Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

I. VEGA[1], CH. SCHÜTTE[1,2] AND T. CONRAD[1,2]

[1]*Department of Mathematics and Computer Science, Freie Universität Berlin, Germany*

[2]*Zuse Institute Berlin, Germany*

SAIMeR: Self-adapted method for the identification of metastable states in real-world time series

ZIB-Report 14-16 (May 2014)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

SAIMeR: Self-adapted method for the identification of metastable states in real-world time series

I. Vega[1], Ch. Schütte[1,2] and T. Conrad[1,2]

[1] *Department of Mathematics and Computer Science, Freie Universität Berlin, Germany*
[2] *Zuse Institute Berlin, Germany*

Abstract

In the framework of time series analysis with recurrence networks, we introduce SAIMeR, a heuristic self-adapted method that determines the elusive recurrence threshold and identifies metastable states in complex time series. To identify metastable states as well as the transitions between them, we use graph theory concepts and a fuzzy partitioning clustering algorithm. We illustrate SAIMeR by applying it to three real-world time series and show that it is able to identify metastable states in real-world data with noise and missing data points. Finally, we suggest a way to choose the embedding parameters used to construct the state space in which this method is performed, based on the analysis of how the values of these parameters affect two recurrence quantitative measurements: recurrence rate and entropy.

Keywords. Time series analysis, application in statistical physics, recurrence quantification analysis, threshold, metastability, non-linear dynamics

AMS subject classifications. 37M10, 62H30, 46N55

1 Introduction

The need to understand the dynamics of complex data coming from the biological, the financial, the environmental or the medical fields, has promoted the development of many visualization and analysis methods.

Some of the main problems these methods face arise from the high-dimensionality, non-linearity, noise or sparsity of measurements of the real-world data they analyze. As mentioned by van der Maaten and van den Herik [1], some of the linear methods — such as PCA or Classical Multi-dimensional Scaling — and non-linear methods — such as Stochastic Neighbor Embedding or Isomaps — used for this purpose, can have some drawbacks, like not preserving both local and global scale properties of complex data or depending on many undetermined parameters. These problems can lead to leaving large part of the analysis open to subjective interpretation.

One approach that gives information about the local, medium and global scales in high-dimensional, non-linear time series, is recurrences analysis.

The study of recurrences in measure preserving dynamical systems dates back to Poincaré's studies at the end of the nineteenth century. Such phase space studies led to the development of the concept of recurrence plot by Eckmann et al. in 1987 [2], which focused on high-dimensional phase space trajectories. Over the years, the study of recurrences moved from the qualitative to the quantitative analysis, which in turn led to the introduction of recurrence quantitative analysis by Zbilut and Webber in 1992 [3]. This allowed the analysis of non-linear, non-stationary time series data and broadened the concept of recurrence.

One of the problems of recurrence plots analysis is the selection of the parameter necessary to compute them: the recurrence threshold. The recurrence threshold controls how close two phase space trajectories, or state space vectors, should be in order to consider them as neighbors. Therefore, it determines the size of neighborhoods in phase space that can be associated with the existence of stable dynamical states.

More recently, Krishnan et al. in 2008 [4, 5] made the analogy of recurrence plots with graph theory and introduced the concept of recurrence network. In 2012, Donges et al. [6] introduced some graph theory concepts to the study of recurrence networks in order to address the problem of selecting a recurrence threshold appropriate to analyze time series with uniform probability density distributions. For these cases, they set bounds in terms of the critical edge density of a recurrence network.

However, selecting an appropriate recurrence threshold for real-world time series is still an open problem [7,8], due to some properties of these time series, like: a non-necessarily uniform probability distribution, frequently having noise or missing some measurement points, and showing *metastability* — a property of physical phenomena with multiple time scales in which some time scales are in equilibrium and produce the so called metastable states, while others are not.

In this paper we introduce a heuristic method, based on recurrence network analysis, which identifies different metastable states in real-world time series data. This method is called the *Self-adapted method for the identification of metastable states in real-world time series* (SAIMeR).

The main components of SAIMeR are: computing an appropriate recurrence threshold for the analysis of real-world time series with recurrence networks theory, identifying metastable states in real-world time series, and providing the possibility of identifying the transitions between these states due to the use of the MSM clustering algorithm [9–12].

This paper contains the detailed explanation of the construction of this method and illustrates its performance in the following way:

Section 2 presents the theoretical foundations of recurrence plots and recurrence networks: it explains the construction of the state space and the problem of selecting a recurrence threshold. Furthermore, it contains a brief review of network clustering theory.

Section 3 explains the two parts in which SAIMeR is divided (summarized in Algorithms 1, 2 and 3).

Sections 4 and 5 validate the ability of this method to identify metastable states in real-world time series in a robust way.

Section 4 illustrates the performance of SAIMeR in application to three time series showing metastability. The time series analyzed are: (1) the one-dimensional movement of a particle under the gradient of a double well potential and a random force, (2) the two-dimensional molecular dynamics of Trialanine, and (3) the one-dimensional real-world time series containing the average daily temperatures in Berlin from 1937 to 2010.

Fig. 1 contains the results of applying SAIMeR to the time series containing the average daily temperatures of Berlin, in the period from January 1st, 1942 to December 31st, 1943. Temperature time series are likely to have trends, possibly associated to climate change, and several missing measurement points during some periods of time (non-equally spaced measurements), possibly related to historical events. In Fig. 1, the gray-scale color code indicates three different groups of time points. Two of them, the metastable states, indicating broadly a warm and a cold season. The third group indicates the transition regime between warm and cold seasons.

The validation of the robustness of SAIMeR is contained in Section 5. For this purpose, we measure the similarity between the results obtained from two different time series: a *control* time series and a time series produced (a) by adding a percentage of noise or (b)

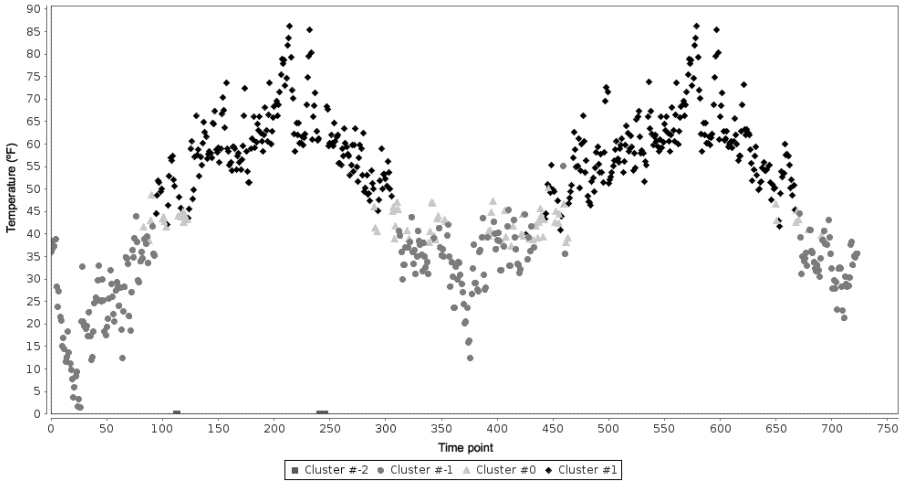


Figure 1: This figure shows the time series containing the daily average temperatures in Berlin (Tempelhof) from January 1, 1942 to December 31, 1943. The grayscale color code represents the different metastable states identified using $\varepsilon^* \simeq 0.2933$, $\tau = 2$ and $m = 2$. For more details see Sections 4.3 and 2.

by removing a percentage of data measurements from the *control* time series. These two features (noise and missing data points) are called *artifacts*.

Finally, in appendix A we suggest a way to determine the better suited embedding parameters for the construction of state space based on two RQA measurements: recurrence rate and entropy.

2 Background

The method introduced in this paper, SAIMeR, is based on recurrence analysis and network clustering analysis. Therefore, we will briefly introduce the main concepts of both theories.

2.1 Overview

Recurrence Plots were introduced by Eckmann et al. in 1987 [2] with the aim of understanding the dynamics of complex data sets. These are inspired by Poincaré's phase space studies [13]. A phase space, or state space, contains all the dynamical states of a system. Therefore, recurrence plots are computed in state space.

A recurrence plot is a tool to obtain meaningful empirical information from high-dimensional data sets which depends on few parameters. It is given by a square binary matrix, $R_{ij}(\varepsilon)$, which contains information about the recurrences of phase space trajectories, or state space vectors, to neighborhoods that can be associated with the existence of stable dynamical states. The size of these neighborhoods is given by the Recurrence Threshold, ε. This way, variations in the recurrence threshold will reveal different scales of structure in the state space.

In order to obtain not only empirical but quantitative information from a recurrence plot, Zbilut and Webber [3] introduced in the mid-nineties the so called Recurrence Quantitative Analysis (RQA) measurements. These measurements are computed from the recurrence plot of a given system and give information about its local, medium and global scales.

About a decade ago, Marwan et al. [14, 15] started studying the similarities between the geometry of the phase space and the RQA measurements. Their results identified recurrence plots as a convenient tool to analyze non-linear data [16].

In 2008, Krishnan et al. in 2008 [4, 5] introduced Recurrence Networks (RN), the graph representation of recurrence plots, which improved the intuitive and quantitative understanding of the dynamics of complex systems [17].

It is worth to notice that, since the structure of a recurrence plot depends on the recurrence threshold, the structure of a recurrence network will also do.

In a recurrence network, variations in the recurrence threshold will produce changes in its connectivity, modifying this way the size and number of its dense groups of interacting nodes, also known as clusters [11]. There are several algorithms that identify clusters in networks, focusing in different network's properties.

Recently, Donges et al. [6] introduced some graph theory concepts to the study of recurrence networks to address the problem of selecting an appropriate recurrence threshold. They set boundaries for the recurrence threshold in terms of the critical edge density of a recurrence network for the analysis of time series with uniform probability density distributions.

However, selecting an appropriate recurrence threshold for real-world time series is still an open problem [7,8] and the search for boundaries for the recurrence threshold in this case, led to the development of SAIMeR.

Keeping the previous considerations in mind, let us proceed to a more detailed explanation of the concepts behind SAIMeR.

2.2 The State Space

It is know that, when the time series of a dynamical system is embedded or mapped into a space of adequate dimension, this space contains all the dynamical information of the system, preserves determinism and creates a diffeomorphism for the attractors [18]. Therefore, the state space can be reconstructed with the appropriate embedding space for the time series.

The recurrence analysis of a time series is performed in state space. Therefore, constructing the state space is fundamental for a good recurrence analysis. But, how to do it?

The embedding space can be constructed either with the m time derivatives of the time series. However, when the computation of the time derivatives of the time series is not possible, one of the most common ways to build the state space is using the time delay embedding method or delay mapping.

The time delay embedding method is based on Taken's theorem of embedding [19] and requires the setting of two parameters: the embedding delay, τ, and the embedding dimension, m [18,20,21]. For a time series u_i of length N, the $N^* = N - \tau(m-1)$ state space trajectories $\overrightarrow{x_i}$ resulting from the time series are given by:

$$\overrightarrow{x_i} = (u_i, u_{i+\tau}, ..., u_{i+(m-1)\tau}), \text{ for } i = 0, ..., N^* \tag{1}$$

Time delay and embedding dimension can be determined through the geometrical, dynamical and topological analysis of a time series data [22].

In order to set the embedding delay, one must guarantee that the vector built with all the i-th entries of the state space trajectories is linearly independent from the vector built with all j-th entries of the state space trajectories, for all $i \neq j$. For periodic time series the embedding delay can not be a multiple of the period, in order to guarantee that the state space constructed does not contain more dimensions than necessary and therefore the state space trajectories do not intersect between each other.

The embedding delay can also be chosen in terms of the linear autocorrelation function or in terms of the average mutual information, which is a non-linear generalization of the

first and tells us how much information about $u_{i+\tau}$ we get when we observe u_i. Since two measurements are completely independent when the mutual information is zero, the time delay τ can be chosen as the one for which we obtain the first minimum in average mutual information. However, for some systems, the mutual information might not have a minimum. In these cases, a deeper analysis is required. For an extended discussion on how to determine the embedding time delay, see the article of Abarbanel in 1996 [21].

