Higher Education
Australian Mathematical Sciences Institute

Analysis and Comparison of the Dynamics of Irregularly Sampled Time Series

Ben B. Luo
Supervised by Dr. Thomas Stemler
The University of Western Australia

Australian Government
Department of Education

Abstract

In the observation of a non-linear system it is often difficult to detect changes in the dynamics of the system from recorded measurements due to several issues. These can include the presence of measurement noise and possibly irregular/random time steps between measurements. A time series or dataset with such random time steps between measurements is called an irregularly sampled time series. One method for analysing such a time series is by defining a metric between two consecutive time frames in the dataset, we used the metric defined by Victor and Purpura in 1997 which contains several cost constants. The set of all such metric measurements is now a regularly sampled series of data points. We then apply recurrence analysis methods to this series to examine the dynamics of the system. This method was applied to several geological datasets (rainfall data) to detect dynamical changes and to determine what values of the constants was optimal. Then a modified version of the code was used to compare two datasets from different locations to see if a similar method could be used to determine relationships between the climate dynamics of the two locations.

1 Introduction

Real world measurements are rarely structured as a regularly sample time series (RSTS) especially if the data is being taken from proxies. For example, in the measurement of rainfall data a core sample of a speleothem (stalagmite/stalactite) is taken and the ratio of stable isotopes ^{18}O to ^{16}O (also known as $\delta^{18}O$)in each layer of the speleotherm is measured. This data can be used as a measurement of the rainfall levels at the time. Each of these layers is then dated using the Uranium-thorium dating technique to provide a time value corresponding to each rainfall measurement.(ref) These time values are not regularly spaced due to how speleothems can form irregularly creating an irregularly sample time series (ISTS).

For such ISTSs we cannot apply the typical methods of time series analysis since they are only applicable to RSTSs and we must consider alternative methods of analysis to determine features of the underlying dynamical system.

In this research project I examined the usage of a metric to compare 'frames' of the dataset, i.e. sets of datapoints within a given time frame. By dividing up the entire dataset into such 'frames' it becomes possible to construct a RSTS of the 'distances' between consecutive frames. Since this dataset is based on the original ISTS it retains some of the underlying dynamics and we can now analyse these dynamics using regular non-linear time series analysis techniques.

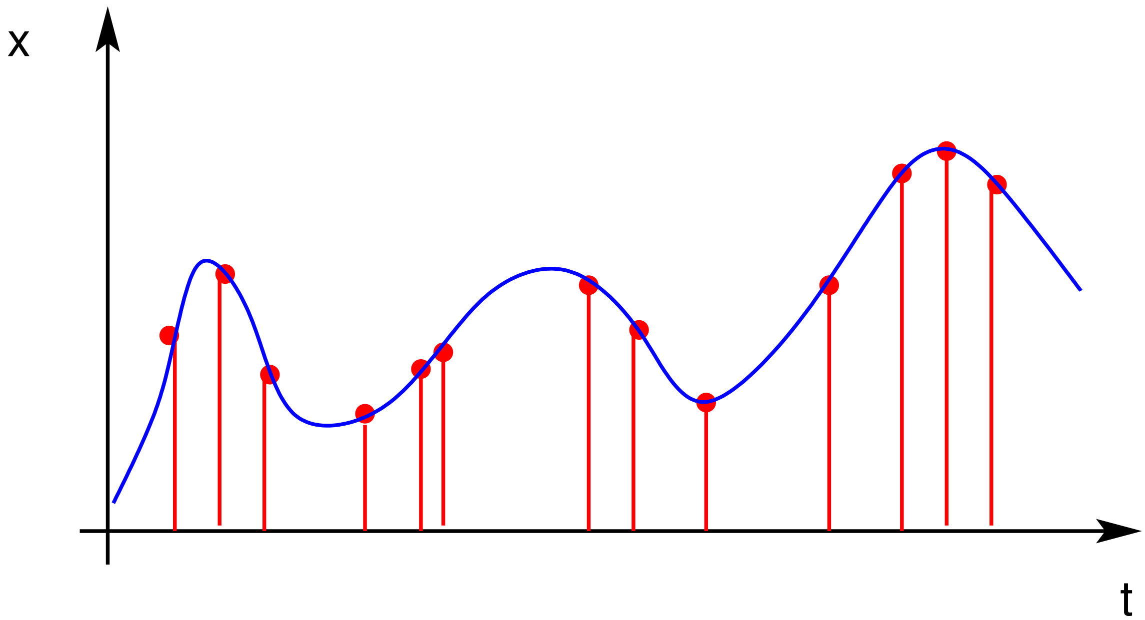

Figure 1: Diagram of example system with variable $\boldsymbol{x}$ plotted against time $\boldsymbol{t}$. Underlying dynamical system represented in blue with irregularly sampled datapoints in red.

