Details regarding the computation of the trajectories and their RPs are given in Section 5.1.

**Hierarchical versus Partitional Clustering** The quality of a clustering under a given similarity measure was measured as adherence to a reference clustering. Both partitional and hierarchical reference clusterings were considered. An apparent option for hierarchical clustering seems to be a three-level hierarchy where the first level refers to kind (deterministic, stochastic), the second level refers to system type (Oscillator, Rössler, Lorenz, Autoregressive, Noise), and the third level refers to the nine classes as defined above. However, expecting that two dynamics are similar, just because they are both produced by e.g. the Rössler System, is not correct. For certain parameters, the Rössler system produces perfectly predictable trajectories, which are very similar to the oscillator trajectories. These trajectories expose strongly different dynamics compared to the chaotic trajectories generated under other parameter ranges of the Rössler System. Thus, this hierarchical clustering fits better an artificial view of the data, rather than the actual dynamics of the trajectories. To assess these, domain knowledge is required, which was incorporated by consulting domain experts, as described below.

In general, hierarchical clusterings allow for exploring a data set at different levels of detail, which is desirable for explorative tasks. Nonetheless, their impractical properties were found to outbalance this advantage. Firstly, with respect to the overarching goal of exploring large sets of RPs, the scalability of the clustering algorithm is important. Hierarchical clusterings depend on the pairwise distances between objects, which implies increased computing and memory requirements as compared to centroid-based clustering schemes. Secondly, comparing a hierarchy to a reference hierarchy was found to be much more difficult than for flat partitionings. For the latter, several well known and chance-adjusted indices exist, whereas for the former, the literature was found to be relatively sparser. Thus, a partitional clustering scheme was decided on for the evaluation framework.
Expert Assessment of Similarities  In [Mil96], several questions regarding clustering evaluation are highlighted. For instance, “How is the reader to believe the author’s a priori grouping of the data?” This question is especially important when no data sets with a well known structure are available, as in this thesis. In principal, the data set was designed such that all systems expose unique dynamics. However, to verify this assumption, two reference similarity matrices were kindly provided by Dr. Marwan and Dr. Riedl from the PIK. The similarity matrices are the result of a manual assessment of similarity between example RPs. During the assessment, 36 inter-class pairs and 9 intra-class pairs of RPs were presented, similar to the RPs in Figure C.1. For each pair of RPs, the similarity could be rated on a scale from 0 to 100%, either using a slider control, or preset buttons for 0, 20, 50, 70 and 90% similarity. The results were converted to a dissimilarity matrix by subtracting the similarities from 100%, as shown in Figure 4.2. Firstly, there seems to be a relatively strong disagreement between the two assessments, which is attributed to the experimental setup. The participants were asked for a general similarity assessment without providing further instructions about which aspects should be considered in particular. However, two commonalities of the matrices are the strong intra-class similarities and the separation between the systems of deterministic and stochastic kind. Because Dr. Marwan was the domain expert consulted through the course of this thesis, his assessment is used as primary reference. For better legibility, the similarities are listed in Table 3. Returning to the question of whether the systems expose unique dynamics, Dr. Marwan’s assessment supports an affirmative answer. In contrast to Dr. Riedl’s assessment, it does not suggest that two classes should be merged. For instance, in both assessments, two candidate classes for merging are AR- and AR+, because of their high inter-class similarity. Still, in
Table 3: Percentual similarities between classes as assessed by Dr. Marwan.

<table>
<thead>
<tr>
<th></th>
<th>OSC</th>
<th>R</th>
<th>RF</th>
<th>L</th>
<th>LO</th>
<th>AR-</th>
<th>AR+</th>
<th>N</th>
<th>NG</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSC</td>
<td>90</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>90</td>
<td>50</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RF</td>
<td>90</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>70</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LO</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AR-</td>
<td>70</td>
<td>70</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90</td>
<td>20</td>
</tr>
<tr>
<td>N</td>
<td>90</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90</td>
</tr>
</tbody>
</table>

Dr. Marwan’s assessment, the intra-class similarity of the AR+ Recurrence Plots was considered higher than the inter-class similarity between AR+ and AR-. Merging both classes would level these different degrees of similarity. The same applies to the R and L classes. Thus, in accordance with the assumption of unique dynamics, a nine class reference clustering was decided on, consisting of the classes described above.

Note that the collected data is not intended as training data, in a machine learning sense. In that case, much more data would have had to be collected, and the design of the assessment procedure would have had to be carried out more carefully. Rather, the similarity matrices were collected to verify basic assumptions about the structure of the artificial data set and as a reference for the preliminary experiments with different similarity measures in Section 4.2.

In summary, the central quality criterion for similarity measures is the adherence to the nine-class reference clustering. A secondary criterion is performance. Since the overarching goal is the exploration of large sets of RPs, the similarity measure should be scalable, preferably linear in the size of the RP, that is, at most quadratic in the size of the trajectory. In this regard, feature vectors are preferable over direct comparison of the RPs, because they allow for centroid-based clustering algorithms, which are faster than medoid-based or distance matrix-based clustering algorithms. The memory requirement should preferably be sublinear in the size of the RP, for instance to avoid memory shortages in parallelization scenarios. Furthermore, the similarity measure should be able to compare RPs of different sizes. The reason is that even a data set of equally long time series will result in RPs of different size if the embedding parameters are estimated. The length of an embedded trajectory is \( N - (m - 1)\tau \), where \( N \) is the number of observations in the original time series, \( m \) is the embedding dimension and \( \tau \) is the embedding delay. Estimation methods will usually result in different \( m \) and \( \tau \) for the time series of a data set.

6 The length of an embedded trajectory is \( N - (m - 1)\tau \), where \( N \) is the number of observations in the original time series, \( m \) is the embedding dimension and \( \tau \) is the embedding delay. Estimation methods will usually result in different \( m \) and \( \tau \) for the time series of a data set.
4.2. Preliminary experiments

As outlined in the introduction of this section, RQA constitute a natural choice for assessing the similarity between RPs. When discussing the problem with Dr. Marwan, a comparison with other similarity measures was regarded an interesting inquiry. However, the experiments suggested focusing on RQA measures for an in-depth evaluation.

Through the course of this thesis, the artificial data set was repeatedly refined, to provide as lifelike experimental conditions as possible. The data shown here was updated using a small version of the final data set, since the preliminary experiments were not performance-optimized. It consists of 4 trajectories per class, where each trajectory comprises 1500 observations.

**Hamming Distance**  As a baseline distance measure, Hamming Distance was chosen. Given two RPs \( R \) and \( R' \), the Hamming Distance between them is defined as the fraction of matrix cells for which \( R_{i,j} \neq R'_{i,j} \). The pairwise distances are shown in Figure 4.3. The results are clearly not acceptable, for instance because the measure leads to strong intra-class dissimilarities for the RF, AR-, AR+, N, and NG trajectories. In fact, the stochastic systems are more similar to the deterministic systems than to each other. One defect of the measure is that it does not account for structures beyond single recurrence points. Even with regard to single recurrence points, it is completely inflexible concerning their position. Consider adding one observation to the beginning of the trajectory and removing one from the end; this would result in an RP that is shifted by one cell along the main diagonal. Due to the inflexibility of directly comparing matrix cells, the impact would be largely unpredictable, resulting in a substantial lack of robustness.

Another disadvantage is that it requires RPs to be of equal size. These weaknesses are in principle shared by all measures that are based on element-wise comparison of the Recurrence Matrices, like Jaccard distance, correlation, or inter-rater agreement. Finally, while theoretically in range \([0, 1]\), the maximum observed distance is about 0.18, which is another indicator for poor performance.

**Spatiogram Distance**  Spatiograms are extensions of histograms, which add higher order moments to each bin. Instead of recording only the count (the zeroth order moment), the mean and covariance matrix of the objects in each bin can also be recorded. In the style of RQA measures, the objects of interest were diagonal and vertical lines. The mean and covariance matrix were computed based on the coordinate of the left- and lowermost point of each line. The length of the line determines the bin of the line. Two spatiograms were created, one for the diagonal and one for the vertical lines. As a distance measure, the Spatiogram distance from \([\text{COS07}]\) was chosen. Figure 4.3 shows the distances of the diagonal line Spatiograms. The results are much better than for Hamming Distance, but still not convincing. A positive aspect is that there is a relatively strong separation between deterministic and stochastic systems. The distance measure also exploits almost its full theoretical range of \([0, 1]\). On the other hand, it is seemingly not able to distinguish between any of the stochastic systems. Additionally, the intra-class similarities are much too low for all except the LO trajectories, and maybe
the RF trajectories. Finally, the measure was found to be extremely sensitive to the choice of the minimum line length parameters. This is problematic, since no generally accepted heuristic for the choice of the minimum line lengths exists.\footnote{Private conversation with Dr. Marwan.} The reason for the partly superior performance over Hamming distance should be due to incorporation of line structures.

**Compression Distance** Referring to the work of Silva et al. [SDSB13], compression distance was evaluated as another proximity measure. The similarity between two RPs is rated proportional to the compression factor that is achieved on a two-frame video consisting of the two RPs. The compression rate is computed by applying MPEG-4 video encoding. MPEG-4 is able to align parts in two subsequent images, which should help to overcome the inflexibility of Hamming Distance. Even so, the distance matrix exposed a strong lack of intra-class similarity and the response of the measure appeared overall spurious. Another downside of the measure is that it is intransparent with respect to what factors lead to a given proximity between two RPs.

**RQA Vectors** Finally, the 16 RQA measures from Chapter 2.3 were used as components of a vector. The distance between RPs was then measured as euclidean distance between the RQA vectors. This lead by far to the best results, as shown in Figure 4.4. The distances expose better intra-class similarities and inter-class distances than any of the previous proximity measures. Furthermore, this measure clearly matches best Dr. Marwan’s manual proximity assessment.

As an additional approach related to RQA-measures, the line length histograms were compared directly. This seems to be reasonable, as they form the basis of the RQA-
measures. Three distance measures were evaluated: Intersection Distance, Earth-Mover Distance, and Dynamic Time Warping. The former two are described in [RTG00], the latter one is described for instance in [KP00]. However, none of the measures gave promising distance matrices. This suggests that the RQA measures comprehensively summarize the salient aspects of the line length histograms.