To set the embedding dimension, different geometrical, dynamical and topological tests can be used. The geometrical tests indicate the variations in distance between two close points when the embedding dimension increases, e.g. the computation of fractal dimensions or false nearest neighbors. The dynamical tests are used to select the embedding that provides a unique future for every data point, e.g. the implementation of predictability tests or the estimation of Lyapunov exponents. The topological tests look for the embedding dimension m that avoids intersections of stable periodic orbits. One-dimensional chaotic data, for example, have embedding dimension $m \geq 3$. Generally, for n-dimensional dynamical systems with fractal dimension d_A, the embedding dimension is $m > 2d_A$. Another general estimation given by Whitney et al. [20] states that $m < 2n$.

Different selections of embedding parameters will reconstruct state spaces with different dynamical information quality. The recurrences in these spaces will also vary and the structure of recurrence plots and networks will differ as well.

2.3 Recurrence Plots

The recurrence states of the state space reconstructed from complex, high-dimensional data sets can be identified with recurrence plots. A recurrence plot is defined in terms of a square binary matrix $R_{ij}(\varepsilon)$ containing information about the recurrences of state space trajectories $\overrightarrow{x_i}$ to a set of states:

$$R_{ij}(\varepsilon) = \Theta\left(\varepsilon - d(\overrightarrow{x_i}, \overrightarrow{x_j})\right) - \delta_{ij} \tag{2}$$

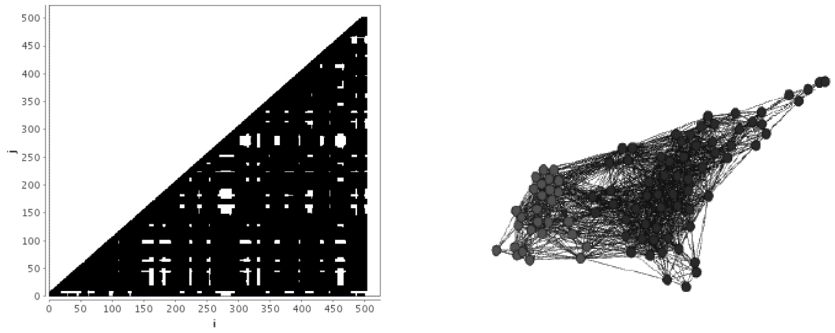
In this expression, $\Theta(\cdot)$ is a Heaviside function, $d(\overrightarrow{x_i}, \overrightarrow{x_j}) = d_{ij}$ is a metric and ε is the *recurrence threshold* – a cutoff distance that determines the size of a recurrence neighborhood –. Throughout this article, we will use the adequately scaled Euclidean metric, so that every variable of a time series is min-max normalized. The selection of norm implies that recurrence neighborhoods are hyperspherical. For a detailed explanation of the effects of choosing a different metric, see article of Donner et. al from 2010 [23].

In a recurrence plot, rows represent each of the state space vectors associated to the time series. This way, every entry (column) j of row i represents the closeness between state space vectors i and j.

2.4 Recurrence Networks

Every recurrence plot, $R_{ij}(\varepsilon)$, has an associated recurrence network, $G_{ij}(\varepsilon)$. In a recurrence network, every node represents one of the state space vectors associated to the time series and every edge represents the belonging of a pair of state space vectors to a same recurrence neighborhood. Due to the symmetry of recurrence plots, recurrence networks are unweighted, undirected and have the same number of nodes as the number of state space vectors built from the data set.

The information about the local, medium and global geometric properties of a system, can be recovered from the recurrence network through measurements based on neighborhoods or on paths. Donner et al. [24] have provided a summary of the definition and meaning of path- and neighborhood-based measurements for recurrence networks.



(a) Recurrence Plot (b) Recurrence Network

Figure 2: The left figure shows a recurrence plot, $R_{ij}(\varepsilon)$, in which every column corresponds to a different state space vector. If the distance between state space vectors i and j is less than the recurrence threshold ε, then $R_{ij}(\varepsilon) = 1$. Otherwise, $R_{ij}(\varepsilon) = 0$. Half of the plot is shown because $R_{ij}(\varepsilon) = R_{ji}(\varepsilon)$. The right figure shows a recurrence network, $G_{ij}(\varepsilon)$, in which every node represents a state space vector and an edge between nodes i and j indicates that $R_{ij}(\varepsilon) = 1$.

Since the structure of a recurrence network depends on the closeness between state space vectors, we suggest that the regions that phase space trajectories visit the most, or recurrence regions, should originate clusters in a recurrence network.

2.5 The Problem of Selecting an Appropriate Recurrence Threshold

The recurrence threshold, ε, determines whether two state space vectors are close or not and, therefore, it also determines the structure – its size and number of clusters – of the recurrence network associated to a time series. This way, an adequate recurrence threshold could reveal structures in different scales within the state space, provide good estimations of the network's properties and assure a better understanding of a complex system.

The problem of selecting an appropriate recurrence threshold has been largely studied. A summary of the problems associated to the selection of the recurrence threshold is given in an article of Donner et al. from 2010 [7].

Initially, the recurrence threshold was set "using rules of thumb" [23,24] over some dynamical measurements such as the correlation integrals [25], correlation dimensions [26], second order Rènyi entropy [27, 28] or attractor dimensions [7].

Generally, the recurrence threshold was kept as small as possible. Recurrence networks with low edge densities were preferred because higher edge density values tend to hide important dynamical structures. Additionally, it was desired that a small variation in the recurrence threshold did not produce noticeable differences in the dynamical analysis results.

Recently, Donges et al. [6] introduced an analytical framework, based on random geometric graphs (RGGs) theory, to analyze recurrence networks. For an extended discussion on random graphs, see the article of Dall and Christensen form 2002 [29].

Considering RGG theory, the recurrence threshold for one-dimensional, non-noisy time series with uniform probability density distribution, is determined in terms of the percolation threshold ϵ_c, which points out the limit in which the network's giant component breaks down and makes impossible to recover information about mesoscopic and path-based measures [30].

For too large ε, the recurrence network becomes too dense, and for too small ε, the recurrence network's giant component breaks down into smaller disconnected components.

In both cases the fine geometry of the time series is not well represented by the neighborhood- and path-measurements. This way, Donges et al. focused on the study of the average path length, which relates to the network's giant component, to set a range of values for the recurrence threshold.

Some approaches to the analysis of recurrence networks constructed from time series with non-uniform distributions are the study of changes in connectivity by Hsing and Rootzé [31], and more recently by Cooper and Frieze [32]. On the other hand, Kong and Yeh [33] have investigated the problem of characterizing the critical density and critical mean degree of random geometric graphs with non-uniform probability distributions and, based on probabilistic methods and clustering analysis, they have provided lower bounds for the critical density of a Poisson RGG in an m-dimensional Euclidean space.

2.6 Clustering Analysis

As we mention before, we are interested in identifying different metastable states in real-world time series. Thus, we combine the idea that there is a recurrence threshold for which a recurrence network's giant component breaks down into smaller disconnected components, with the idea that recurrence network's clusters correspond to metastable states.

This way, we suggest to look at the results of performing clustering analysis on a recurrence network in order to determine the recurrence threshold that allows the identification of all relevant metastable states in a system.

As mentioned before, the problem of finding clusters, or modules, in complex networks has been approached in several ways and many clustering algorithms exist for this purpose [11]. However, we use the Markov State Model (MSM) clustering algorithm introduced by Sarich, Djurdjevac and collaborators [9, 12].

The MSM clustering algorithm is based on spectral analysis of random walks on modular networks and identifies modules as the metastable states in the random walker and a transition region, composed with the nodes that do not belong to metastable states of the random walker or outliers.

In computational terms, this algorithm scales linearly with the size of the network, making it also useful for analyzing large networks.

Since the MSM clustering algorithm gives more refined information about a network, we suggest that using it to analyze a recurrence network will allow us to perform a more refined quantitative analysis of recurrences in state space. And this, in turn, will help us to set boundaries for a recurrence threshold.

3 SAIMeR: Self-adapted method for the identification of metastable states in real-world time series

Summarizing, SAIMeR is a method for the identification of metastable states in complex real wold time series. It is based on the theories of recurrence analysis and network clustering and is therefore divided in two parts, each one focused on one of these theories. This method is described in Algorithms 1, 2 and 3, given below.

The recurrence analysis part of SAIMeR is divided in two steps: (a) the reconstruction of state space from a time series and (b) the definition of a set of recurrence thresholds and the construction of its associated recurrence networks.

The clustering part of SAIMeR is divided in three steps: (a) the clustering of the set of recurrence networks previously obtained, (b) the selection of a subset of recurrence networks with "similar characteristics" and the recurrence thresholds producing them, and (c) the

Table 1: Recurrence Analysis Algorithm

• Construct the state space
Construct $N^* = N - (m-1)\tau$ state space vectors with specific embedding parameters τ and m as in Eq. 1 from min-max normalized time series containing N data points.

• Define set of recurrence thresholds and its associated networks
Compute initial recurrence threshold ε_0 as in Eq. 4.
for $\nu = 0$ **to** $\nu = \nu_f$ **do**
▷ Compute recurrence threshold ε_ν as in Eq. 5.
▷ Compute its associated recurrence plot R_ν.
▷ Compute its associated recurrence network G_ν.
end for
return Sets of recurrence thresholds $\{\varepsilon_\nu\}$ and recurrence networks $\{G_{ij}(\varepsilon_\nu)\}$

computation of a *final recurrence threshold* from the subset of recurrence thresholds producing such subset of networks and the identification of metastable states (and transition region) in a time series.

3.1 Part I: recurrence analysis

As mentioned above, the first part of SAIMeR is divided in two steps. These are explained in Algorithm 1.

The first step consists on the construction of state space from a time series. The second step consists on the definition of a set of recurrence thresholds – based on basic statistical analysis of the data set – and the construction of its associated recurrence networks. Let us look at these steps more carefully.

3.1.1 Constructing the state space

The first step in the recurrence analysis part of SAIMeR is the construction of the state space from a given time series normalized to the maximum in an interval going from zero to one. We reconstruct the state space using the time delay embedding method, mentioned in Section 2.2.

The time delay embedding method requires setting two parameters: the embedding delay and the embedding dimension. Different selections of embedding parameters lead state spaces containing different dynamical information. Thus, the recurrence regions identified in state spaces constructed with different embedding parameters will vary. Consequently, its recurrence quantitative information will vary as well.

This way, we suggest that the analysis of RQA measurements can be used to determine the embedding parameters that better describe the dynamics of a time series. In Appendix A we analyze the changes in entropy and recurrence rate of the recurrence plots associated to the same time series but in state spaces constructed with different embedding parameters. There, we suggest a way to choose the embedding parameters. This suggestion is used throughout this paper to determine the embedding parameters in each of the examples used to illustrate SIMeR.

3.1.2 Defining a set of recurrence thresholds

The second step in the recurrence analysis part of SAIMeR is the definition of a set of recurrence thresholds. For this purpose, we require some knowledge about the time series analyzed.

One of the distinctive features of real-world time series is the presence of *artifacts* such as noise, missing or wrong measurement points, or non-uniform probability distributions. Due to these artifacts, the results of the recurrence analysis of real-world time series can be very different from the results obtained when analyzing time series without artifacts.

We suggest to compute the recurrence threshold, ε, in terms of the second moment of the time series distribution. However, we do not use the standard deviation of the time series, σ, but assume that our data is a sample of a larger distribution and therefore use the standard error of the mean $SE_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$. The standard error of the mean measures the probability of a sample's mean to be close to the data set's mean. Even for non-uniform probability distributions, the standard error of the mean defines boundaries for the uncertainty in the value of a random variable with finite variance.