2 Construction of Proxy Time Series

To analyse a irregularly sample times series, we create a 'proxy' dataset that is regularly sampled. This is done by utilising a metric first designed by Victor and Purpura in 1997 [1] which we can use to compare any given pair of time frames. This metric measures the 'cost' of transforming one frame into another using three operations, changing the time coordinate of a datapoint (also called a 'spike), raising or lowering the 'height' of a spike and creating/deleting a spike. The distance between two frames using this metric is described by the following equation:

$$d(U,V) = \Sigma_{\theta\in\Theta}\lambda_t\Delta t_\theta + \Sigma_{\phi\in\Phi}\lambda_x\Delta x_\phi + (n_c + n_d)\lambda_{cd} \tag{1}$$

where U and V are sets of ordered pairs (t, x) with t being within two different given ranges (i.e. all the spikes in a time frame), Θ is the set of time change operations, Φ is the set of height change operations and n_c and n_d are the number of creation and deletion operations respectively.

We are then required to determine the constants λ_t, λ_x and λ_{cd}. These constants need to be such that they produce appropriate results. For instance, λ_{cd} cannot be too low otherwise the distance between frames would depend purely on the difference in number of spikes in each frame. Intuitively, λ_t should be inversely proportional to the average spacing of spikes as the more spread out the spikes the less it should cost to shift their time position. Similarly λ_x should be inversely proportional to the average spike height. We will discuss how to determine an appropriate λ_{cd} later on.

Using this cost metric we can now create a proxy regularly sample dataset. To do this we split the dataset up into equally sized frames and calculate the minimum distance between each pair of consecutive frames. We then assign a time value to each calculated distance based on the starting time value of the first frame. This means we now have a regularly spaced time series which we can analyse in place of our original ISTS. To determine the appropriate value for λ_{cd} we need to look at the distribution of costs. This should approximate a normal distribution as testing this method on generated datasets showed that the dynamics inferred from the proxy dataset were most accurate and representative of the original dynamics if the costs were distributed normally (possibly with a slight skew). Note that we need to determine the minimum cost needed to transform one frame into the next, this is done with an optimisation algorithm.

Dist.jpg

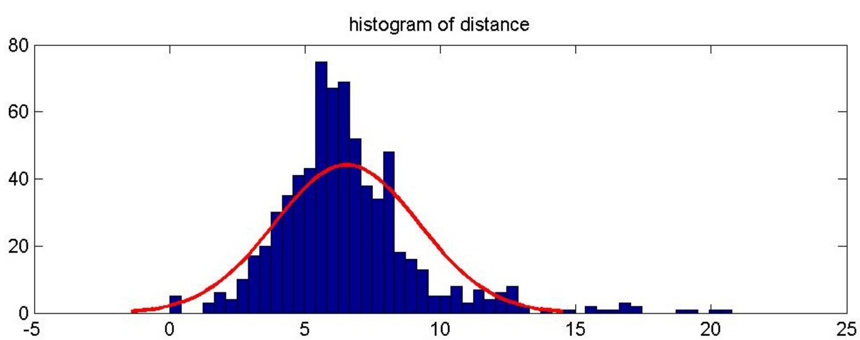

Figure 2: Distribution of costs for a speleotherm dataset overlayed with a normal distribution with the same mean and standard deviation.