4.3. Geometric and Temporal Abstractions in RQA Measures

The preliminary experiments have focused on different aspects of the data. First, the RPs were directly compared, using Hamming and Compression distance. Then, the line length histograms of the RPs were compared, both directly and by extending them to Spatiograms. Finally, statistics of the line length histograms were compared, the RQA measures of the RPs. Two key abstractions were identified in the experiments that are hypothesized to explain the superior performance of the RQA measures.

**Geometric Abstraction** An RP abstracts from its underlying trajectory by focusing on its patterns of recurrence. These are, to a certain degree, independent of phase space points that constitute the trajectory. Figure 4.5 shows two trajectories with different geometries but identical recurrence structures, i.e. two repetitions of a pattern. A very small recurrence threshold \( \varepsilon \) reveals this recurrence structure as two diagonal lines above the main diagonal. These diagonals obviously contain no information about the geometry of the underlying trajectories. Only for larger thresholds, the difference between the motifs becomes apparent, as additional patterns around the diagonal lines emerge. In [MW14], it is stated that “in all cases it is desirable that the smallest threshold pos-
possible is chosen.” It is concluded that in the context of Recurrence Analysis, a certain abstraction from the geometry of the phase space trajectories is desirable.

Regarding the question of whether the recurrence matrix or its underlying distance matrix should be used, a similar argument applies. The distance matrix is intimately related to the geometry of the trajectory, as there is “no loss of dynamical or geometric information in going from the original locus of points to the set of distances between them” [MAS97]. It is hypothesized that the thresholding of the distance matrices introduces an abstraction that allows to focus on the recurrence structure rather than the geometry of the trajectory. This requirement disqualifies traditional time series similarity measures, like euclidean distance and Dynamic Time Warping, since they do not abstract from the geometry of the trajectory.

Temporal Abstraction The line length histograms abstract from their underlying RPs by discarding temporal information. Only the length of a line is recorded, but not its location in the RP. This is necessary, since the Butterfly Effect causes trajectories with identical dynamics to diverge quickly. This is here termed time locality, and refers to the impossibility of long-term predictions. As a result, two timely distant points can be practically unrelated. Still, a fundamental property of dynamical systems is recurrence, which indeed relates timely distant observations. This is formalized in the Poincaré recurrence theorem, that expresses the idea that certain dynamical systems keep returning to some states in sufficiently long, but finite intervals. In consequence, two timely distant points of a trajectory might or might not be related, which justifies the approach of computing the distance between all of them. Specifically, this implies that an element-wise comparison of the trajectories does not make sense. Even the temporal flexibility that is introduced by Dynamic Time Warping does not suffice, because time
locality and recurrence allow for permutations of patterns. As an example, all trajectories of the Lorenz System expose the two characteristic patterns, sometimes referred to as scrolls, that are shown in Figure 2.2.\textsuperscript{8} These patterns recur endlessly, independently of the initial conditions [Spa82]. However, the smallest variation in the initial conditions of two trajectories will lead to different traversal orders of the scrolls.

The line length histograms compensate for the Butterfly Effect using the recurrence property. When working with line length histograms, temporal closeness matters only where a trajectory exposes recurrent behavior, i.e. in the line structures of the RP. Beyond the limits of a line, the histograms discard all temporal information. As a result, the disruptive Butterfly Effect is compensated and an abstraction from initial conditions is established. Clearly, traditional time series distance measures are not designed for the special role that time plays in dynamical systems.

Two further approaches for a similarity measure can be ruled out by the following thought experiment. Partition a Lorenz System’s trajectory according to whether the trajectory is in one or the other scroll. Then, rearrange the traversals of one scroll and the other in alternating order. The resulting trajectory is more predictable and thus exposes different dynamics. Consequently, describing dynamics by the set of its recurring patterns (the two scrolls, or Unstable Periodic Orbits, see Section 7.3) is not sufficient, since the traversal order determines the predictability. Note that both trajectories are formed by the same points in phase space, only their visiting order differs. Therefore, considering only the distribution of the points in phase space is also not sufficient.

5. Clustering Recurrence Plots

5.1. Artificial Data Set

5.1.1. Computation of Trajectories

The selection of the classes in the artificial data set has been motivated in Chapter 4. Table 4 shows the equations that were used to generate the trajectories within the nine classes of the data set. Within each class, 100 trajectories were generated, each comprising 25,000 observations. The numerical integration of the chaotic systems was performed using a fourth-order Runge-Kutta method, as in [MW14]. The first 1000 observations were discarded to avoid artifacts due to settling behavior. All trajectories of chaotic systems were generated by partitioning a single long integration run into the required number of trajectories.\textsuperscript{9} This has the same effect as varying the initial conditions\textsuperscript{10}, but settling behavior does not occur, because the trajectory is already

\textsuperscript{8}Given that the parameters are chosen in certain ranges, for instance using the standard parameters $\sigma = 10$, $\beta = 8/3$, $\rho = 28$.

\textsuperscript{9}Let $x_i, 1 \leq i \leq 25000 \cdot 100 + 1000$ be the result vectors of the numerical integration. The $k$-th trajectory consists of the vectors $x_{(k-1) \cdot 25000+1001}, \ldots, x_{k \cdot 25000+1000}$.

\textsuperscript{10}Private conversation with Dr. Marwan.
within the normal range of the attractor. As initial conditions for the Lorenz Systems, the point $(-8, 8, 27)$ was used. The Lorenz equations are:

$$\frac{dx}{dt} = \sigma(y - x) \quad \frac{dy}{dt} = x(\rho - z) - y \quad \frac{dz}{dt} = xy - \beta z$$

For the Rössler Systems, the point $(1, 2, 3)$ was used as initial condition. The Rössler equations are:

$$\frac{dx}{dt} = -y - z \quad \frac{dy}{dt} = x + ay \quad \frac{dz}{dt} = b + z(x - c)$$

As integration step size, the values $\Delta t$ from Table 4 were applied. For the Autoregressive processes and the noise trajectories, the variation of initial conditions corresponds to varying the seed of the random number generator. As the first value for the Autoregressive processes, a random number from $\mathcal{N}(0, 0.1)$ was used. For the oscillators, $\lambda$ was varied such that the circle is traversed at least three times and such that each traversal consists of at least 16 observations.

<table>
<thead>
<tr>
<th>System</th>
<th>Abbr.</th>
<th>Equations</th>
<th>Parameterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Periodic motion</td>
<td>OSC</td>
<td>$x(t) = \cos(\lambda t)$, $y(t) = \sin(\lambda t)$</td>
<td>see text</td>
</tr>
<tr>
<td>Rössler System, Standard Parameters</td>
<td>R</td>
<td>Rössler Equations</td>
<td>$a=0.2$, $b=0.2$, $c=5.7$, $\Delta t=0.085$</td>
</tr>
<tr>
<td>Rössler System, RF</td>
<td>RF</td>
<td>Rössler Equations</td>
<td>$a=0.2925$, $b=0.1$, $c=8.5$, $\Delta t=0.085$</td>
</tr>
<tr>
<td>Lorenz System, Standard Parameters</td>
<td>L</td>
<td>Lorenz Equations</td>
<td>$\sigma=10$, $\beta=2.66$, $\rho=28$, $\Delta t=0.025$</td>
</tr>
<tr>
<td>Lorenz System, Noisy Oscillation</td>
<td>LO</td>
<td>Lorenz Equations</td>
<td>$\sigma=10$, $\beta=2.66$, $\rho=198$, $\Delta t=0.025$</td>
</tr>
<tr>
<td>Autoregressive, Negative Correlation</td>
<td>AR-</td>
<td>$x(t) = \phi x(t - 1) + \mathcal{N}(0, 0.1)$</td>
<td>$\phi = -0.95$</td>
</tr>
<tr>
<td>Autoregressive, Positive Correlation</td>
<td>AR+</td>
<td>$x(t) = \phi x(t - 1) + \mathcal{N}(0, 0.1)$</td>
<td>$\phi = 0.95$</td>
</tr>
<tr>
<td>Normal Noise</td>
<td>N</td>
<td>$x(t) \sim \mathcal{N}(\mu, \sigma^2)$</td>
<td>$\mu = 0$, $\sigma = 1$</td>
</tr>
<tr>
<td>Gamma Noise</td>
<td>NG</td>
<td>$x(t) \sim \Gamma(k, \theta)$</td>
<td>$k = 0.5$, $\theta = 1$</td>
</tr>
</tbody>
</table>

In a real-world scenario, RPs that are not relevant for the application at hand could be considered noise. These can, but must not be RPs produced by noise trajectories. In the
context of the artificial data set, the two noise classes primarily constitute just another type of system dynamics. In this sense, there is no noise in the artificial data set. However, a notion of noise that is present in the artificial data set is an intra-class variation of the RQA measures. Since this variation can not be controlled, it is considered noise. An ideal feature set would give the exact same values across all trajectories of a class, and different values for different classes. To eliminate as much uncontrolled variation as possible, the trajectories in each class should be as representative as possible. The more uncontrolled variation is eliminated, the more reliable can the clustering quality be attributed to the experimental parameters.

5.1.2. Selection of Trajectory Lengths

To get a representative distribution of line lengths, the RPs must have a minimum size. The more stable the line length histograms are, the less variation will be present in the RQA measures. The problem is that a dynamical system’s trajectory is in principal infinite, such that there is no obvious minimum trajectory length to choose. To find a suitable minimum trajectory length, a method similar to that in [LB14] was applied. The RQA measures are plotted for one trajectory of each class, gradually increasing the length of the trajectory. Based on a visual analysis, the RQA measures were found to be sufficiently convergent for a length of 25,000 observations per trajectory. Figure 5.1 shows the data for selected RQA measures of a trajectory of the Rössler System. The chosen length of 25,000 observations is marked in grey. The full data is shown in Figure C.2. Note that the choice of long trajectories does not lead to a loss of generality. RPs are generally suited to analyze short trajectories [MW14, p.9]. Choosing as few as 500 or 2000 observations per trajectory leads to comparable results, see Figure C.10. Note that details regarding the interpretation of Figure C.10 are given in subsequent chapters.