We determine the recurrence threshold to be equal to a fraction α of the smallest standard error of the mean $SE_{\overline{x}}$ of the data sample. This way, an initial guess for a recurrence threshold ε_0, considering the previous restrictions, is given by:

$$\varepsilon_0 = \alpha \frac{\sigma}{\sqrt{N}} \tag{3}$$

Varying the fraction α implies varying the size of the minimum number of recurrences. This can also be understood as varying the number of nodes required for a neighborhood to be recurrent. We set $\alpha = 0.05N$ since this corresponds to the 5% error usually accepted as error in accurate statistical analyses. This way the initial guess for the recurrence threshold is:

$$\varepsilon_0 = 0.05\sqrt{N}\sigma \tag{4}$$

For multi-dimensional data we compute the standard error of the mean for every variable, or dimension, and take the largest value to compute the initial recurrence threshold, using the same expression as for one-dimensional data. Selecting the largest standard error of the mean from all variables, we lose smaller scale information. This could be overcome by previously normalizing all variables to the value of the smallest standard deviation.

Another possible treatment could involve computing a different recurrence threshold for every variable of the time series. This way we would compute a matrix of standard errors of the mean. However, we decide to focus in the dynamics of the variable that varies the most.

Once we have computed the initial recurrence threshold, we compute a set of thresholds $\{\varepsilon_\nu\}$. Every element of $\{\varepsilon_\nu\}$ is given by:

$$\varepsilon_\nu = (1.5 - 0.1\nu)\,\varepsilon_0, \text{ for } \nu = [0, \nu_f] \tag{5}$$

The size of set $\{\varepsilon_\nu\}$ is determined by ν_f. We set $\nu_f = 14$ in order to vary the recurrence threshold from $0.005N$ to $0.075N$ times the standard error of the mean. We suggest that, this way, small variations in fraction α in Eq. 3 will not affect dramatically our results.

Finally, with every recurrence threshold $\varepsilon_\mu \in \{\varepsilon_\nu\}$ we will compute a recurrence plot, $R_\mu = R_{ij}(\varepsilon_\mu)$, and an associated recurrence network, $G_\mu = G_{ij}(\varepsilon_\mu)$, as described in Sections 2.3 and 2.4. This way, we obtain the set of recurrence networks, $\{G_\nu\}$, computed with each of the recurrence thresholds in set $\{\varepsilon_\nu\}$.

Table 2: Clustering analysis algorithm (Part 1)

- Cluster set of recurrence networks

for $\nu = 0$ **to** $\nu = \nu_f$ **do**
  - ▷ Perform clustering analysis of the associated recurrence network $G_{ij}(\varepsilon_\nu)$.
  - ▷ Compute number of clusters $C(\varepsilon_\nu)$ and number of nodes in each cluster $|C_k(\varepsilon_\nu)|$ of G_ν.

end for

- Select subset of similar networks

Select subset of thresholds $\{\varepsilon_\nu\}^-$ such that conditions in Eq. 6 hold.
for $\chi_j = [\chi_0, \chi_{j^*}]$ and χ_j as in Eq. 8 **do**
  for all $\varepsilon_\lambda \in \{\varepsilon_\nu\}^-$ **do**
    if $|C_k(\varepsilon_{\lambda+1})| - |C_k(\varepsilon_\lambda)| < \chi_j$, as in Eq. 8 **then**
      Add recurrence threshold ε_λ to subset $\{\varepsilon_\nu\}^{\chi_j}$
    end if
  end for
  if $\{\varepsilon_\nu\}^{\chi_j} \neq \emptyset$ **then**
    if $j \neq j^*$ **then**
      Continue
    else
      return $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_j}$
    end if
  else
    $\chi_{j^!} = \chi_{(j-1)}$
    return $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_{j^!}}$
  end if
end for

3.2 Part II: clustering analysis

The second part of SAIMeR has three steps. These are explained in Algorithms 2 and 3.

The first step is the clustering analysis (number and size of clusters) of the recurrence networks, obtained in the previous part.

The second step consists on the selection of the subset of networks with similar number and size of clusters. The set of recurrence thresholds producing such set is denoted by $\{\varepsilon_\nu\}^*$.

Finally, the third step consists on computing the *final recurrence threshold* ε^*, as the average value of recurrence thresholds in $\{\varepsilon_\nu\}^*$.

The clustering analysis of the recurrence network produced with the final recurrence threshold, $G_* = G_{ij}(\varepsilon^*)$, leads to the identification of metastable states (and transition region) in the time series.

3.2.1 Clustering analysis of the set of associated recurrence networks

Every recurrence network in the set $\{G_\nu\}$ of all recurrence networks associated to recurrence thresholds in set $\{\varepsilon_\nu\}$, will have a different structure.

The clustering analysis of a recurrence network $G_\mu \in \{G_\nu\}$ will indicate the number of clusters in the network and the number of nodes in each cluster.

Table 3: Clustering analysis algorithm (Part 2)

- Identify Metastable States
  - ▷ Compute *'final recurrence threshold'*, ε^*, as the average value from subset $\{\varepsilon_\nu\}^*$.
  - ▷ Perform clustering analysis of the associated recurrence network $G_* = G_{ij}(\varepsilon^*)$.
  - ▷ Classify time points into different dynamical states as in Section 3.2.3, according to their belonging to a particular cluster in the recurrence network G_*.

Every cluster identified in a recurrence network will represent a different metastable state in the time series producing such network. When we also want to know the number of nodes identified as part of the transition region (using the MSM clustering algorithm), we assign them to an extra cluster. What we obtain with this extra cluster is that the sum of all nodes in a cluster for every recurrence network $G_\mu \in \{G_\nu\}$ is the same and equal to $N^* = N - \tau(m-1)$.

A comparison of the clustering results for every recurrence network in $\{G_\nu\}$ can be represented with Sankey diagrams. A Sankey diagram is a flow diagram showing the change in clustering results between networks. For an example of a Sankey diagram, see Fig. 17 in Appendix C.

In a Sankey diagram, every network is represented as a column and every column is divided into blocks. The number of blocks in a column represents the number of clusters identified in a network. The size of a block in a column is determined by the number of nodes such cluster contains. This way, if a group of nodes in network A are assigned to a different cluster in network B, the Sankey diagram of these networks will show, as an arrow, the *flow* of such nodes from one block in column A to a different block in column B. The thickness of such arrow will be determined by the amount of nodes *flowing*.

3.2.2 Tuning the Final Recurrence Threshold

Analyzing the different clustering results of recurrence networks in $\{G_\nu\}$, we will obtain a final recurrence threshold, ε^*. The way to compute this final recurrence threshold constitutes the *self-adaptive* part of SAIMeR.

The first step to obtain the final recurrence threshold is to identify in $\{G_\nu\}$ a subset of recurrence networks with the same number of clusters, $\{G_\nu\}^-$. From such set of recurrence networks one can define the subset of recurrence thresholds $\{\varepsilon_\nu\}^-$.

Given two recurrence networks G_μ and $G_{\mu+1} \in \{G_\nu\}$, the number of clusters in them is the same if the following conditions hold:

$$\begin{aligned} C(\varepsilon_{\mu+1}) - C(\varepsilon_\mu) &= 0 \\ C(\varepsilon_\mu) - C(\varepsilon_{\mu-1}) &= 0 \\ C(\varepsilon_\mu) &> 1 \end{aligned} \tag{6}$$

Where $C(\varepsilon_\mu)$ is the number of clusters in recurrence network $G_\mu \in \{G_\nu\}$

The next step consists of selecting from $\{G_\nu\}^-$ a subset of recurrence networks with clusters of similar sizes, $\{G_\nu\}^{\chi_j}$. Here, χ_j is a value, or tolerance, measuring the similarity between the size of clusters. The set of recurrence thresholds producing the networks with clusters of similar sizes, is denoted by $\{\varepsilon_\nu\}^{\chi_j}$.

For $C_k(\varepsilon_\lambda)$ denoting the k-th cluster of recurrence network $G_\lambda \in \{G_\nu\}^-$, and $|C_k(\varepsilon_\lambda)|$ denoting the number of nodes in such cluster. Then, the size of every cluster in a pair of

consecutive recurrence networks $G_\lambda, G_{\lambda+1} \in \{G_\nu\}^-$ varies less than a specified tolerance χ_0, if:

$$|C_k(\varepsilon_{\lambda+1})| - |C_k(\varepsilon_\lambda)| < \chi_0, \text{ for } \varepsilon_\lambda \in \{\varepsilon_\nu\}^- \tag{7}$$

In Eq. 7, the tolerance depends on the number of nodes in a recurrence network, N^*, so that $\chi_0 = \chi_0(N^*)$. Initially, two k-th clusters in $G_\lambda, G_{\lambda+1} \in \{G_\nu\}^-$ will have similar size if the number of nodes they contain is different in no more than ten percent the total number of nodes in the recurrence network, or $\chi_0(N^*) = 0.1N^*$.

Initial tolerance χ_0 is later decreased in order to restrict the condition of similarity between clusters. The extreme of similarity will be reached with tolerance χ_{j^*}, when the number of nodes in the k-th clusters of $G_\lambda, G_{\lambda+1} \in \{G_\nu\}^-$ is different in no more than one percent the total number of nodes in the recurrence network, or $\chi_{j^*}(N^*) = 0.1N^*$.

The reduction of tolerance we propose consists of only ten steps, which implies that $j^* = 9$. This way, every reduced tolerance $\chi_j \in [\chi_0, \chi_{j^*}]$, is given by:

$$\chi_j = \chi_0(1 - j), \text{ for } j = [0, j^*] \tag{8}$$

If the subset of recurrence thresholds sufficing the maximum decrease of tolerance χ_{j^*} is not empty, then $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_{j^*}}$.

However, this will not always occur, since not all the sets of recurrence networks associated to $\{\varepsilon_\nu\}^-$ will suffice the maximum tolerance decrease. If the subset of recurrence thresholds sufficing the decrease of tolerance is empty for a $j > j^!$, then $\{\varepsilon_\nu\}^*$ will be the last subset of recurrence thresholds that suffices Eq. 7, it means the average of $\{\varepsilon_\nu\}^{\chi_{j^!}}$.

The final recurrence threshold, ε^*, will be the average of thresholds in $\{\varepsilon_\nu\}^*$.

3.2.3 Identification of Metastable States in the Time Series

Once that the final recurrence threshold ε^* has been computed, we generate the recurrence network associated to it, $G_* = G_{ij}(\varepsilon^*)$. We identify the different metastable states in the time series by clustering this recurrence network.

As we mentioned in Section 2.4, each node in a recurrence network represents a state space vector in the state space reconstructed from the time series. Since we use the delay mapping to obtain the state space from the time series, each component of a state space vector corresponds to a data point in a time series. Therefore, each node in a recurrence network represents a collection of time point measurements, and the size of this collection depends on the embedding dimension.

In this paper, for simplicity, if node i of G_* has been assigned to a specific cluster C_k, the data point u_i in the first component of state space vector $\overrightarrow{x}_i$ is assigned to the k-th metastable state.

This metastable state assignment approach is naïve since, for an embedding dimension m, every data point u_i appears in up to $m+1$ state space vectors. The total number $M(u_i)$ of state space vectors in which a data point appears u_i, is variable.

Therefore, the metastable state to which data point u_i is assigned, should be determined from the cluster assignment of $M(u_i)$ nodes in the final recurrence network. This means that the metastable state of data point u_i is given in terms of an average cluster number $\overline{C_i}$ and a threshold θ^* by:

$$\begin{aligned} \overline{C_i} &= \frac{1}{M(u_i)} \sum_{\{\overrightarrow{x}_j\}} C_j \\ c(u_i) &= \delta(C_i - \theta) \end{aligned} \tag{9}$$

Given that the state space vectors are computed using time delay embedding, $M(u_i)$ for data points u_i such that $\beta\tau \leq i \leq (\beta+1)\tau$, for $\beta \in \mathbb{N}$, $0 \leq \beta < m$, is equal to $\beta + 1$. For data points u_i with $N - (m-\alpha)\tau \leq i \leq N - (m-\alpha-1)\tau$, for $\alpha \in \mathbb{N}$, $0 \leq \alpha < m-1$, then $M(u_i) = m - \alpha + 1$. Any other data point has $M(u_i) = m + 1$.