A comprehensive approach to convergence would have to run the analysis on multiple trajectories of each class. However, the goal here was to obtain a rough estimate of a suitable minimum trajectory length, not an analysis of convergence behavior. Based on the limited data, almost all RQA measures converge rather quickly, except for the maximum line lengths $L_{\text{max}}$, $V_{\text{max}}$, and $W_{\text{max}}$. Most measures were found to behave roughly like the diagonal entropy, whereas the maximum line lengths clearly do not converge in the investigated range. Especially the measure $W_{\text{max}}$ did not stop to show large increases up to 100,000 observations in all systems. This is shown in the last column of Figure C.2.

Remarkably, initial experiments on trajectories of as few as 500 or 2000 points showed similar outcomes to the ones presented below, which suggests that working with shorter trajectories should also work. Note, in this respect, that the diagonal entropy converges already around 3,000 observations. In general, the necessary amount of cohesion depends, among other factors, on the distance of the cluster centers and the clustering algorithm of choice. As such, the concept of convergence is of interest mostly in terms of experimental design, and less with regard to real-world applications.
Figure 5.1: Evolution of $L_{\text{entr}}$, TT, $L_{\text{max}}$, and $W_{\text{max}}$ with increasing trajectory length of a Rössler System. The measures are plotted relative to the maximum value across all trajectory lengths. The selected length of 25,000 observations is marked in grey.

5.2. RQA Measures of the Artificial Data Set

5.2.1. Computation of RPs and RQA Measures

The data that all experiments are based on, are the RQA measures of the RPs of the trajectories in the artificial data set. To compute the RPs, a recurrence threshold of 5% of the trajectory diameter, the maximum distance between any two points on the trajectory, was chosen. This is a popular criterion in the literature, see for instance [MCRTK07, YYC14, SDM08]. Another popular criterion is to choose the threshold such that the resulting RP exhibits a given Recurrence Rate [MCRTK07, p.248]. Using the diameter criterion has the advantage that the Recurrence Rate can be used as an additional feature to distinguish the systems, while it would be almost constant using the latter method. To compute the recurrence matrix, the euclidean norm was used. Note that there is no need to choose an embedding dimension and embedding delay when working on multivariate trajectories in the first place. The effect of embedding is considered separately in Section 6.2.

To compute the RQA measures of the RPs, the minimum line lengths and the Theiler Corrector need to be chosen. For the minimum line lengths of the diagonal and vertical lines, no well established heuristic exist.\(^{11}\) Nonetheless, the minimum line lengths are critical for the estimation of Determinism and Laminarity. The apparent lack of an in-depth study of these two parameters is probably due to the fact that these parameters are often chosen based on visual inspection of the RP. However, to analyze large numbers of RPs, a mostly automatic processing of the trajectory is required. As a pragmatic solution, the minimum diagonal and vertical line lengths were set to 2. In a similarly

\(^{11}\)Private conversation with Dr. Marwan.
pragmatic way, the Theiler Corrector $w$ was set to 1. While this is certainly not optimal, optimizing the choice of these parameters should only improve clustering performance. However, the influence is expected to be small and is thus left as possible future work.

5.2.2. Qualitative Results

According to John Tukey, “There is no excuse for failing to plot and look”. In clustering, visual analysis can be used prior to the actual clustering to get an impression of the separation of the data and the shape of the clusters. A simple visualization of the 16 dimensional feature space of RQA measures is the scatter plot matrix. This visualization shows the distribution of the 900 RPs in each two-dimensional subspace of the feature space. These plots can be used as a first indicator which features separate the classes well. Among all 120 two-dimensional subspaces, the projection to the entropy features shows best the separation of the classes. Since the entire plot is large, only these two subspaces are shown in Figure 5.2. Regarding the interpretation of the scatter plots, note that the euclidean distance in each two-dimensional subspace lower bounds the distance in the full feature space. The reason is that the sum of squares can become only larger when considering additional dimensions. It is however possible that a seemingly large distance is small compared to the distance in the full feature space, so the information must be interpreted with care.

Instead of plotting all two-dimensional subspaces of the data, a common approach is to reduce the dimensionality of the feature space before plotting it. Principal Component Analysis (PCA) is a standard method to reduce the dimensionality of a feature space. As Milligan [Mil96] and Everitt [ELLS11] point out, the method must be applied with care, because variance along an axis does not necessarily imply separation along that axis. In general, every projection to lower dimensional spaces can obscure the clustering structure. Thus, for the purpose of detecting clusters graphically, different approaches to finding better projections of the data have been proposed, as described in Chapter 2 of [ELLS11]. However, rotating the data in feature space can also reveal clusters that are hidden in axis parallel projections. Starting from the assumption that different classes are located in distant positions in feature space, the idea of projecting the data set onto

Figure 5.2: Scatter plots of $L_{\text{entr}}$, $V_{\text{entr}}$, and $W_{\text{entr}}$ of the RPs in the artificial data set.
Figure 5.3: Scatter plots of the first two principal components of two selections of RQA measures.

(a) All RQA measures
(b) Without maximum line lengths

With these caveats in mind, applying PCA to the data was found to be generally useful in the present scenario. Figure 5.3a shows the scatter plot of the first principal components of the data, which reflects the two most salient characteristics of the full scatter plot matrix. Firstly, most RQA vectors seem to fall into well separated clusters, which is clear only in few 2D subspaces, for instance the projections in Figure 5.2. Secondly, the RQA measures of the oscillators are scattered strongly. The large variation is due to the increase of vertical and white vertical line lengths with decreasing $\lambda$. As the system traverses the circle slower, more subsequent observations fall within each other’s $\varepsilon$-neighborhood, which causes the vertical lines to become thicker. Likewise, more observations elapse between two visits of a point on the circle, which causes the white vertical lines to become longer. Referring back to the convergence discussion above, removing the maximum line lengths from the feature set seemingly results in a much stronger separation, as shown in Figure 5.3b. The influence of this factor is thus evaluated in more detail in Section 6.

Finally, a way to entirely avoid projections to lower dimensional spaces is to visualize the pairwise distances between all points. Another advantage of the similarity matrix approach is that it takes into account the chosen similarity measure and that the objects need not be of vector form. To visualize for instance the clustering structure of the RPs under compression distance, the scatter plots are not an option. A disadvantage of the approach is that it provides no information about the features that cause the proximity of two given objects. Note that RPs also utilize pairwise distances to visualize high-dimensional point sets in two dimensions. In contrast to the points on a phase space trajectory, there is usually no natural order for the points in feature space. In a supervised scenario, the points can be grouped by their class labels, in which case the
distance matrix visualizes the intra- and interclass distances. However, without an appropriate ordering of the rows and columns, the visualization will not be able to convey a sensible impression of the distribution in feature space. As described in Section 2.4.2, the iVAT algorithm attempts to find such an order. Figure 5.4 shows the 900×900 euclidean distance matrix and its transformation as produced by the iVAT algorithm. The improved intra-class similarity and inter-class separation as compared to the preliminary experiments in Chapter 4 is most likely due to the strongly increased trajectory length. As discussed in Section 5.1.1, this leads to less variation in the RQA measures which increases both cohesion and separation of the classes. The iVAT presentation suggests an increased inter-class similarity between the L, R, and LO trajectories and the AR+, NG, and N trajectories. The former complies with the expert similarity assessment in Chapter 4. In summary, all visualizations suggest a pronounced clustering structure of the data.

5.3. Clustering RQA Measures

5.3.1. Experimental Design

The results in this chapter are based on a total of 1.8 million clusterings of the RQA measures described in Section 5.2. The description of the experimental design follows the seven step outline of a cluster analysis given in [Mil96]. Table 5 provides an overview of the parameters that were chosen for the systematic evaluation. In this chapter, the effects of observational noise and embedding are not yet considered. However, it was found that the allegedly best conditions did not lead to the best results. In fact, small amounts of noise consistently improved the results, before they abruptly deteriorated when exceeding a certain amount of noise. Therefore, the discussion of the results is
partly delayed until Chapter 6, where the influence of noise and embedding is taken into consideration.

Subsampling Milligan [Mil96] points out that, if demonstrating the performance of a clustering algorithm on a single data set, “the validation information has been based on a sample size of one.” For practical application, it is important to demonstrate that the method is robust against moderate perturbations of the data set. Since the generation of a data set took several hours\textsuperscript{12}, even if parallelized, a subsampling approach was chosen. A total of 500 subsamples was generated. Each subsample contained 600 of the total 900 RPs, which were selected randomly and without replacement. Monti et al. [MTMG03] also opt for a subsampling approach, since they argue that bootstrapping a data set for clustering is “artificially inflating the compactness”. Using 600 RPs per subsample, two data sets have an expected 400 elements in common\textsuperscript{13}. The expected fraction of $2/3$ of shared elements was considered a good trade-off between variability and comparability. The departure from the standard method of bootstrapping is on the grounds of the argument of Monti et al. For a sample of size $N$, the basic bootstrapping procedure consists of drawing with replacement $N$ random elements from the sample and then to recompute the statistic on that sample. While duplicate elements do not seem to pose a problem for estimating e.g. the sampling distribution of a mean, it is hypothesized that clustering algorithms are in principle amenable to such subtleties.

Note that the outcomes of clustering a subsample with regard to different clustering parameters form a paired observation. Varying the subsampling in each parameterization would not allow to attribute qualitative differences to the influence of a given parameter, since the experiment would not control for the variation induced by the subsampling.

Table 5: Experimental Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Choices</th>
<th>Cardinality</th>
<th>Number of Clusterings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure Selection</td>
<td>Standard, Convergent</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Dimensionality Reduction</td>
<td>No Reduction, PCA</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Clustering Algorithm</td>
<td>K-Means, GMM</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>2, ..., 16</td>
<td>15</td>
<td>120</td>
</tr>
<tr>
<td>Data Set Resamples</td>
<td>random</td>
<td>500</td>
<td>60,000</td>
</tr>
<tr>
<td>Number of Initializations</td>
<td>random</td>
<td>30</td>
<td>1,800,000</td>
</tr>
</tbody>
</table>

\textsuperscript{12}Partly because the trajectories are relatively long and no sophisticated RQA algorithm, like a GPU implementation, was used. The second reason is that each data set is subjected to all 27 combinations of embedding methods and noise ratios, as described in Chapter 6.