4 Examples

To illustrate the ability of SAIMeR to identify metastable states in complex time series, we present and analyze three cases. We identify the different metastable states in each of these time series.

The first example corresponds to the one-dimensional time series describing the motion of a particle under the gradient of a double well potential and a random force. This is one of the simplest systems showing metastability.

The second example corresponds to a two-dimensional time series describing the molecular dynamics of trialanine, i.e. the variation of two of the three torsion angles describing its conformation (see Fig. 6), in order to identify its main molecular conformations.

In the third example we analyze a one-dimensional real-world time series containing the average daily temperatures of Berlin from June 12th, 1936 to January 9th, 2008. This time series is likely to have trends, possibly associated to climate change, and several missing measurement points during some periods of time (non-equally spaced measurements), possibly related to historical events.

As the results of these three examples suggest, having more accurate data improve the identification of metastable states in the time series with SAIMeR.

4.1 Double Well Potential

The double well potential is a simple one-dimensional system showing metastability. For this reason, this is the initial example to illustrate SAIMeR.

The time series analyzed in this section corresponds to the simulated motion of a particle, in a heat bath with temperature T, under the gradient of a double well potential and a random force. Such motion can be modeled with the following equation:

$$dX_t = -\nabla V(x)dt + \sqrt{2\epsilon}dB_t \tag{10}$$

In this equation, B_t is a Brownian motion, $\epsilon = \nu T$, $\nu > 0$ is a friction parameter and T is the temperature of the heat bath. $V(x) = (x^2 - a^2)^2$, is a double well potential with two local minima at $x_1 = a$ and $x_2 = -a$. In this case we set $a = 1$.

The double well potential model in Eq. 10 is one of the first models for metastability. It was proposed by Kramer in 1949 [34], during his studies on chemical reactions. Fig. 3 shows a representation of the double well potential $V(x) = (x^2 - 1)^2$. In this figure, ΔV is the trap depth difference between the potential wells which controls how metastable the system is.

The double well potential time series is shown in Fig. 4 and results from integrating the double well potential's Langevin dynamical equations. For this, we use the Euler Maruyama integrator with lag time $\lambda = 0.001$, 7500 iterations, initial positions $q_{init} = (0, 1)$ and temperature $T = 100$. Additionally, we sample this time series every 10 time points.

In this time series, we expect to find two main dynamical states and a transition region. Every metastable state should correspond to each of the wells in the potential and the transition region should indicate the moments of transition between potential wells. We want to stress that, for the purposes of this article, the transition region is represented as another cluster, named 'cluster 2'.

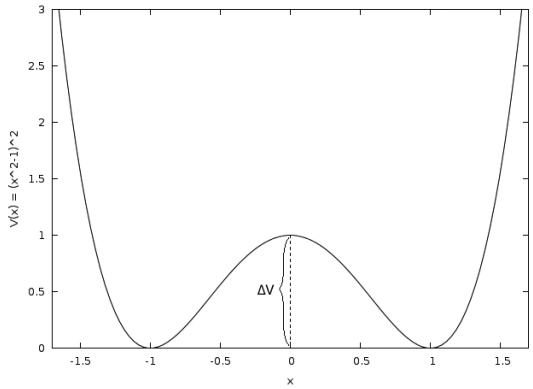


Figure 3: Scheme representing a double well potential $V(x) = (x^2 - 1)^2$, with two wells centered in $x = -1$ and $x = 1$, and with $\Delta V = 1$ the trap depth difference between them. This is a simple system showing metastable behavior.

As mentioned in Sec. 3.1, SAIMeR starts by constructing the state space from the time series. In this case, the state space is built with embedding parameters $\tau = 7$ and $m = 2$. Embedding parameters are determined as explained in Appendix A.

The next step consists on defining a set of recurrence thresholds, using Eq. 4 and Eq. 5, and performing the clustering analysis of the recurrence networks associated to each of the recurrence thresholds in this set. The clustering results for this example are shown in the Sankey diagram of Fig. 17 in Appendix C. In this figure, the size of the clusters (and transition region) do not vary too much for recurrence thresholds $\varepsilon < \varepsilon_{10}$. For details about the tolerance in variation see Section 3.2.2.

Then, as mentioned in Sec. 3.2, the clustering results of every recurrence network are used to compute the final recurrence threshold, $\varepsilon*$. The final recurrence threshold computed for this time series is $\varepsilon^* \simeq 0.29035$.

Finally, clustering the recurrence network computed with the final recurrence threshold leads to the identification of metastable states in the time series. Clustering results are shown in Fig. 5, where clusters 0 and 1 can be associated to the two expected metastable states, one for every potential well. Cluster 2 is where the transition paths between metastable states are allocated.

4.2 Molecular Conformations of Trialanine

In this section we analyze with SAIMeR the second time series example, which corresponds to the simulation of molecular configurations of trialanine. Trialanine is one of the simplest systems that exhibits the typical features of biomolecules, such as having a backbone with various stable conformations. A ball-and-stick diagram of this molecule and its torsion angles are shown in Fig. 6.

The conformation of a molecule is a mean geometric structure which is conserved on a large time scale compared to the fastest molecular motions, such that the associated subset of configurations is metastable. Characterizing a molecule with its central peptide dihedral angles, or torsion angles, has the advantage of producing a reference system invariant to translations and rotations of the molecule, reducing this way the dimensionality of the description.

At low temperatures, for example $T = 300K$, the different molecular conformations of

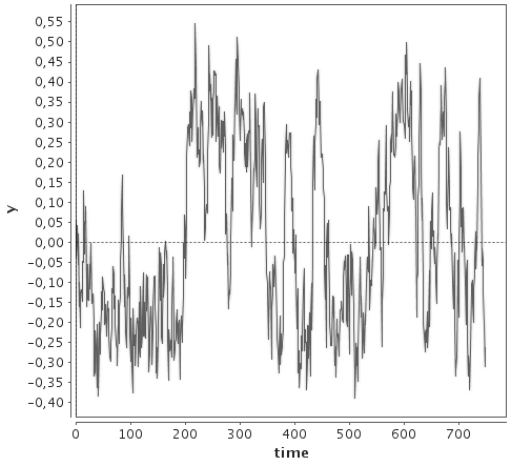


Figure 4: Time series for a particle in a double well potential. This time series result from normalizing and sampling every 10 time points the time series computed by integration of its Langevin equations using an Euler Maruyama integrator with lag time $\lambda = 0.001$, 7500 iterations, initial positions $q_{init} = (0, 1)$ and temperature $T = 100$.

trialanine can be sufficiently characterized by the two central peptide dihedral angles: ϕ and ψ. At higher temperatures, for example $T = 700K$, one should also take into consideration changes of the peptide bond angle Ω. According to Prei et al. [35] and Metzner, Putzig and Horenko [36], clustering the state space of trialanine at higher temperatures results in the identification of five metastable states.

The time series analyzed in this section is simulated with JGromacs [37], in which trialanine is represented by 21 united atoms. It is simulated with 5000 steps, in vacuum and at constant temperature $T = 300K$, to produce time series that can be considered stationary. Additionally, we sample such time series with rate $\Delta t = 10$, which does not hide transitions between states for any torsion angle. For more details about the simulation, see the article of Prei et al. from 2004 [35].

Since the time series is simulated at $T = 300K$, the following analysis considers only the two central peptide dihedral angles, ϕ and ψ.

The molecular conformations of trialanine can be shown in a two-dimensional plot, called the Ramachandran plot, which contains the dependency between ϕ and ψ only.

The state space associated to trialanine's molecular conformations is constructed using the time delay embedding, with embedding dimension $m = 2$ and embedding delay $\tau = 7$ (see Appendix A).

As mentioned above, the state space of trialanine at higher temperatures has resulted in the identification of five metastable states [35, 36]. For this reason, we guess the number of clusters MSM should identify (see Section 2.6).

This way, the final recurrence threshold computed for this time series is $\varepsilon^* \simeq 0.2796$ and the clustering analysis of the recurrence network associated to this ε^*, results in the identification of the five clusters shown in Fig. 8.

In Fig. 8, we see three larger sets of points and two smaller sets. Due to their location in the Ramachandran plot, one can identify the three main sets with the three main molecular conformations for trialanine mentioned by Fischer et al. in 2006 [38]. The two smaller sets could be a consequence to the way we assign every time point data to a metastable state, as mentioned in Section 2.6.

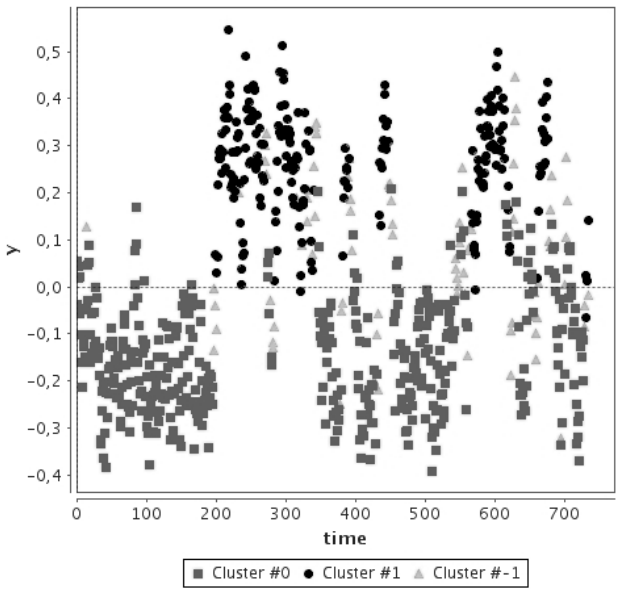


Figure 5: Metastable states identified with SAIMeR on the time series from Fig. 4. The grayscale color code shows the different metastable states identified. The state space associated to such time series is built with the delay mapping using embedding delay $\tau = 7$ and embedding dimension $m = 2$. The metastable states are identified as the clusters found in the recurrence network computed from the state space using recurrence threshold $\varepsilon^* \simeq 0.29035$.

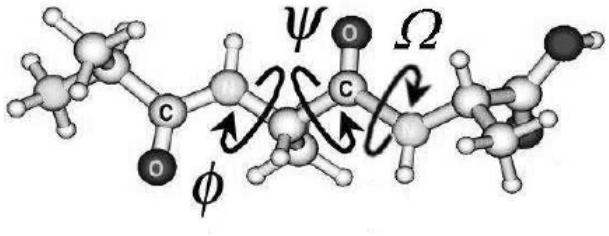


Figure 6: Ball-and-stick representation of a trialanine dipeptide molecule and its torsion angles ϕ, ψ and Ω. At low temperatures, its stable molecular conformations can be sufficiently characterized by the central peptide dihedral angles ϕ and ψ, but at higher temperatures one should also consider the peptide bond angle Ω.

4.3 Weather data

The last example corresponds to the observations of the average daily temperatures in Berlin-Tempelhof weather station (located near Tempelhof Airport) from June 12, 1936 to January 9, 2008.

The Berlin-Tempelhof measuring station is located in N 52°47', E 13°40', at 49m a.m.s.l. The time series is taken from the Rimfrost database [39], which collects information from the German Weather Service [40] (Deutscher Wetterdienst) and the NASA Goddard Institute for Space Studies [41] (NASA-GISS).

This time series has several periods without measurements, as is shown in Fig. 9. We will refer to it as the *complete* time series. The relationship between some of these periods

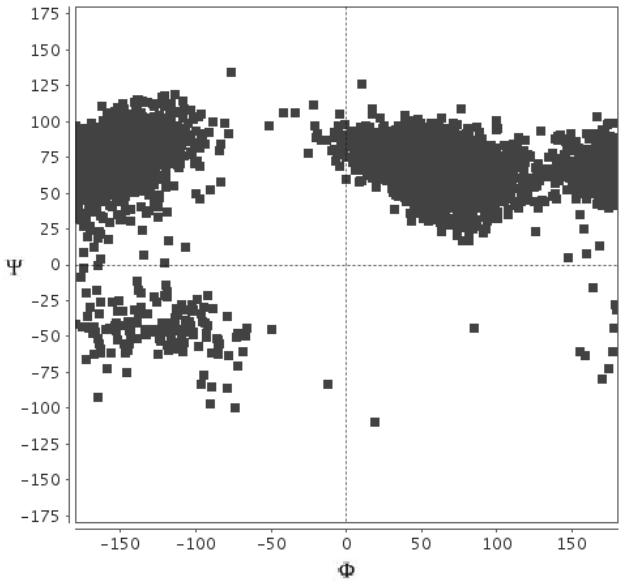


Figure 7: Ramachandran plot containing a sample of the molecular conformations of trialanine, simulated in vacuum at $T = 300K$ (for details go to the text). Conformations are given by the dependency between torsions angles ϕ and ψ.

without measurements and historical events, will be addressed in Appendix D.

Due to the large amount of missing data points in this time series data, its analysis illustrates how SAIMeR can identify metastable states in complex time series.

To start our analysis, we ignore the data points in which no measurements were taken and produce this way a merged time series. This merged time series, shown in Fig. 10, has a measurement for every time point.

To simplify the computations, we sample the merged time series every 14 time points to produce a *coarse* time series. In the periods in which measurements are regularly taken, this sampling rate corresponds to taking the daily temperature every second week and therefore we suggest that season transitions could be sufficiently represented. Evidently, this is not the case in the periods in which measurements are irregular and we can not guarantee the appropriate representation of seasons. For this reason, we use the *coarse* time series to compute a final recurrence threshold but later identify metastable states in different sections of the merged time series.

Using the *coarse* time series, we reconstruct the state space, using embedding parameters $\tau = 2$ and $m = 2$ (see Appendix A). This way, the final recurrence threshold we compute is $\varepsilon^* \simeq 0.2933$.

Now, we use the same final recurrence threshold ($\varepsilon^* \simeq 0.2933$) and embedding parameters ($\tau = 2$ and $m = 2$) to analyze three segments of the merged time series, which were produced by ignoring time points in which no measurements were taken. The results from such analysis is then used to reconstruct the analysis of the *complete* time series in the same periods of time, by adding the missing time points to the time series.

The three segments of the merged time series we mention, correspond to three periods of time: 1937 and 1938, 1942 and 1943, and 1991 and 1992. In these time series we expect to identify yearly seasons and the transit between them.

The first period of time we analyze goes from January 1, 1937 to December 31, 1938. The

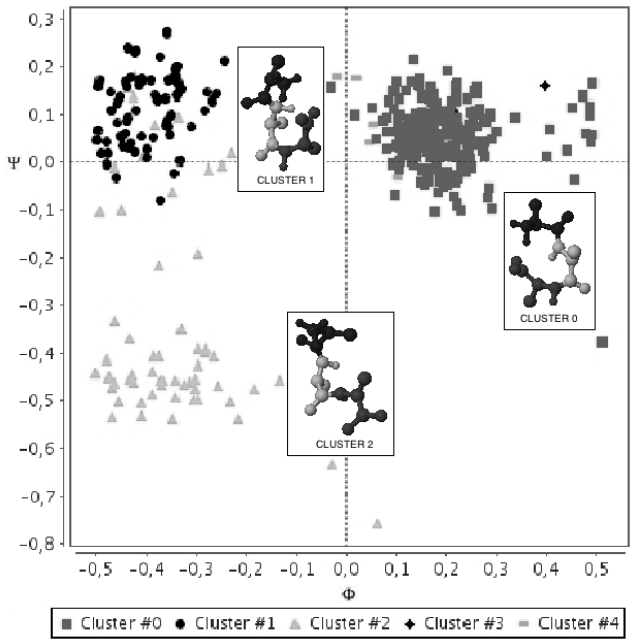


Figure 8: Ramachandran plot containing a sample of the molecular conformations of trialanine simulated at $T = 300K$ (for details go to the text). Time series for ϕ and ψ are normalized. The grayscale color code identifies the five different metastable states (or main molecular conformations) identified with SAIMeR (using $\varepsilon^* \simeq 0.2796$, $\tau = 7$ and $m = 2$). In the plot, every metastable state is called a cluster. For each cluster we show an example of molecular conformation of trialanine belonging to it.

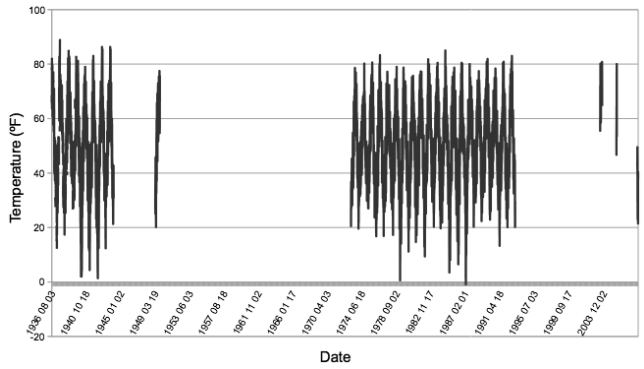


Figure 9: Daily average temperatures in Berlin - Tempelhof measuring station from June 12, 1936 to January 9, 2008. Measurements are irregularly taken before August 31, 1939 and measuring techniques previous to 1943 are not provided. Empty spaces in the plot correspond to periods in which no measurements were taken, due to historical or technical reasons.

result of this analysis is shown in Fig. 11 (a).

This period has several missing measurements – around 30% of the time points – and there is no information about the way in which measurements were taken. In this time series we identify one metastable state corresponding to a colder season (cluster 0 in the figure), which lasts around six months. A second metastable state (cluster 1 in the figure) and

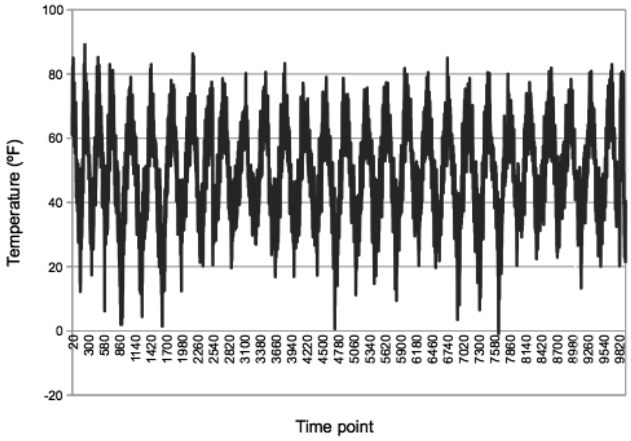


Figure 10: *Merged* time series. Obtained from the time series in Fig. 9 (containing the daily average temperatures in Berlin-Tempelhof measuring station from June 12, 1936 to January 9, 2008) by ignoring the time points in which no measurements were taken.

the time points associated to the transition region (cluster -1 in the figure) do not seem to correspond to any yearly season. These results might originate from the large amount of missing measurement points in the time series, which would require a different recurrence threshold and embedding dimensions to be analyzed. Another reason for these results might be the dispersion of the temperature measurements data, which might originate from a non-systematic measuring technique. A suggestion to improve the identification of metastable states in this time series is to analyze it with different recurrence threshold and embedding parameters, specific for these data and not for the *coarse* time series.

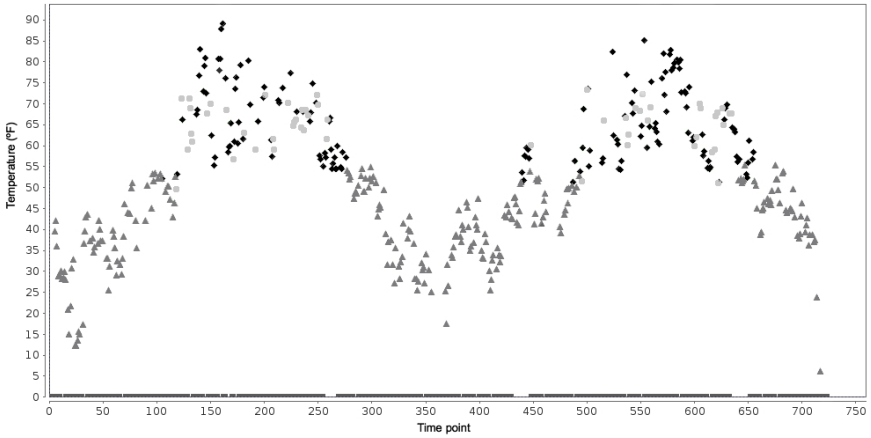
The second period of time we analyze goes from January 1, 1942 to December 31, 1943. The result of this analysis is shown in Fig. 11 (b).

This period does not have many missing measurements – less than 1% of the time points – . The temperature measureents data in this region are less disperse than in the previous region, which suggests a more systematic measuring technique. In this time series we identify one metastable state (cluster 0 in the figure) corresponding to a colder season, which lasts around six months, and another metastable states (cluster 1 in the figure) corresponding to a warmer season that also lasts around six months. Cluster -1 indicates points in between the warmer and the colder seasons coming from the transition region in the final recurrence network associated to this period of measurements. Time points in cluster -1 relate to the periods of transition between the colder and the warmer seasons.

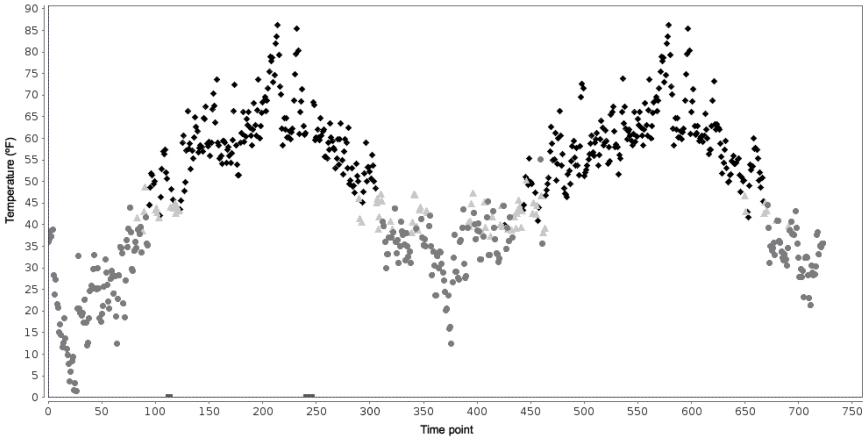
The third period of time we analyze goes from January 1, 1991 to December 31, 1992. The result of this analysis is shown in Fig. 11 (c).

This period does not have missing measurements. Additionally, temperature data in this region were obtained with a more systematic measuring technique. In this time series we identify one metastable state (cluster 0 in the figure) corresponding to a colder season, which lasts around six months, and another metastable state (cluster 1 in the figure) corresponding to a warmer season that also lasts around six months. Cluster -1 indicates the periods of transition between the colder and the warmer seasons.

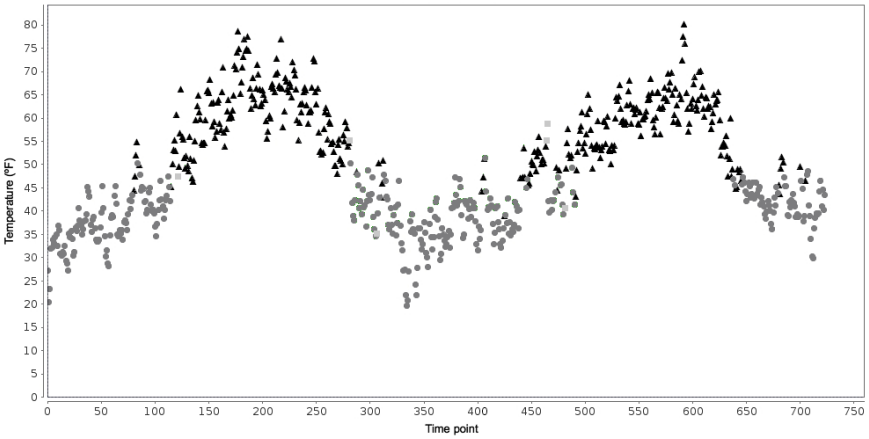
To the best of our knowledge, the time series data analyzed in this section has not been previously analyzed in any similar fashion. However statistical analysis and interpretations of such analysis have been performed [42].