\textsuperscript{13}The expected cardinality of the intersection set of two subsets of cardinality $k$, drawn from a set of cardinality $N$, is $k^2/N$. Assuming a uniform probability of $k/N$ for each element $e$ of the superset to be part of either subset, $P(e \in A \cap B) = k^2/N^2$. Since we have $N$ elements, the expected size of the intersection is $Nk^2/N^2 = k^2/N$. 

38
Element and Feature Selection  According to the seven-step outline of clustering, the first steps comprise the selection of elements and features. A central aspect of element selection is outlier removal. The RQA measures appear to be very reliable in this respect except for the class of oscillators. As eight out of nine systems appear to expose no outliers at all, the problem was not considered substantial enough to include for systematic evaluation. Moreover, it is not clear how the general problem of outlier removal poses special requirements in the context of clustering RPs, such that artificially introducing outliers does not seem to be sensible.

Regarding feature selection, according to Milligan [Mil96], “a variable should be included only if strong justification exists that the variable helps to define the underlying clustering.” Without prior knowledge, selecting RQA measures according to this criterion is not possible. Therefore, the option Standard in Table 5 refers to including all 16 RQA measures defined in Section 2.3. Drawing from the observation of non-convergent behavior of the maximum line lengths, the option Convergent refers to removing $L_{\text{max}}$, $V_{\text{max}}$, $W_{\text{max}}$, DIV, and $W_{\text{div}}$ from the feature set.

It is desirable to have features selected automatically, and a variety of algorithms exists for this task. One problem is that many of them come with specific assumptions, which complicates their application and comparison. Since the focus of this thesis is not the evaluation of feature selection algorithms, only PCA was chosen as a basic approach. The options No Reduction and PCA in Table 5 refer to whether PCA is applied to the given feature set before applying the clustering algorithm. With the caveats from Section 5.2.2 in mind, the idea is that the influence of variables that do not contribute to the clustering should be reduced in the transformed data. The criterion was to retain the minimum number of principal components to explain 95% of the variance in the data set. This resulted in a median of 5 retained dimensions for the Standard feature set and a median of 3 retained dimensions for the Convergent feature set.

Standardization and Similarity Measure  The choices of a variable standardization method and a similarity measure constitute step 3 and 4 of the outline. Reasonable defaults have been determined in preliminary experiments. Concerning standardization, peak-to-peak normalization was found to give slightly better results than scaling the standard deviation of each RQA measure to one. This is in accordance with the results of the generic standardization study reported by Milligan [Mil96]. Specifically, each measure is divided by the difference between the maximum and minimum value of that measure. In general, standardization is strictly necessary, since the RQA measures are located on largely different scales. Despite that, the specific method seems to have only little impact on the clustering results.

The RQA measures RR, DET, and LAM are by definition within range $[0, 1]$ and in principle need not be standardized. Even so, the measures often do not exploit their full range, thus normalizing them tends to increases their influence. In the preliminary experiments, normalizing all measures was found to slightly increase the separation of the classes, which is the option that was chosen throughout the experiments.
As similarity measure, the $L_2$ norm was chosen. Experiments with other norms have either shown little variation or much worse performance. Distance measures like correlation and cosine distance can be ruled out using a domain knowledge argument. High and low values of Determinism are for instance a reliable indicator to discriminate stochastic from deterministic systems. However, under cosine distance $d_{\text{cos}}(a, \lambda a) = 0$, thus, different degrees of Determinism do not necessarily induce distance. Likewise, under correlation distance $d_{\text{cor}}(a, a + c) = 0$. Surprisingly, the similarity matrices under cosine and correlation distance retain a notable amount of intra-class similarity and inter-class separation. This indicates that the relationship between the RQA measures of an RP is another indicator for the system dynamics, independently of their absolute values. An illustration of the profiles of the RQA vectors is shown in Figure C.3.

**Clustering method and Number of Clusters**

A central choice to be made is which clustering algorithm to apply to the data. As a baseline, the popular K-Means algorithm was chosen, having the advantage of being scalable and widely covered in the literature. As a competitor, the Gaussian Mixture Model (GMM) was chosen. This choice is motivated by the seemingly elongated shape of the clusters, as shown in Figure 5.2. Elongated clusters are assumed to be better modeled by an elliptical normal distribution than by the voronoi-like partitioning of feature space that is induced by the nearest-centroid principle of K-Means. A critical parameter in both clustering algorithms is the choice of the number of clusters $k$. A common method to determine the best $k$ is to cluster the data for all $k$ in a selected range. The number of clusters can then be selected as the $k$ for which the improvements of an internal validation measure stagnates. Ideally, the criterion peaks at a given $k$, which avoids the subjectivity of visual assessment or the definition of a quantitative stagnation criterion. To evaluate in how far the correct number of clusters can be found without prior knowledge, the data was clustered in the range of $k \in [2, 16]$ clusters. For both algorithms 30 initializations were chosen and a maximum of 100 iterations. These values were found to give stable results preliminary experiments.

### 5.3.2. Quantitative Results

The overall clustering quality in terms of Adjusted Mutual Information (AMI) for $k = 9$ is shown in Figure 5.5. The median AMI of all clusterings under PCA is 92.7%. The other external validation criteria, Adjusted Rand Index and V-Measure [RH07], were found to be strongly and significantly correlated ($\rho > 0.95, p \ll 0.001$), regardless of the chosen parameters. A visual comparison of the external validation criteria also suggests that all three measures behave very similar. The subsequent discussion uses the Adjusted Mutual Information (AMI) as external measure of clustering quality.

The most striking difference in quality exists between the dimensionality reduced and original feature sets. Applying PCA results in consistently better clustering results. Secondly, the Gaussian Mixture Model always performs at least as good as the K-Means algorithm. Considering the measure selection, discarding the maximum line lengths from
the feature set does not consistently improve the results. However, this observation will be revised when observational noise is present in the data.

The quality of the clustering depends on the correct choice of the number of clusters. Therefore, the practical applicability of the method also depends on the degree to which a suitable number of clusters can be estimated without prior knowledge. Figure 5.6 shows the distribution of the mean Silhouette Scores depending on the number of clusters. Compared to the Dunn Index and the Davies-Bouldin Index, the mean Silhouette Score was found to be most robust and conclusive and is thus chosen as internal validation criterion in the subsequent discussion. For the clusterings without dimensionality reduction, estimating the correct $k$ based on internal validation is not possible. In the PCA case, the stagnation of the Silhouette Score provides moderate evidence for $9 \leq k \leq 12$. Furthermore, when applying PCA, the internal validation matches the external validation much better. The strongest divergence between internal and external validation occurs in the case of using GMM without dimensionality reduction. Here, the internal validation results in much better scores for the Convergent feature set, whereas the external validation clearly favors the Standard feature set. An explanation of the unclear results in this case might be that the Silhouette Score is based on euclidean distances between objects and their cluster centers. The K-Means algorithm minimizes these distances by design, whereas the GMM assesses the distance of an object to its cluster based on the covariance of the cluster. This is related to the Mahalanobis distance, which can differ strongly from the euclidean distance for elongated clusters. The conclusion would be that the dimensionality reduction makes the clusters more spherical. While in line with Figure 5.3, such interpretations should be made with care, since in general, clustering results are influenced by several, partly very subtle factors. Furthermore, it is not obvious how well the lower-dimensional projection in Figure 5.3 reflects the true 16-dimensional shape of the cluster. Nonetheless, the superior performance of
the Silhouette Score and the poor performance of the Dunn Index are in line with the systematic study of internal validation measures in [BC12].

The maximum line length features clearly contribute to variance of the data set, as shown by Figure 5.3. Thus, they have potential to be weighted stronger in the first Principal Components, although they are hypothesized to negatively affect the clustering quality. A negative impact of applying PCA to the Standard feature set was however empirically not confirmed.

6. Influence of Noise and Embedding

Real-world data sets typically suffer from observational noise and often, only a single dimension of the phase space can be observed. In this section, the influence of observational noise and phase space reconstruction on the clustering results is evaluated. The clustering was repeated under nine noise levels and for three embedding methods, as summarized in Table 6. The Time Delay Embedding option was evaluated only for noise ratios 0 and 0.1, because the results are identical to the Single Dimensions results for noise ratios > 0.

6.1. Artificial Noise Model

The proposed noise model simulates a multivariate, normally distributed observational error. Each phase space trajectory point is displaced by a random perturbation vector
with independently and identically distributed components. This can be interpreted as measurement error, which manifests itself as difference between the true and the observed location of a point. The standard deviation of the components is chosen such that the expected magnitude $E[d]$ of the perturbation vector is $r\varepsilon$, where $\varepsilon$ equals 5% of the trajectory diameter and $r$ is called noise ratio. This formulation is inspired by the work of Thiel et al. [TRK+02], where the noise ratio is measured in units of the recurrence threshold $\varepsilon$. Here, the noise ratio is specified relative to 5% of the diameter of the original trajectory, rather than the selected threshold. This avoids circular dependencies when choosing $\varepsilon$ according to the diameter criterion. The problem would be that $\varepsilon$ depends on the trajectory, but the distortion implied by a given noise ratio, and thus the trajectory, would depend on $\varepsilon$.

Under the proposed noise model, a noise ratio of $\leq 20\%$ corresponds to moderate distortions. A noise ratio of 80% leads to strong deformations of the trajectory and the line structures in the RPs. The impact of different noise ratios on the trajectories, RPs and RQA measures is illustrated in Figures C.4, C.5, C.6, and C.7. The distortions introduced by the noise model were considered visually plausible by Dr. Marwan.