(a) January 1, 1937 to December 31, 1938.



(b) January 1, 1942 to December 31, 1943.



(c) January 1, 1991 to December 31, 1992.

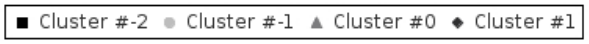


Figure 11: The grayscale color code in these figures represents the different metastable states (clusters 0 and 1) and transition region (cluster -1) identified with SAIMeR (using $\varepsilon^* \simeq 0.2933$, $\tau = 2$ and $m = 2$) in different periods of the time series containing the daily average temperatures measured in Berlin-Tempelhof from June 12, 1936 to January 9, 2008. Time points with missing temperature measurements are assigned temperature $T = 0°F$ and indicated by cluster -2. In (a), temperature measurements are irregularly taken and there is no information about the measuring technique. In (b) and (c), the measuring technique is specified and there are few or none missing measurements.

5 Robustness

In this section we measure the robustness of SAIMeR. We define robustness as the similarity between the metastable states identified in a (original) time series and in the same time series when some artifacts have been added. By artifacts we understand noise or missing data points.

We measure robustness with the Adjusted Rand Index [43] (ARI), developed by Hubert and Arabie in 1985. The ARI is a measurement of agreement between two (clustering) partitions that ranges from 0 – when the partitions are not similar at all – to 1 – when the partitions are the same –. This can be used even if the number of clusters in the two partitions compared is different, assigns a constant value of zero to the expected value of agreement between two random partitions and does not get affected when comparing partitions with a high number of clusters. The expression of this index can be found in Appendix B.

Since we use the MSM clustering algorithm in SAIMeR, we need to adapt the measurement of similarity in order to account for the different partition of the networks into modular and transition regions. This treatment is based on the work of Hueffner et al. from 2013 [11], in which every node identified as part of the transition region of a recurrence network is assigned to an independent cluster in order to create a full partition, in which the ARI is computed. In the following analysis, we will use this type of partitioning.

In our case, the two clustering partitions used to compute the ARI are the ones coming from the MSM clustering analysis of the final recurrence networks associated to a time series and the same time series with added artifacts.

We analyze the robustness of SAIMeR in two cases: when a time series has a percentage of noise added and when a percentage of time points has been removed from a time series. For these analysis, we take as example a double well potential time series.

5.1 Noisy time series

To test the robustness of SAIMeR to analyze time series with noise, we measure the similarity between (a) the clustering partition obtained from analyzing a time series and (b) the clustering partition obtained from analyzing a *noisy* time series. We compare two different types of noisy time series, whose construction we describe below.

The results we show in the following two sections, suggest that SAIMeR is able to identify metastable states in time series with noise of amplitude up to 20% the amplitude of the original time series or with amplitude up to 200 times the minimum variation (different to zero) in consecutive measurement points in the time series.

Our results also confirm Zbilut's statement of inflation of the embedding dimension when reconstructing the state space from noisy time series [3]. Thus, the ARI is higher when selecting different embedding parameters (as explained in Appendix A) to reconstruct the state space from noisy time series.

5.1.1 Noise as a fraction of minimum change between consecutive measurement points

The first definition of noisy time series we use is the one indicated by Hassona [44] for the analysis of variations of RQA measurements when adding noise to time series data. In this, a noisy time series is computed by adding Gaussian white noise (mean $\mu = 0$ and standard deviation $\sigma = 1$) with amplitude equal to a multiple, α, of the minimum variation different to zero in consecutive measurement points to the time series. We vary the amplitude of noise from $\alpha = 1$ to $\alpha = 100$ in intervals $\Delta\alpha = 10$. For every increase in the amplitude of noise,

we compute 10 different time series, in order to get rid of the bias produced by the selection of noise.

We simulate a time series for the double well potential, as described in Section 4.1, and analyze it with SAIMeR. The recurrence threshold and the embedding parameters used in this case are $\varepsilon \simeq 0.3922$, $\tau = 3$ and $m = 2$.

Using the same recurrence threshold and embedding parameters for the construction of the state space associated to each noisy time series, we obtain the results shown in Fig. 12. In this plot, the ARI indicating the similarity between the original and the noisy partitions is lower than 0.6 for $\alpha < 25$.

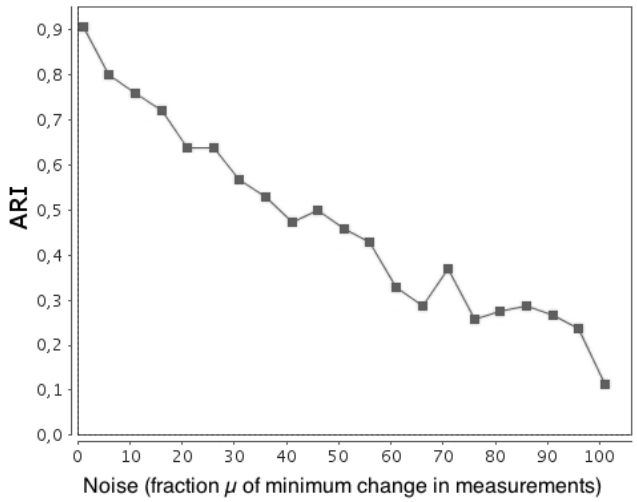


Figure 12: Similarity, measured with the Adjusted Rand Index (ARI), between the metastable states identified in the original time series and in the time series where white Gaussian noise has been added. In this case, the amplitude of noise is equivalent to a fraction, α, of the minimum change between consecutive measurement points. All partitions are computed with the same recurrence threshold $\varepsilon \simeq 0.3922$ and embedding parameters $\tau = 3$ and $m = 2$.

As mentioned by Zbilut in 1992 [3], having noise in a time series has an effect of inflation of the embedding dimension when reconstructing the state space. The low resistance to noise shown in Fig. 12 could indicate the necessity to increase the embedding dimension as we increase α. To confirm this suggestion, we perform a second experiment where every noisy time series is analyzed with different recurrence threshold and embedding parameters. In this case we vary α from 1 to 200, in intervals $\Delta\alpha = 10$. The results we obtain are shown in Fig. 13.

As Figs. 12 and 13 suggest, adapting a recurrence threshold and embedding parameters to every noisy time series, increases the ARI measured with the noisy and the original time series. Without adapting the analysis parameters, ARI < 0.6 for $\alpha < 30$, but adapting them, the ARI does not drop lower than 0.6 even for α equal to 200.

5.1.2 Noise as a percentage of the amplitude of the time series

The second definition of noisy time series we use also corresponds to the addition of Gaussian white noise to the time series (mean $\mu = 0$ and standard deviation $\sigma = 1$), but with amplitude equal to a percentage, α', of the amplitude of the original time series. In this case, the amplitude of noise varies from $\alpha' = 0$ to $\alpha' = 100$ in intervals $\Delta\alpha' = 10$. Once more, in order

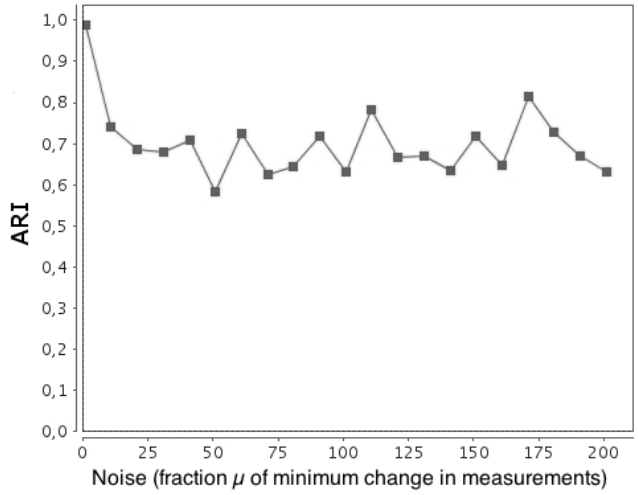


Figure 13: Similarity, measured with the Adjusted Rand Index (ARI), between the metastable states identified in the original time series and in the time series where white Gaussian noise has been added. In this case, the amplitude of noise is equivalent to a fraction, α, of the minimum change between consecutive measurement points. Every partition is computed with different embedding parameters and recurrence thresholds. The range of α is larger than in Fig. 12, showing that this approach provides more similar results (ARI > 0.6).

to get rid of the bias produced by the different noise introduced, we try 10 different noisy time series for every increase in the amplitude of noise analyze each of them with different recurrence threshold and embedding parameters.

We simulate a time series for the double well potential, as described in Section 4.1, and analyze it with SAIMeR. According to the results obtained in Section 5.1.1, the recurrence threshold and embedding parameters we use are different for every noisy time series analyzed. In Fig. 14 we observe the similarity in clustering results between (a) the noisy and (b) the original time series.

The ARI measuring the similarity between the clustering partitions for the noisy and for the original time series, is higher than 0.6 for $\alpha' \leq 20$. ARI values remain higher than 0.9 for $\alpha' \leq 10$.

These results suggest that, when adapting a recurrence threshold and embedding parameters to every noisy time series, SAIMeR enables the identification of metastable states in the system even for time series with noise with amplitude given by $\alpha = 20$, or 20% the amplitude of the original time series.

5.2 Removing data points

One of the typical features of real-world time series is having observations irregularly taken. This irregularity can be understood as if a percentage of measurement points, randomly distributed in the time series, had been removed from a time series containing a set of measurements regularly taken.

To analyze this type of irregular time series, we produce a time series with regularly spaced measurements, called the original time series, and assign to a percentage of randomly distributed data points a "null" value. We refer to the time series resulting from this process as the *trimmed* time series. Since we are not ignoring time points but only assigning a new value to some time points, the length of the original and the trimmed time series is the same.

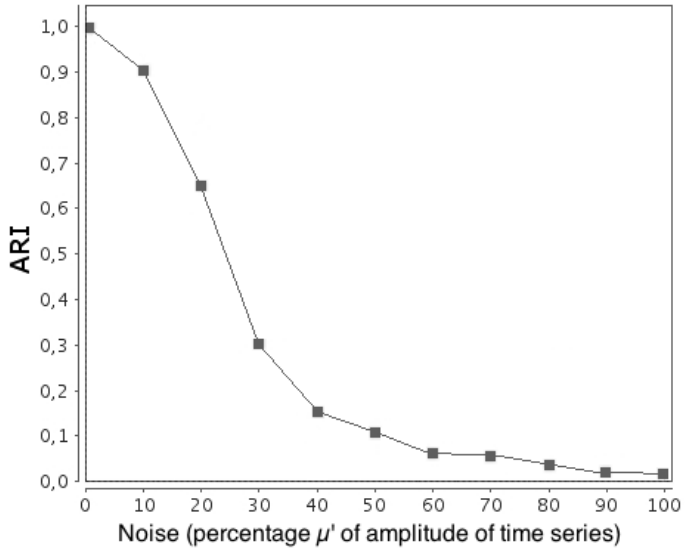


Figure 14: Similarity, measured with the Adjusted Rand Index (ARI), between the metastable states identified in the original time series and in the time series where white Gaussian noise has been added. In this case, the amplitude of noise is equivalent to a percentage α' of the amplitude of the original time series. Every partition is computed with different embedding parameters and recurrence thresholds.

We vary the percentage of time points being removed from 0% to 19%, in intervals of 1%. For every percentage of data points being removed, we compute 10 different time series, in order to get rid of the bias produced by the selection of data points to remove.

We compare the partition obtained by analyzing the original time series and the partition obtained by analyzing the trimmed time series.

All the clustering partitions associated to the trimmed time series are computed using the same recurrence threshold originally computed with the complete time series, $\varepsilon \simeq 0.3678$.