### 6.1.1. Alternative Models

A simpler noise model would be to randomly change elements in the recurrence matrix. This model would guarantee equal affection of systems in terms of an equal average number of errors in the RPs. However, this noise model is considered unrealistic, since the defects do not depend on the geometry of the phase space trajectory. But two distant points should be less likely to be falsely considered recurrent, than two nearby points. As such, this model might be appropriate for errors that could arise during an unreliable data transmission, but it does not seem to be particularly useful in the context of Recurrence Analysis.

Another alternative would be the model of Thiel et al. [TRK+02]. A problem here is that Thiel et al. focus on univariate trajectories and their embeddings, but not directly on multivariate trajectories. A natural relationship between a multivariate trajectory and a univariate trajectory is given through time delay embedding. A multivariate trajectory $(x_{1,1}, \ldots, x_{1,d}), \ldots, (x_{n,1}, \ldots, x_{n,d})$ of length $n$ and dimensionality $d$ can be

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Choices</th>
<th>Cardinality</th>
<th>Number of Clusterings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise ratio</td>
<td>0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8</td>
<td>9</td>
<td>16,200,000</td>
</tr>
<tr>
<td>Embedding Method</td>
<td>Original Trajectory, Time Delay Embedding, Single Dimension</td>
<td>3</td>
<td>36,600,000</td>
</tr>
</tbody>
</table>
written as a univariate trajectory \( x_{1,1}, \ldots, x_{n,1}, \ldots, x_{1,d}, \ldots, x_{n,d} \) of length \( nd \). Embedding this de-embedded trajectory using a dimension of \( d \) and a time lag of \( \tau = n \) will result in the original multivariate trajectory. Thiel et al. have extended their univariate noise model to embedded trajectories. Applying this noise model to a de-embedded multivariate trajectory is considered its natural multivariate extension. It was however noted that high-dimensional systems will be more affected by noise. This can be shown by calculating the expected magnitude of the perturbation vector. Assume without loss of generality, that perturbation in each of \( k \) dimensions is drawn independently and identically distributed from a standard normal distribution. The expected magnitude of the perturbation vector is the square root of the sum of squares of \( k \) standard normal random variables. The author was pointed to the fact that this magnitude follows the \( \chi(k) \)-distribution, where \( k \) is the number of degrees of freedom, or dimensions. Mathematical details are given in Appendix A. Calculating the expected value for \( k = 1, 2, 3 \) gives \( E[\chi(1)] \approx 0.8 \), \( E[\chi(2)] \approx 1.25 \), and \( E[\chi(3)] \approx 1.6 \). Since the expected magnitude of the perturbation vectors is larger in higher dimensions, the expected difference between two perturbation vectors is also assumed to be larger. Thus, higher dimensional systems should be more strongly affected by the noise model. However, showing this analytically turned out to be complicated. In this respect, Thiel et al. also moved from euclidean to maximum norm to simplify the evaluation of the involved integrals and used some approximations. Basically, the errors depend on the magnitude and angle of the perturbation vectors and on the distribution of distance vectors between pairs of points on the trajectory. Nonetheless, the independence of the perturbation vector magnitudes from the dimensionality of the trajectory is considered an advantage of the proposed noise model.

### 6.1.2. Equal Impairment of Classes

Contradicting the proposed noise model, one could argue that making more measurements will naturally result in increased error rates, and thus, higher-dimensional systems should be more influenced by noise. This raises the question of what requirements an artificial noise model should fulfill. From the evaluation point of view, any unequal treatment of systems seems debatable. It was thus assumed that the RQA measures of each system should be comparably affected by noise to avoid spurious effects due to the noise model. However, elaborating on the subtleties of the implications, it was found that the question whether all systems should be equally impaired by noise, and what that exactly means, seems to be best answered implicitly, by specifying a realistic noise model. But a realistic noise model must take into account the processes that generate the data, which depend on the application. As a generic model for measurement error, the multivariate perturbation model was considered an acceptable choice in that respect. However, since this is a generic artificial noise model, one should be cautious about spurious effects and as such, the working hypothesis of approximately equal impairment of RQA measures was adopted again.

Figure C.7 shows the impact of different noise ratios on the distributions of the RQA measures. Comparing the plots for each measure, two indicators of equal impairment
can be seen. Firstly, most measures stop changing for a noise ratio of 80\% and higher. Secondly, the trend is usually the same among all classes, for instance DET and $L_{\text{entr}}$ decrease with increasing noise, regardless of the class. An exception is the class of normally distributed noise, which is not affected by the chosen noise model. Regarding the RPs, noise ratios $\geq 160\%$ lead to hardly distinguishable RPs, and at $\geq 320\%$ almost all RPs resemble normally distributed noise, again regardless of the class. Compare Figure C.5 for an illustration of the distortion caused by different noise ratios on different classes.

Regarding the equal impairment criterion, the following subtleties were encountered. Firstly, the noise model cannot be derived in isolation of the remaining processes. Its interrelation with e.g. the chosen feature set is for instance demonstrated by the fact that $W_{\text{entr}}$ was found in general to be more robust to noise than $L_{\text{max}}$. Depending on which features contribute most to the profile of a given class, the susceptibility to noise might differ across classes. Even if the feature set comprises only a single feature, the different geometries of the phase space trajectories can cause different responses to the noise model. This is analytically shown in the work of Thiel et al., where the rates of false positive and false negative recurrence points are derived based on the density of pairwise distances of trajectory points. As an empirical example, Figure C.7 shows that $W_{\text{mean}}$ of the R trajectories decreased under noise whereas it increased for LO trajectories. The only controlled difference between these trajectories is the generating system, thus, the effect should be due to the different phase space geometry of the trajectories.

6.1.3. Dependence on Dimensionality

In spite of the magnitude correction of the perturbation vector with respect to the dimensionality of the trajectory, the noise model was found to react to the dimensionality of the trajectories. The experimental conditions of 640\% and 1280\% noise were included only to verify the assumption of a total collapse of clustering structure under very high noise ratios. This was empirically not confirmed. Up to seemingly arbitrary amounts of noise, a three-cluster structure is maintained, which is formed by the one-dimensional (AR-, AR+, N, NG), two-dimensional (OSC) and three-dimensional (R, RF, L, LO) trajectories. This is subsequently referred to as persistent clustering structure problem. The separation of the classes is attributed to the dimensionality of the trajectories, because their dynamics should be too obscured at these noise levels to explain the separation. Indeed, according to a visual inspection, all RPs resemble normally distributed noise under a noise ratio of 1280\%. Additional evidence is provided by the fact that the clustering quality indeed drops to zero when reducing the dimensionality of the trajectories to one, as described in Section 6.2.2.

The persistent separation is explained by the consistently different Recurrence Rates, as shown in Figure 6.1. Note that high Recurrence Rates also increase the formation of longer diagonal lines by chance. In a similar way, the vertical line length distributions are affected by higher Recurrence Rates. Thus, differences in the Recurrence Rates tend to propagate to other RQA measures for noisy trajectories. The propagation can be e.g. seen by focusing on the measures RR, DET, L, LAM, and TT in Figure 6.1.
There is strong evidence that the persistent separation is an interaction effect of the 5% diameter criterion for selecting $\varepsilon$ and the proposed noise model. Again, an analytical formulation is difficult, but the diameters of the trajectories were empirically found to suffer from consistently different expansion rates, depending on their dimensionality. For instance, the one-dimensional trajectories under a noise ratio of 12.8 exposed $\approx 12$ times the diameter of their noise-free equivalents. The two-dimensional trajectories exposed an expansion factor of $\approx 9$ and the three-dimensional trajectories an expansion factor of $\approx 7$. An intuitive explanation is that the probability of drawing approximately orthogonal perturbation vectors, which is necessary to maximally expand the diameter, reduces in higher dimensional spaces. An analytical derivation of the expected expansion rate is difficult because one needs to consider the distribution of distances and angles between trajectory points. In the second step, a stronger expansion of the trajectory diameter in combination with the diameter criterion for threshold selection is expected to yield higher Recurrence Rates.\(^{14}\)

The persistent separation under high noise ratios raises the question whether the noise model is flawed and, more importantly, whether the effect biases the clustering results at lower noise ratios. It is argued that this is not the case, since the observed spurious behavior supports a three-cluster solution, whereas the method usually produces much more differentiated clusterings. Further support is provided by Figure 6.2, which shows that there is a difference in the Recurrence Rates of the classes already at 0% noise. Furthermore, these relative differences are only marginally influenced by the noise model up to a noise ratio of 80%. Only for very high noise ratios, the Recurrence Rates are dominated by the influence of noise.

\(^{14}\)The exact influence on the Recurrence Rate depends on the distribution of the pairwise distances between the points on the trajectory.
6.1.4. Clustering Quality under Observational Noise

Figure 6.3 shows the distance matrix and its iVAT transform for the RQA measures under 80% noise. The observed amount of separation is highly remarkable, since at this noise level, almost all diagonal and vertical lines have already vanished, cf. Figure C.6. Note that for distinguishing different systems, the RQA measures do not have to be correct, they just have to change consistently within classes and maintain a certain separation.

An interesting effect is that clustering quality increases consistently up to 40% noise. This is explained by a smoothing effect that reduces the variance OSC class. Figure 6.4 illustrates this in the space of the first two Principal Components at 0% and 80% noise. In this projection, the RQAs measures of the OSC trajectories expose no more variance than the other classes.

An important observation is that the RQA measures change gradually and in a systematic way with increasing noise ratio. This supports the selection of noise ratios, and improves confidence that the influence of different noise ratios can be reliably assessed based on the results. If the effect on the RQA measures would be less systematic, one could not be sure whether the results are representative, or whether different effects occur in other parameter ranges for the noise ratio. However, given the smooth evolution of RQA measures under noise, this seems highly unlikely. Figure C.8 shows a scatter plot of DET and $L_{\text{entr}}$ that gives an impression how the clusters move in feature space.

The clustering quality under observational noise is plotted in Figure 6.5. The main finding is that, although the RQA measures change strongly under noise, their separation is maintained. Specifically, good clustering quality was achieved up to 80% noise. For the PCA case, $k = 9$ and a noise ratio of 0.8, the median AMI is still 82.6%, before it drops to a median of 61.7% for a noise ratio of 1.6.

The clustering results under moderate noise ratios of 0.1, 0.2, and 0.4 even outperform the results under noise-free conditions. The PCA transformed feature sets give close to
Figure 6.3: Left: Distance matrix of the euclidean distances between RQA measures under 80% noise. Right: The iVAT transform of the distance matrix.

Figure 6.4: Scatter plots of the first Principal Components of the data under 0% and 80% noise. The variance of the OSC RPs is strongly reduced under 80% noise.
optimal results. Regarding the other experimental parameters, most observations still hold. The GMM still performs at least as good as the K-Means algorithm. Concerning measure selection, the Convergent feature set shows mixed influence. In the PCA case, the Convergent feature set consistently outperforms the Standard feature set, but in case of applying GMM without PCA, it often stays behind. With respect to the internal validation measures, the correct number of clusters is indicated by the mean silhouette score up to about 40% noise, but again only in the PCA case.

### 6.2. Influence of Embedding

When working with real-world data, often only a single component of a trajectory is available. The method of Time Delay Embedding (see Section 2.1.2) can be used to approximately reconstruct the full trajectory. Alternatively, the RP can be computed directly from the single component, as advocated by [IB98]. The primary focus of this thesis is to evaluate the separability of RPs according to their RQA measures, irrespective of the quality with which the underlying theoretical aspects are captured. In this context, the question of which option is favorable may have a different answer. The main question is what clustering quality can be achieved when only one dimension of the trajectories is available.
6.2.1. Estimation of Embedding Parameters

The option Time Delay Embedding in Table 6 refers to computing the RPs from the embedded x-component of the trajectories. The embedding delay $\tau$ and the embedding dimension $m$ were estimated using the Mutual Information algorithm and the False Nearest Neighbors algorithm, respectively. The implementations in the TISEAN software package [HKS99] were used. The estimated delay was used as input for the False Nearest Neighbor algorithm.

Both algorithms do not give the desired statistics directly. A suitable embedding delay is often selected as the delay that results in the smallest local minimum of the mutual information. [MW14] As the maximum delay, half of the trajectory length, i.e. 12,500, was specified. The resulting embedding delays have a median of 2020. Too large delays are problematic because the measurements will be only loosely related, for instance as a result of the Butterfly Effect. The large delays might be due to the simplistic definition of a local minimum, which takes into account only the mutual information at the preceding and the succeeding delay. Using this definition, small, possibly even random fluctuations will be regarded as local minima. When visually assessing the mutual information curve, such small fluctuations would be naturally filtered out. However, incorporating a more sophisticated strategy for detecting local minima is left as possible future work. As a simple alternative, the first local minimum of the mutual information function was chosen, which leads to a median $\tau$ of 15. The estimated delays are less scattered within classes and lead to better estimations of the embedding dimension. As shown in Figure 6.6, the estimated dimension match the true dimensionality of the trajectories well. In contrast, using the smallest local minimum, almost all trajectories are considered one-dimensional.

For the estimation of the embedding dimension, another parameter was introduced. The usual criterion is to select the smallest embedding dimension for which no more false nearest neighbors appear. However, for noisy trajectories, a small amount of false nearest neighbors usually persist. In these cases, a dimension for which the false nearest neighbors almost disappear, must be chosen. As a simple formalization of this criterion, the smallest embedding dimension was chosen for which the amount of false nearest neighbors is below 5%. If the amount does not drop below this threshold for $m \leq 10$, the system is considered to be noisy and a dimension of 1 is assumed. This is also a precaution to avoid high-dimensional embeddings of noise, which can lead to spurious structures in the RPs. [MCRTK07] While this lead to good dimensionality estimations in the noise-free case, it did not show satisfying results on noisy trajectories. Already for a noise ratio of 10%, all trajectories are estimated to be one-dimensional, because the amount of false nearest neighbors remains above 5%. Choosing a threshold based on the data set was avoided to prevent overfitting. Also, experiments suggest that a more sophisticated criterion than a threshold is necessary, because for higher noise ratios, the base levels of false nearest neighbors are also higher.

\[\text{Specifically, the program mutual and the program false_nearest as of version 3.0.1.}\]
\[\text{The threshold was considered reasonable by Dr. Marwan.}\]
Figure 6.6: Estimated embedding delays and dimensions according to the first and the smallest local minimum of the mutual information function.
An alternative to embedding is to compute the RPs directly from the univariate trajectories. The option Single Dimension in Table 6 refers to this approach, based on the x-component of the trajectories. Note that the RPs of the one-dimensional systems will be the same as when computed according to the Original Trajectory option. The RPs under the Time Delay Embedding and Single Dimension options are also the same for a noise ratio $\geq 10\%$. This is because the embedding delay becomes irrelevant for an embedding dimension of one, as it was estimated for all trajectories under observational noise. This is not desirable with respect to the initial goal to comparing both options. However, no simple solution to the problem was found. Adding the requirement $m \geq 2$ lead to increased scattering of the RQA measures, outliers and split-up of clusters into two or more clusters. Hence, this option was regarded impractical.

### 6.2.2. Clustering Quality for Univariate Trajectories

The results under the Time Delay Embedding and Single Dimension options are highly correlated for the case of PCA and noise-free trajectories ($\rho > 0.98$, $p \ll 0.001$). As described above, they are the same for a noise ratio $> 0$. Thus, the discussion of embedding influence focuses on the Single Dimension option. The results are shown in Figure 6.7.

![Figure 6.7: Distributions of AMI scores for Single Dimension trajectories in dependence of the noise ratio.](image)
In the experiments, the seed of the random generator that generated the observational noise was not fixed. This gives slightly different RQA measures for the Single Dimension option and Time Delay Embedding option with $m = 1$, although they should be identical. Fixing the seed verified the source of variation. However, the clustering was not repeated because the issue affects only the comparison between embedding methods. The results of applying different embedding methods should be a paired observation on an entire data set (the 900 RPs that result from a fixed noise ratio). Ideally, only the embedding method would have been varied, but here, also the seed of the observational noise was varied. However, the influence was found to be small, such that it is expected to be well within the limits of uncontrolled variance that would also arise when working with real world data. For 99% of the clusterings, the absolute difference in AMI is < 0.13. For 90% of the clusterings, the absolute difference is < 0.04.

With this in mind, the Single Dimension option performs nearly as good as the Original Trajectory option in Section 6.1.4. In fact, the main difference is that the persistent clustering structure problem (cf. Section 6.1.3) vanishes. Additionally, the beneficial influence of noise is limited to noise ratios of 0.1 and 0.2.

In summary, working with the Single Dimension option gave good results. The K-Means algorithm performed well in terms of cluster quality, when used in conjunction with PCA. Additionally, it was faster than the GMM and lead to less variation in the clustering quality. Using the Convergent feature set improved the results under a wide range of parameters. These observations were confirmed for shorter trajectories by repeating the clustering under selected parameters. The results were similar, even for trajectories with as few as 500 observations per trajectory (instead of 25,000); see Figure C.10.

7. Extending Recurrence Quantification Analysis

7.1. Lines of Definite and Indefinite Length

There is a subtlety in the computation of the line length histograms. The length of lines that are adjacent to the border of the recurrence matrix is unknown, since at least one of their endpoints is located outside the recurrence matrix. These lines are subsequently referred to as indefinite, and all others as definite. Following the definition in [MCRTK07] (see Equation 2 and following), indefinite lines should be ignored. However, many implementations that compute RQA measures count these lines. A simple test case that demonstrates this is shown in Figure 7.1, and the results for different implementations are listed in Table 7. The chosen RP contains only indefinite diagonal lines and an indefinite white vertical line of length 215. The web application [MD] is excluded because of obviously wrong results, such as DET > 1. As the measures $L_{\text{max}}$, L, and $L_{\text{entr}}$ show, all implementations count the indefinite diagonal lines. PyRQA and pyunicorn also count the indefinite white vertical lines, whereas the Command Line RPs implementation does not. For long time series, the number indefinite lines is usually small. For short input data however, indefinite lines might cause
<table>
<thead>
<tr>
<th>Implementation</th>
<th>Reference</th>
<th>RR</th>
<th>DET</th>
<th>$L_{\text{max}}$</th>
<th>L</th>
<th>$L_{\text{entr}}$</th>
<th>$W_{\text{max}}$</th>
<th>$W_{\text{mean}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyunicorn</td>
<td>[DHR13]</td>
<td>0.011</td>
<td>1.0</td>
<td>104</td>
<td>78</td>
<td>0.6931</td>
<td>215</td>
<td>62.3</td>
</tr>
<tr>
<td>Matlab CRP Toolbox</td>
<td>[Mara]</td>
<td>0.011</td>
<td>1.0</td>
<td>104</td>
<td>78</td>
<td>0.6931</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Command Line RPs</td>
<td>[Marb]</td>
<td>0.011</td>
<td>1.0</td>
<td>104</td>
<td>78</td>
<td>0.6931</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>PyRQA</td>
<td>[RSMD14]</td>
<td>0.011</td>
<td><strong>0.59</strong></td>
<td>104</td>
<td>78</td>
<td>0.6931</td>
<td>215</td>
<td>62.7</td>
</tr>
<tr>
<td>recurrence-plot.tk</td>
<td>[MD]</td>
<td><strong>0.03</strong></td>
<td><strong>1.03</strong></td>
<td>-</td>
<td><strong>77.04</strong></td>
<td>0.73</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>present author</td>
<td></td>
<td>0.011</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

The lengths of the indefinite lines are then recorded in the histograms $P_{L}^{\ast}, P_{V}^{\ast},$ and $P_{W}^{\ast}$. To define the histograms of indefinite diagonal and vertical line lengths, it is convenient to define an extended recurrence matrix $\mathbf{W}$ such that $\mathbf{W}_{i,j} = R_{i,j}$ for $1 \leq i, j \leq N$ and $\mathbf{W}_{i,j} = 0$ elsewhere. Loosely spoken, this adds a “white” border around the RP. Using this definition, $P_{L}(\mathbf{W}, l)$ is the histogram of definite and indefinite diagonal lines. By subtracting the count of the definite diagonal lines, the counts of the indefinite lines are obtained. To define the white vertical lines, an analogous matrix $\mathbf{B}$ is defined as...
\( B_{i,j} = R_{i,j} \) for \( 1 \leq i, j \leq N \) and \( B_{i,j} = 1 \) elsewhere. The number of indefinite lines is then given as:

\[
\begin{align*}
P^*_L(R, l) &= P_L(W, l) - P_L(R, l) \\
P^*_V(R, l) &= P_V(W, l) - P_V(R, l) \\
P^*_W(R, l) &= P_W(B, l) - P_W(R, l)
\end{align*}
\]

Within this framework, the definitions of DET and LAM can be modified to be compliant with both the software implementations and the formal definitions. The length of the indefinite lines can be lower bounded by the length of the part of the line that lies within the recurrence matrix. Since the filter ratio depends on a lower bound on the line lengths, the indefinite lines can be included in the computation of DET and LAM.

\[
\begin{align*}
RR &= \frac{1}{N^2} \sum_{l=1}^{N} l(P_L(l) + P^*_L(l)) \\
DET &= \frac{\sum_{l=l_{\text{min}}}^{N} l(P_L(l) + P^*_L(l))}{N^2 RR} \\
LAM &= \frac{\sum_{l=l_{\text{min}}}^{N} l(P_V(l) + P^*_V(l))}{N^2 RR}
\end{align*}
\]

All RQA measures in this thesis were computed based on a custom implementation that uses only definite lines, except for the filter ratios, as in equations 6 and 7. However, since the RPs are large, the results should be almost identical to the results of other implementations.