However, we interpret the case of removing measurement points from the time series as another case of noise and therefore we use different embedding parameters for the reconstruction of the the state space from every time series with artifacts. Every recurrence network associated to a trimmed time series is computed with different embedding parameters. Fig. 15 shows the ARI values obtained.

In Fig. 15 we see that the ARI has values higher than 0.9 for time series with up to 5% of time points removed. The ARI does not have values lower to 0.6 even for time series with up to 19% of time points being removed. These results suggest that the recurrence threshold computed with SAIMeR enables the identification of the metastable states even for time series with up to 19% of randomly distributed missing points.

We believe that these results could be improved by selecting a different recurrence threshold for every time series with missing points with which the comparison is performed. However, since the results obtained by using the same recurrence threshold already allow a good identification of the metastable states, we maintained this methodology due to its simplicity.

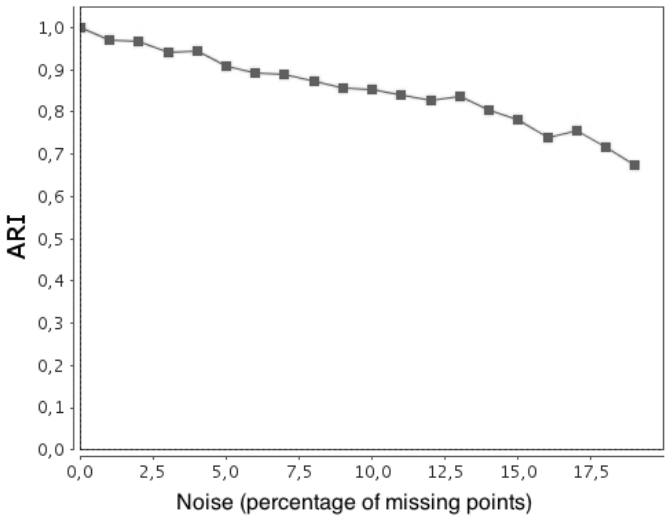


Figure 15: Similarity, measured with the Adjusted Rand Index (ARI), between the metastable states identified in the original time series and the *trimmed* time series, which contains a percentage of randomly distributed missing data points. All partitions are computed with different embedding parameters but with the same recurrence threshold $\varepsilon \simeq 0.3678$.

6 Conclusions

In this paper, we present SAIMeR, a self-adapted method for the identification of metastable states in real-world time series based on recurrence networks analysis.

SAIMeR uses particular statistical information of the time series analyzed in order to produce a recurrence threshold. Clustering the recurrence network associated to such recurrence threshold results in the classification of time points into different metastable states.

We use three examples to illustrate the performance of SAIMeR, where we identify metastable states that correspond to the different dynamical behaviors in such time series. The first example is a double well potential, where we identify two metastable states, which correspond to moments in which the simulated particle is in each of the potential wells, plus a region of transition between such states. The second example is the molecular conformations of trialanine. We identify five main metastable states in this data, which seem to correspond to the different conformation states of the molecule mentioned by Fischer et al. in 2006 [38]. The third example corresponds to the daily average temperature measured in Berlin-Tempelhof between June 12, 1936, and January 9, 2008. In this time series, despite the several missing data points in some regions of the time series, we identify two main metastable states, corresponding to the warmer and the colder seasons.

Additionally, we show that SAIMeR gives similar results (measured with the Adjusted Rand Index [43] [ARI]) when identifying metastable states on time series with artifacts (where noise have been added or data points have been removed). Similarity is measured for time series where a percentage of data measurements have been removed, where $ARI > 0.6$ even for time series with 19% of data points removed. It is also measured for time series with Gaussian white noise added. We consider two definitions of noise. In the first, the amplitude of noise is expressed as a multiple, α, of the minimum variation different to zero between consecutive measurements in the time series. In this case, ARI is higher than 0.6 even for $\alpha = 200$. In the second definition, the amplitude of noise is expressed as a percentage, α',

of the amplitude of the original time series. In this case, ARI is higher than 0.6 even for $\alpha' = 20$.

Finally, since any recurrence analysis requires the reconstruction of the state space and since we use the delay mapping for this purpose, we propose a methodology to determine appropriate embedding parameters (delay and dimension). We propose that the embedding parameters must provide the first minimum of entropy and maximum of recurrence rate during the recurrence analysis of the time series with SAIMeR. The selection of these parameters, prior to the recurrence analysis, is still an open problem that we aim to approach in future work.

All the results of the experiments mentioned above suggest that SAIMeR is an efficient tool for the analysis of real-world time series.

A Comments about Taken's embedding parameters.

Recurrence quantitative analysis (RQA) measurements are sensitive to variations in the recurrence threshold and embedding parameters. Here, we analyze the sensitivity of two quantitative measurements when varying the embedding parameters for SAIMeR on the double well potential time series in Fig 4: the entropy (Eq. 12) and the recurrence rate (Eq. 11).

A.1 Recurrence rate

The recurrence rate [3, 45], RR, is a recurrence measurement that indicates the percentage of recurrence points in a recurrence plot. In terms of the recurrence network, it indicates the relative frequency of edges a node contributes to [24]. This way, higher values in this measurement indicate that the nodes are more connected. Or, in other words, that a larger number of state space vectors are inside a same state space neighborhood.

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{ij}(\varepsilon^*) \tag{11}$$

A.2 Entropy

Entropy, $S(\varepsilon^*)$, refers to the Shannon entropy and indicates the probability to find a diagonal line of length l in a recurrence plot $R_{ij}(\varepsilon^*)$. In other words, it indicates the complexity of a recurrence plot, with respect to its diagonal lines. As a recurrence plot depends on the recurrence threshold, variations in this parameter will modify the value of the entropy. Variations in the embedding parameters will also modify this value [46].

Being $R_{ij}(\varepsilon^*)$ a recurrence plot computed from N^* state space vectors and $P(l) = P(\varepsilon^*, l)$ its histogram of diagonal lines of length l, the relative frequency of diagonal lines with length l is given by $p(l) = P(l)/N_l$. This relative frequency is equal to the number of diagonal lines with length l divided by the total number of state space vectors. This way, according to Marwan et al. [16], entropy is given by:

$$S = -\sum_{l_{min}}^{N^*} p(l) \log p(l) \tag{12}$$

In Eq. 12, l_{min} is the minimum length of the diagonal lines in a recurrence plot. This length can be defined for a recurrence plot computed from non-noisy time series. However, we want to work with real-world time series data and expect to have noise.

According to Marwan et al. [16], entropy is small for recurrence plots computed from noisy time series. This is expected because noisy time series produce recurrence plots with many short and thin diagonal lines and single points. Since we wand to distinguish noise from the rest of our time series data, we would like to remove short diagonals from the entropy computation. Therefore, we compute a new minimum length of the diagonal lines, l^*_{min}, as:

$$l^*_{min} = \frac{\sum_{l=0}^{N^*} lp(l)}{\sum_{l=0} p(l)} \tag{13}$$

In general, a lower value in entropy indicates that a recurrence plot has thinner diagonal lines, which in turn indicates less time intervals with similar evolution in the time series originating such recurrence plot. This way, a minima in entropy could indicate the recovery of more dynamical structure in the associated recurrence plot. For this reason, we would like to find a recurrence threshold and embedding parameters that produce recurrence plots with a lower value of entropy.

A.3 Embedding parameters

We suggest that selecting the embedding parameters that first provide a simultaneous local minima in entropy and local maxima in recurrence rate, and that also give the lowest minimum in entropy, construct a state space in which more nodes are closer for smaller neighborhoods. This kind of space would provide more structure in the associated recurrence networks produced when analyzing recurrences in the state space.

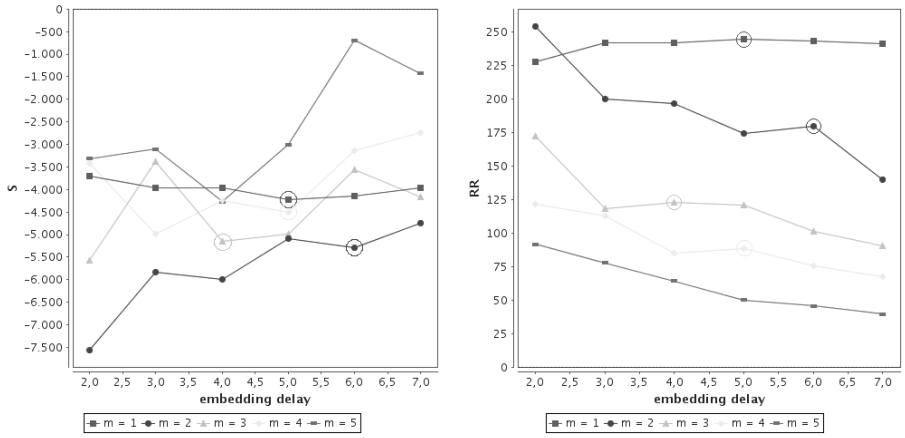
To illustrate our suggestion, we use the time series for a double well potential. Determining a selection of embedding delay and embedding dimension, we can construct the state space associated to this time series and compute a final recurrence threshold. With this recurrence threshold and the state space constructed with the selected embedding parameters, we can compute the recurrence plot associated to the time series. In this recurrence plot we measure the recurrence rate and the entropy.

Graphs in Fig. 16 show the different recurrence rate and entropy values obtained for different selections of embedding parameters when analyzing a double well potential time series. Embedding parameters that first provide a simultaneous local minima in entropy and local maxima in recurrence rate, are pointed in circles. From these combination, we select those that produce the recurrence plot with minimum entropy and use them in SAIMeR in order to identify metastable states.

B The Adjusted Rand index

This index measures the agreement between any two (clustering) partitions, even if the number of clusters in each of them is different. It assigns a constant value of zero to the expected value of agreement between two random partitions and ranges between zero and one.

Let us imagine $S = \{O_1, ..., O_N\}$, a set of objects. The number of combinations of pairs that are possible to make from set S is $\binom{N}{2}$. Set $P = \{p_1, p_2, ..., p_A\}$ and $Q = \{q_1, q_2, ..., q_B\}$ two partitions (or collections of subsets) of S such that $\cup_{a=1}^{A} p_a = \cup_{b=1}^{B} q_b = S$, $p_a \cap p_{a'} = \emptyset$ for any $a \neq a'$, and $q_b \cap q_{b'} = \emptyset$ for any $b \neq b'$. If t_{ab} represents the number of objects in S that were classified in the a-th subset of P and in the b-th subset of Q, then the ARI, as



(a) Entropy (b) Recurrence rate

Figure 16: Plot showing the variation in the entropy (computed as in Eq. 12) and recurrence rate (computed as in Eq. 11) for the different recurrence plots associated to the same time series of a double well potential resulting from constructing different state spaces with different embedding parameters (embedding delay and dimension). The circles in the plot point at the combination of embedding dimension and delay for which we find a simultaneous local minimum in entropy and a first local maximum in recurrence rate.

defined by Santos [47], can be expressed as the quotient F_1/F_2, where:

$$F_1 = \binom{n}{2} \sum_{a=1}^{A} \sum_{b=1}^{B} \binom{t_{ab}}{2} - \sum_{a=1}^{A} \binom{t_{a\cdot}}{2} \sum_{b=1}^{B} \binom{t_{\cdot b}}{2}$$

$$F_2 = \frac{1}{2}\binom{n}{2} \left[\sum_{a=1}^{A} \binom{t_{a\cdot}}{2} + \sum_{b=1}^{B} \binom{t_{\cdot b}}{2} \right] - \sum_{a=1}^{A} \binom{t_{a\cdot}}{2} \sum_{b=1}^{B} \binom{t_{\cdot b}}{2}$$

C Sankey diagram for two well potential time series analysis.

Sankey diagrams are a visual tool that shows the number of clusters as well as the nodes distribution for each of the different recurrence networks computed from the tuning set $\{\varepsilon_\nu\}$.