### 7.2. Line Types and Descriptors

Through the course of the work, the definition of RQA measures have been examined for extension opportunities. From a Computer Science point of view, there are two evident starting points for defining further RQA measures. Firstly, the definition of further line types and secondly, developing other descriptors for the line length histograms. Experiments regarding both were performed, but showed limited success. As another line type, the diagonal lines that run orthogonal to the main diagonal were counted. In Recurrence Analysis, these correspond to the palindromic evolution of a trajectory. [MW14] They are sometimes used as an indicator of an insufficient embedding. [MCRTK07, p. 251] The orthogonal line length histograms were described using the standard statistics. The filter ratio and the entropy of the orthogonal lines were found to distinguish the chaotic systems in the artificial data set. However, adding them to the feature set did not improve the clustering quality.

A second approach is to describe further aspects of the line length histograms. For example, sparseness was defined as the fraction of empty bins up to the maximum line length. Definiteness was defined as the fraction of points on definite lines. The difference entropy was defined as the entropy the discrete derivative \( P'(l) = P(l + 1) - P(l) \) of the histogram. The qualitative evaluation of the distributions of these measures was
Figure 7.2: Recurrence Plot of a trajectory of the Rössler system which shows three UPOs, which correspond to the peaks on the main diagonal of the CRT Plot.

not promising. Either there were not responding to class membership, exposed strong scatter, have not been robust to noise or depended strongly on the embedding method. Note that the orthogonal entropy performed well with respect to these requirements. Finally, as an alternative to the average line length, the median line length was found to converge much faster in the experiments described in Section 5.1.2. Consequently, replacing the maximum line length with the 95th percentile might be an option, although it is unclear which percentile to chose.

### 7.3. Conditional Recurrence Time

An Unstable Periodic Orbit (UPO) is a phase in which a subsequence of the trajectory is repeated for a limited number of times. This leads to evenly spaced diagonal lines in the RPs. UPOs are an important aspect of chaotic dynamical systems, since they can be considered the building blocks of their trajectories [MCRTK07, p 301]. This is especially clear in the case of the Lorenz System, where the trajectory jumps from one scroll to the other after having stayed in its orbit for some time. As a simple method to detect UPOs, the Conditional Recurrence Time (CRT) is proposed. Three or more evenly spaced diagonal lines will lead to an increased number of consecutive white vertical lines of equal length. For brevity, let the term line, when used without further qualification, refer to white vertical line in this section. Given an RP, the CRT is defined as the 2D histogram $C(l_2 \mid l_1)$ of consecutive lines. This corresponds to the empirical probability of finding a line of length $l_2$, given that the previous line had length $l_1$. The pairs of lines...
of equal length will contribute to regions around the main diagonal in the histogram, ideally leading to a peak corresponding to the period of the UPO. An example for the Rössler System is shown in Figure 7.2a.

The RP shows three UPOs with periods of roughly 70, 140 and 210, respectively. These UPOs correspond to the peaks on the main diagonal of the plot of the CRT histogram. In the CRT histogram, the vertical axis corresponds to the length $l_1$ of the line and the horizontal axis refers to the length $l_2$ of the consecutive line. Note that binning is not necessary, since the line lengths are discrete. However, trimming the histogram to a maximum line length is useful, because rare combinations of line lengths strongly increase the limits of the histogram. In general, log-transforming the histogram was found to be convenient for visualization, because the bin counts can vary by several orders of magnitude. The displayed values are $\log_{10}(C(l_2 \mid l_1) + 1)$.

The proposed technique was found to produce distinctive patterns for the different chaotic systems, as shown in Figure 7.4. These patterns were found to be consistent within classes and non-redundant with regard to the white vertical line length distributions shown in Figure C.11. Additionally, the CRT plot was found to be very robust against noise and choice of embedding method. Figure 7.3 shows a trajectory of the Lorenz System under 80% noise and its CRT plot. Although the structures in the RP are strongly distorted, the CRT plot shows a picture similar to that in Figure 7.4c.

To assess the use of the techniques in the context of clustering, the compression distance was applied to the CRT plots of the artificial data set. However, the results were not satisfactory. As an alternative approach, statistics of the CRT histogram were computed. In the style of RQA measures, the average $C_{\text{mean}}$ and the entropy $C_{\text{entr}}$
Figure 7.4: Log-transformed CRT plots for trajectories from the artificial data set. The histograms were trimmed to a line length of 500, the values outside these limits decay quickly.
of the bins was computed. Furthermore, the bin $C_{\text{mode}}$ with the highest frequency was computed and the correlation $C_{\text{corr}}$ between $l_1$ and $l_2$. In this context, the correlation will be high if many consecutive lines are of equal length, i.e. the data concentrates around the main diagonal of the histogram. The number of local maxima $C_{\text{lm}}$ was defined as the number of bins that are larger than all bins in their 11x11 neighborhood. The size of the neighborhood was found to give a good separation of the systems but a heuristic for choosing the size automatically would be desirable.

$$C_{\text{mean}} = \left( \frac{l_1}{l_2} \right) = \frac{1}{\sum C} \sum_{l_1=1}^{N} \sum_{l_2=1}^{N} C(l_1 \mid l_2) \left( \frac{l_1}{l_2} \right)$$

$$C_{\text{mode}} = \left( \frac{l_1^*}{l_2^*} \right) = \arg \max_{l_1^*,l_2^*} C(l_1^* \mid l_2^*)$$

$$C_{\text{entr}} = -\sum_{l_1,l_2} p(l_1 \mid l_2) \ln p(l_1 \mid l_2)$$

$$C_{\text{corr}} = \frac{\sum_{l_1=1}^{N} \sum_{l_2=1}^{N} C(l_1 \mid l_2)(l_1 - \overline{l_1})(l_2 - \overline{l_2})}{\sigma_{C(l_1)} \sigma_{C(l_2)}} \sum C$$

$$C_{\text{lm}} = \left| \{(l_1,l_2) \mid P(l_1 \mid l_2) > P(l_1 + i \mid l_2 + j) \text{ for } -5 \leq i,j \leq 5 \text{ and } -1(i = 0 = j) \} \right|$$

These statistics were found to work well both on the Original Trajectory and Single Dimension options, as well as being robust with respect to noise. Considering only the statistics of the CRT histograms, a good separation between the classes of the artificial data set can be achieved, as shown in Figure C.13. Note that using the mean, maximum and entropy of the traditional white vertical line lengths does not lead to a separation of the classes. As such, it is remarkable that a good separation can be achieved by considering the white vertical lines only. An in-depth study of the relationship to the dynamics of the system would be interesting, but should be approached from a Physics perspective. The off-diagonal peaks in the CRT plot seem to correspond to transitions between UPOs and could be another interesting aspect of the dynamics. Note for instance, that the CRT plot in Figure 7.4d is not symmetric, despite its seeming regularity. In the second row of the grid of peaks, only every other peak is present. Variations of the method that also consider the black vertical lines were experimented with, but the CRT was found to be most informative. Another obvious extension is to increase the dimensionality of the histogram to $k$ by counting $k$ consecutive lines. However, two consecutive lines already respond to the presence of UPOs while adding additional robustness compared to single lines. Two dimensional histograms also offer the advantage of simple visualization of the histogram. This suggests that the CRT plot could be used as a visual signature of the RP. Especially for the deterministic systems, the plots seem to capture salient information on much less space than the original RP. The size of that signature depends on the range of most frequent line lengths, but no longer on the trajectory length. For the artificial data set, most patterns were found in
a range of line lengths up to only 500. Limited screen size is a problem when working with large RPs, since interpolation of the images can lead to spurious patterns [Mar11]. The problem is aggravated when working with large numbers of RPs, which cannot be simply displayed side by side on the screen.

8. Conclusions and Future Work

In Chapter 4, an evaluation framework for RP similarity measures was created in cooperation with domain experts from the Potsdam Institute for Climate Impact Research. It consists of trajectories with different dynamics and a reference clustering of these trajectories. Within this framework, similarity measures for RPs were evaluated. It was found that the euclidean distance between vectors of RQA measures comply with an expert’s assessment. A geometrical and a temporal abstraction were identified that are hypothesized to explain the superior performance of the RQA measures. In Chapter 5, the Convergent feature set, a selection of RQA measures, is proposed, based on the observation of inferior convergence properties of maximum line length measures. The impact of feature set, dimensionality reduction and clustering algorithm on the clustering quality is evaluated. K-Means clustering of the PCA-reduced Convergent feature set was found to be a recommendable option, yielding AMI scores of > 0.9. In Chapter 6, the influence of observational noise and Time Delay Embedding was assessed. A variant of the noise model in Thiel et al. [TRK+02] was proposed and interaction effects with the dimensionality of the trajectories were discussed. The RQA measures were found to maintain good separation until 80% noise. Computing the RPs from only one dimension of the trajectories was found to have minor influence on the clustering quality. The Mean Silhouette Score was noticed to perform well and superior to other internal validation measures with regard to detecting the correct number of clusters. In Chapter 7, two extensions of RQA were proposed. Firstly, it was put forward to treat lines of known and unknown length separately, in an attempt to relieve a discrepancy between the definition of line structures and several RQA implementations. Secondly, Conditional Recurrence Time (CRT), an approach to quantify the structures in an RP based on consecutive white vertical lines, was proposed. Statistics of the CRT were noted to well characterize the dynamics in the data set, even if computed from only a single dimension of the trajectories and under noise.

Real World Data and Embedding Parameter Estimation An important task is to evaluate the discussed methods on real-world data sets. Indeed, several experiments were performed on a real world data set, the Bern-Barcelona EEG Database [ASR12]. However, none of the methods was able to reveal a clustering structure in the data. Some kinds of data are simply not suitable for applying Recurrence Analysis, but this should not be the case here. Recurrence Analysis has been used for several applications that use EEG data [FKC+14]. One of the statistical tests of the data set’s original paper [ASR12] is also based on Recurrence Analysis. One basic problem when working with the data set was to determine good parameters for computing the Recurrence Plots.
Several combinations have been tried, including RPs and CRPs of fixed and estimated embedding parameters, using different estimation methods. For individual time series, it was possible to create allegedly more informative RPs by optimizing the parameters manually. However, no single heuristic was found that gave visually convincing RPs for all time series. Regarding the recurrence threshold, the phase space diameter criterion was not very reliable on this data set. Computing the RPs according to a fixed RR might have been more appropriate here. Estimating the embedding dimension according to the False Nearest Neighbor method again seemed to suffer from observational noise, in accordance with the observations in 6.2.1. To explore large sets of time series from a Recurrence Analysis point of view, good heuristics for estimating recurrence parameters are needed. If a manual adaption to a given data set is inevitable, software support should be developed for assessing the influence of embedding parameters on a large data set. The proposed evaluation framework could be used to evaluate other heuristics for embedding parameters, especially under observational noise.

Further Development of Similarity Measures The strong scattering of RQA measures in the OSC class supports an expected sensitivity of RQA measures to variations of time scale. Multiscale applications, which naturally produce time series on difference time scales, probably require the development of similarity measures that are time scale invariant. Variation of time scale could be incorporated into the evaluation framework, to aid the assessment of such measures.

The promising results of the CRT histograms suggest that direct comparison of the line length histograms might have more potential than initially thought. In addition there seem to be important aspects of the CRT histograms that are not yet covered by the proposed statistics. For instance, finding and comparing the grid structures in the CRT histogram seems promising with respect to Figure 7.4. Texture descriptors might be another worthwhile approach to capture structures on still larger scales than CRT.

References


64


Appendices

A. Artificial Noise Model

In the following, the variance of normal distribution, which can be used to simulate a multivariate measurement error of expected magnitude, is derived. As usual, the measurement error is assumed to have zero mean. Furthermore, the expected deviation in each dimension is assumed to be equally large. Thus, the first step is to describe the expected magnitude \( E[d] \) of a vector drawn from a k-variate normal with zero mean and covariance \( \Sigma = \sigma I \), where \( I \) is the identity matrix of size \( k \). Since the distribution’s covariances are zero, this is equivalent to drawing each component \( x_i \) of the vector independently from \( N(0, \sigma^2) \). As is known, \( (x_i - \mu)/\sigma \sim N(0, 1) \), which can be used to express \( E[d] \) as the scaled mean of the \( \chi \)-distribution. The \( \chi \)-distribution with \( k \) degrees of freedom describes the distribution of the square root of the sum of \( k \) squared standard normal random variables.

\[
E[d] = E[\sqrt{\sum_{i=1}^{k} (\sigma x_i)^2}] = E[\sigma \sqrt{\sum_{i=1}^{k} x_i^2}] = \sigma E[\chi(k)] = \sigma \sqrt{2 \frac{\Gamma((k + 1)/2)}{\Gamma(k/2)}}
\]

Accordingly, to get an expected displacement of \( E[d] \), each component’s displacement must be drawn from a normal with standard deviation \( \sigma(E[d]) \).

\[
\sigma(E[d]) = \sqrt{2} E[d] \frac{\Gamma(k/2)}{\Gamma((k + 1)/2)}
\]

One of the central results in Thiel et al. [TRK+02] is an analytical expression for the probability to recognize a recurrence point for a given point of pairs on a univariate time series with gaussian noise applied. The probability follows equation 13.

\[
P_{i,j} = \frac{1}{8} \left\{ \text{erfc}\left(\frac{D_{i,j} - \varepsilon}{2\sigma}\right) - \text{erfc}\left(\frac{-D_{i,j} - \varepsilon}{2\sigma}\right) + \text{erfc}\left(\frac{-D_{i,j} + \varepsilon}{2\sigma}\right) - \text{erfc}\left(\frac{-D_{i,j} - \varepsilon}{2\sigma}\right) \right\}
\]

(13)

Summing the two noise factors into a single effective noise factor \( \sim N(0, \sqrt{2\sigma^2}) \), an equivalent, but simpler formulation can be derived by using the Cumulative Distribution Function of the effective noise factor:

\[
P_{i,j} = \text{CDF}(\varepsilon - D_{i,j}) - \text{CDF}(-\varepsilon - D_{i,j}) = \frac{1}{2} \left\{ \text{erf}\left(\frac{\varepsilon - D_{i,j}}{2\sigma}\right) - \text{erf}\left(\frac{-\varepsilon - D_{i,j}}{2\sigma}\right) \right\}
\]

(14)

B. Experimental Data

The enclosed data medium contains the input data and the experimental results. The RPs that were presented during the expert assessment of similarities are located in
Table 8: Results Database Overview

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>clustering_algorithm</td>
<td>Dimension</td>
<td>IDs for the clustering algorithms</td>
</tr>
<tr>
<td>data_set</td>
<td></td>
<td>IDs for the main, tiny and small data sets</td>
</tr>
<tr>
<td>dimensionality_reduction</td>
<td></td>
<td>IDs for PCA and No Reduction options</td>
</tr>
<tr>
<td>embedding_method</td>
<td></td>
<td>IDs for embedding options</td>
</tr>
<tr>
<td>l_min_method</td>
<td></td>
<td>Deprecated. IDs for the estimation method for $l_{\min}$</td>
</tr>
<tr>
<td>measure_selection</td>
<td></td>
<td>IDs for the Standard and Convergent feature sets</td>
</tr>
<tr>
<td>noise_ratio</td>
<td></td>
<td>IDs for the noise levels</td>
</tr>
<tr>
<td>number_of_clusters</td>
<td></td>
<td>IDs for the number of clusters</td>
</tr>
<tr>
<td>resampling_indices</td>
<td></td>
<td>IDs and index sets for the subsamples used in the clustering</td>
</tr>
<tr>
<td>rqa_measures</td>
<td>Fact</td>
<td>The RQA measures of the trajectories under noise and embedding</td>
</tr>
<tr>
<td>rqa_measures_clustering</td>
<td></td>
<td>The clustering label assignments and validation indices for all clusterings</td>
</tr>
</tbody>
</table>

**expert-assessment-RP-selection.** The raw trajectory data is located in *time-series*. The results in this thesis are based on experiments with trajectories of 25,000 observations. These are found in *corpus-100-25000*. The two data sets with shorter trajectories, which were used in Figure C.10, are located in *corpus-100-500* and *corpus-100-2000*. Most figures in this thesis are created from the MySQL results database that is provided as a dump in *db_wittcarl.sql*. The database is structured using a Star Schema, as summarized in 8. The dimension tables describe the experimental conditions, such as the selected clustering algorithm, noise level, embedding method, etc. The *rqa_measures* table record the computed RQA measures of the trajectories of the three data sets (25,000 observations, 500, and 2500 observations) under different noise ratios and embedding methods. The *rqa_measures_clustering* table contains the cluster label assignments and the internal and external validation scores for all clusterings. The majority of the Figures was created in R, fetching data from the provided database. The scripts are located in *r_scripts*. They require an RData file to be created in this folder, containing a variable *db_password* that gives access to the database.

**C. Data Figures**

This Appendix provides large figures and figures of full data sets, excerpts of which have been used in the text to highlight particular aspects of the results.
Figure C.1: Excerpts of selected Recurrence Plots of the artificial data set.
Figure C.2: Evolution of RQA measures of a trajectory of each system depending on trajectory length.
Figure C.3: The RQA measures of the artificial data set under 10% artificial noise.
Figure C.4: Effect of different noise ratios on a short trajectory of the Lorenz System.
Figure C.5: Effect of different noise ratios on a short trajectory of the Lorenz System.
Figure C.6: RPs for selected classes under a noise ratio of 80%.
Figure C.7: The impact of noise on the minimum, mean, and maximum of distribution of RQA measures.
Figure C.8: The impact of noise ratio on the location of the clusters in the DET, $W_{\text{entr}}$ subspace.
Figure C.9: All clustering results except for the Time Delay Embedding option.
Figure C.10: Clustering results for variations of the artificial data set with fewer than the original 25,000 observations per trajectory. Clustering conditions were selected according to the results in Sections 5.3.2, 6.1.4, and 6.2.2.
Figure C.11: White vertical line length histograms of selected trajectories. The histograms have been divided by the maximum bin count.
Figure C.12: Selected CRT statistics of the artificial data set under the Single Dimension option and 20% noise. The measures show good separation between classes. C2 Mode refers to $l_2^*$ in Equation 9, the column of the mode of the histogram.
Figure C.13: Distance matrix of all CRT statistics under the Single Dimension option and 20% noise. This shows that the classes have characteristic CRT histograms.

Figure C.14: Scatter plot of two CRT statistics under the Single Dimension option and 20% noise.
Selbständigkeitserklärung


Berlin, den November 5, 2015

.................................................................