In these diagrams, each network is represented as a column, the number of clusters in a network is represented as sections of the column whose size corresponds to the number of nodes each cluster has. The amount of nodes whose correspondence to a cluster varies from one network to another, is represented as a flux between columns, and the width of such flux corresponds to the number of nodes whose classification differs between two networks.

Fig. 17 shows the Sankey diagram used for the two well potential time series analysis of Section 4.1. In this particular diagram, we observe a group of networks (columns) with the same number of clusters (size of sections of a column), for which the number of nodes in each cluster is almost the same (low flux of nodes from one column to another). This is the set of networks with which we compute the final recurrence threshold used for the identification of metastable states in the two well potential time series. Recurrence networks fulfilling conditions 6 and 7, have a similar number and size of clusters identified. We suggest that

these networks define a set of recurrence thresholds giving robust results about the dynamics of the time series analyzed.

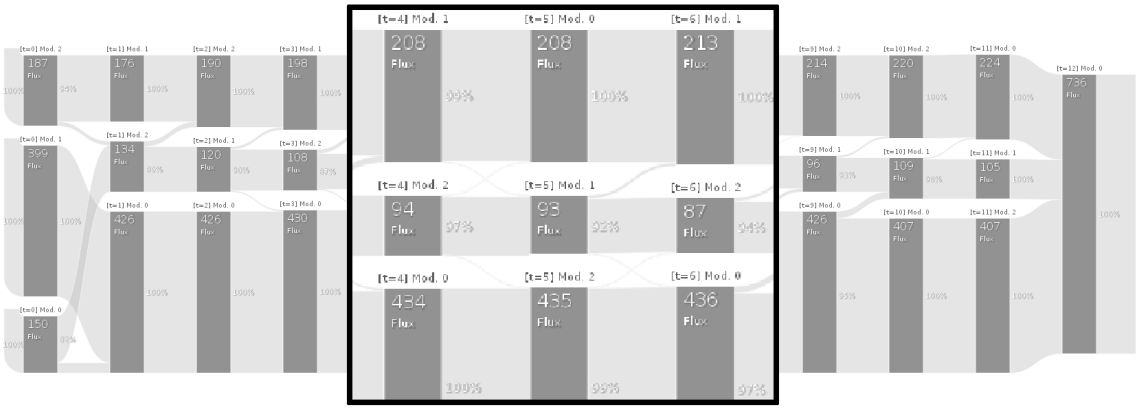


Figure 17: Sankey diagram showing the subgroup of recurrence networks (columns) with the same number of clusters (see eq. 6) and similar number of nodes (see eq. 7). Networks are computed from tuning set $\{\varepsilon_\nu\}$ (see eq. 5) on the state space constructed from a two well potential time series and embedding parameters $\tau = 7$ and $m = 2$. We suggest that this group of networks determines the recurrence threshold giving robust results about the dynamics of the time series analyzed.

D Weather data: Specifications

The time series mentioned in Section 4.3 has several periods of time without temperature measurements. Before August 31, 1939 measurements are irregular. After that day, temperature is taken daily, even though the measuring technique is not specified. Measurements continue until December 31, 1943, after a short interruption from August 29, 1942 to September 5, 1942.

During the three decades after December 31, 1943 There are no measurements. An exception are the seven months from February 1, 1949, to August 27, 1949. These decades include the end of the war, on 1945 and the Cold War.

Measurements start again on January 1, 1973 and are uninterrupted until December 31, 1992. However, after that date, measurements are scarce. Temperature was measured in 2003 only from May 18 to June 28, and from August 10 to August 16. During 2005, temperature was measured from May 16 to May 31. The last measurements we consider in our analysis were taken from December 3, 2007 to January 9, 2008.

References

[1] L. van der Maaten, E. Postma, and J. van den Herik. *Dimensionality Reduction: A Comparative Review*. Tilburg University, Netherlands, 2009.

[2] S. O. Kamphorst J. P. Eckmann and D. Ruelle. Recurrence plots of dynamic-systems. *Europhysics Letters*, 4(9):973–977, 1987.

[3] J. P. Zbilut and C. L. Webber Jr. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171:199–203, 1992.

[4] J.P. Zbilut A. Krishnan, A. Giuliani and M. Tomitah. Implications from a network-based topological analysis of ubiquitin unfolding simulations. *PLoS ONE*, 3, 2008.

[5] M. Tomita A. Krishnan, J.P. Zbilut and A. Giuliani. Proteins as networks: usefulness of graph theory in protein science. *Curr. Prot. Peptide Sci.*, 9:28–38, 2008.

[6] Jonathan F. Donges, Jobst Heitzig, Reik V. Donner, and Jürgen Kurths. Analytical framework for recurrence network analysis of time series. *Phys. Rev. E*, 85:046105, Apr 2012.

[7] Reik V. Donner, Yong Zou, Jonathan F. Donges, Norbert Marwan, and Jürgen Kurths. Ambiguities in recurrence-based complex network representations of time series. *Phys. Rev. E*, 81:015101, Jan 2010.

[8] O. Dimigen S. Schinkel and N. Marwan. Selection of the recurrence threshold for signal detection. *The European Phys. J.*, 164:45–53, 2011. Special Topics.

[9] M. Sarich et al. Modularity revisited: A novel dynamics-based concept for decomposing complex networks. *To Appear, Journal of Computational Dynamics*, 2012. `http://publications.mi.fu-berlin.de/1127/.2,3,5`.

[10] M. Sarich and C. Schuette. Approximating selected non-dominant timescales by markov state models. *Comm. Math. Sci.*, 10(3):1001–1013, 2012.

[11] S. Hueffner, B. Kayser, and T. O. F. Conrad. Finding modules in networks with non-modular regions. *Lecture Notes in Computer Science*, 7933:188–199, 2013. Proceedings of the 12th international symposium, SEA 2013.

[12] N. Djurdjevac, S. Bruckner, T. O. F. Conrad, and C. Schuette. Random walks on complex modular networks. *Journal of Numerical Analysis Industrial and Applied Mathematics*, 6(1-2):29–50, 2012.

[13] H. Poincaré. Sur le probeéme des trois corps et les ´quations de la dynamique. *Acta Mathematica*, 13(1-270), 1890.

[14] M. Thiel N. Marwan, M. C. Romano and J. Kurths. Crossed recurrence plot based synchronization of time series. *Nonlinear Processes in Geophysics*, 9:325–331, 2002.

[15] Norbert Marwan, Niels Wessel, Udo Meyerfeldt, Alexander Schirdewan, and Jürgen Kurths. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. *Phys. Rev. E*, 66:026702, Aug 2002.

[16] M. Thiel N. Marwan, M. C. Romano and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438:237–329, 2007.

[17] N. Marwan et al. Complex network approach for recurrence analysis of time-series. *Phys. Letters A*, 373:4246–4254, 2009.

[18] J. A. Yorke T. Sauer and M. Casdagli. Embedology. *J. Stat. Phys.*, 65(3-4):579–616, 1991.

[19] F. Takens. *Detecting strange attractor in turbulence*, volume 898 of *Lecture Notes in Mathematics*. Springer Verlag, Berlin, 1981.

[20] H. Whitney. Differentiable manifolds. *Ann. Math.*, 37:645–680, 1936.

[21] H. D. I. Abarbanel. *Analysis of Observed Chaotic Data*. Springer Verlag, Berlin, 1996.

[22] C. Letellier, I. M. Moroz, and R. Gilmore. Comparison of tests for embeddings. *Phys. Rev. E*, 78:026203, Aug 2008.

[23] R. V. Donner et al. Recurrence networks a novel paradigm for nonlinear time-series analysis. *New Journal of Physics*, 12:033025, 2010. DOI:10.1088/1367-2630/12/3/033025.

[24] R. V. Donner et al. The geometry of chaotic dynamics - a complex network perspective. *Eur. Phys. J. B*, 2011. DOI:10.1140/epjb/e2011-10899-1.

[25] Y. Zou et al. Identifying complex periodic windows in continuous-time dynamical systems using recurrence-based methods. *CHAOS*, 20:043130, 2010. DOI:10.1063/1.3523304.

[26] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9(1-2):189–208, 1983.

[27] P. L. Read M. Thiel, M. C. Romano and J. Kurths. Estimation of dynamical invariants without embedding by recurrence plots. *CHAOS*, 14(2):234–243, 2004.

[28] N. Marwan A. P. Schutz, Y. Zou and M. T. Turvey. Local minima-based recurrence plots for continuous dynamical systems. *Int. J. Bifurcation and Chaos*, 21(4):1065–1075, 2011. DOI: 10.1142/S0218127411029045.

[29] Jesper Dall and Michael Christensen. Random geometric graphs. *Phys. Rev. E*, 66:016121, Jul 2002.

[30] M. Penrose. *Random Geometric Graphs*. Oxford University Press, Oxford, 2003.

[31] T. Hsing and H. Rootzén. Extremes on trees. *Ann. Probab.*, 33(1):413–444, 2005.

[32] C. Cooper and A. Frieze. Component structure of the vacant set induced by a random walk on a random graph. *Random Structures and Algorithms*, 42(2):135–158, 2013. DOI:10.1002/rsa.20402.

[33] Z. Kong and E. M. Yeh. On the critical density for percolation in random geometric graphs. *Proceedings of IEEE International Symposium on Information Theory, ISIT 2007*, 1-7:151–155, 2007. DOI:10.1109/ISIT.2007.4557082.

[34] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284–304, 1949.

[35] R. Preis et al. Dominant paths between almost invariant sets of dynamical systems. 2004. preprint available in `http://www.biocomputing-berlin.de/biocomputing/en/?cmd=publication`.

[36] L. Putzig P. Metzner and I. Horenko. Analysis of persistent non-stationary time series and applications. *CAMCoS*, 7(2):175229, 2012. DOI: 10.2140/camcos.2012.7.1753.

[37] M. Munz and P. C. Biggin. Jgromacs: a java package for analyzing protein simulations. *J. Chem. Inf. Model*, pages 255–259, 2012.

[38] A. Fischer et al. Identification of biomolecular conformations from incomplete torsion angle observations by hidden markov models. *J. Comput. Chem.*, 28:2453–2464, 2006. DOI: 10.1002/jcc.20692.

[39] Rimfrost database. `http://www.rimfrost.no/`.

[40] Deutscher wetterdienst. `http://www.dwd.de/`.

[41] Nasa goddard institute for space studies. `http://data.giss.nasa.gov/gistemp`.

[42] Data base: Urban and environmental information system (ueis) berlin department for urban development and the environment. `http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/ed412_03_zusatz.htm`.

[43] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classication*, 2:193–218, 1985.

[44] C. J. Hassona. Influence of embedding parameters and noise in center of pressure recurrence quantification analysis. *Gait and Posture*, 27(3):416–422, 2008.

[45] C. L. Webber Jr. and J. P. Zbilut. *Recurrence quantification analysis of nonlinear dynamical systems.* Tutorials in contemporary Nonlinear Methods for the Behavioral Sciences Web Book. Riley, National Science Foundation US, 1995.

[46] P. Faure and H. Korn. A new method to estimate the kolmogorov entropy from recurrence plots: its application to neuronal signals. *Physica D*, 122:265–279, 1998.

[47] J. M. Santos and M. Embrechts. On the use of adjusted rand index as a metric for evaluating supervised classification. *Proceedings of the 19th International Conference on Artificial Neural Networks: Part II*, pages 175–184, 2009.

[48] J. M. Hammersleyh. The distribution of distance in a hypersphere. *Ann. Math. Statist.*, 21:447–452, 1950.

[49] O. Amidi S. Toal and R. Schweitzer-Stenner. Conformational changes of trialanine induced by direct interactions between alanine residues and alcohols in binary mixtures of water with gycerol and ethanol. *J. Am. Chem. Soc.*, 133:12728–12739, 2011. DOI: 10.1021/ja204123g.

[50] Y. Xu M. (I-Hsien) Tsai and J. J. Dannenberg. Ramachandran revisited. DFT energy surfaces of diastereomic trialanine peptides in the gas phase and aqueous solution. *J. Phys. Chem. B*, 113:309–318, 2009.