3 Time Series Recurrence Plot Analysis

Using this proxy time series we then proceed to use non-linear time series analysis techniques to determine its dynamics.[2] In particularly we created a recurrence plot for segment of the distance time series. A recurrence plot is simply a matrix R with ith and jth entries determined by the following equation:

$$R(i,j) = \begin{cases} 1, |\rho_i - \rho_j| < \epsilon \\ 0, |\rho_i - \rho_j| \geq \epsilon \end{cases} \tag{2}$$

with ϵ being a variable that we can optimise for later on in the process. A recurrence plot is a good visual representation of the dynamics of a system. But it is possible to quantify the dynamics shown in a recurrence plot using a variety of measures, the one we chose to focus on is called the determinism and has the following definition:

$$Determinism = \frac{\Sigma_{l=l_{min}}^{N} l \times P(l)}{\Sigma_{i=1}^{N} \Sigma_{j=1}^{N} R(i,j)} \tag{3}$$

where P(l) is the frequency distribution of the length l of diagonals in the recurrence plot and N is the number of datapoints in our sampled segment. We can now quantify the dynamics of this larger 'window' of successive frames by calculating its determinism. We repeat this for consecutive (overlapping) windows allowing us to observe the evolution of the determinism of the system over time. This provides us with some insight into how the underlying governing factors of the system change over time. We choose ϵ such that our results best correlate with known changes in the system's dynamics.

4 Application

This technique was applied to a dataset of speleotherm $\delta^{18}O$ measurements taken from the KNI-51 cave in the northern part of Western Australia. This data spanned a time period from approximately 8000 years ago to the present day.

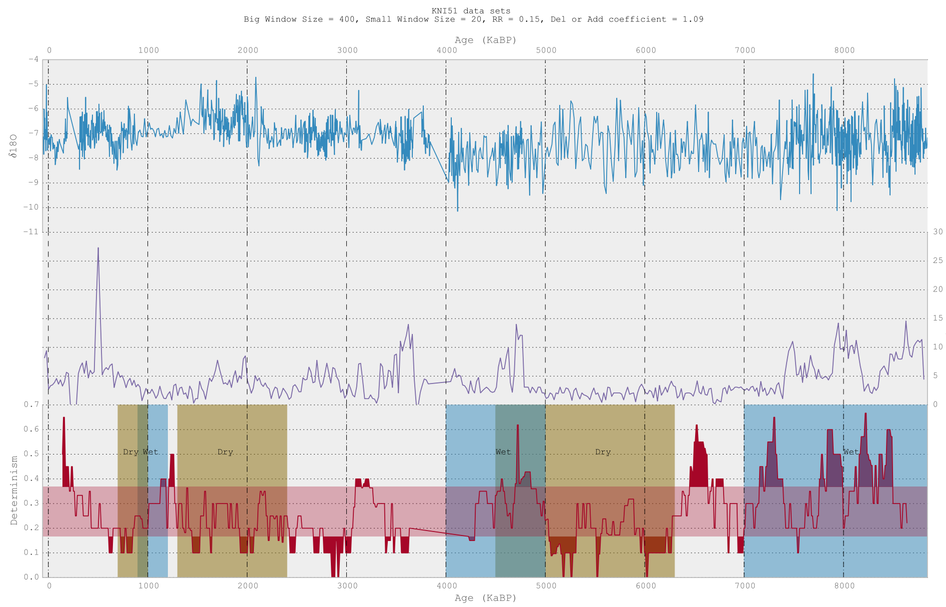

Figure 3: $\delta^{18}O$ data taken from the KNI-51 site along with the cost proxy dataset and a plot of the determinism over time. The highlighted areas on the determinism plot indicate known information from other sources about the local weather patterns.

As is evident from Figure 3. there is a high correlation between known periods of wet weather and high determinism values and similarly for dry weather and low determinism. This allows us to infer information about the time periods for which there is no other known information for instance the period between 6300 and 7000 years ago appears to be also a wet period. Keep in mind this changing value of determinism simply shows a strong shift in underlying dynamics, there is not necessarily always a correlation between dryness and low determinism and viceversa. But from this example we can clearly see how this method can be applied to gain additional knowledge from irregularly sampled data.

5 Comparing Datasets

We also explored using a modified version of this method to see if it would be possible to use it to examine the relationships between dynamical system, for instance the climate systems of two different locations. To do this, the main modification necessary is that in the construction of the proxy distance time series we compare corresponding time frames between two datasets and record the 'distance' between them rather than consecutive time frames of the same dataset. This method was used to compare the KNI-51 dataset and a set from the Dongge cave in southeastern China. Preliminary calculations have been carried out but so far are inconclusive.

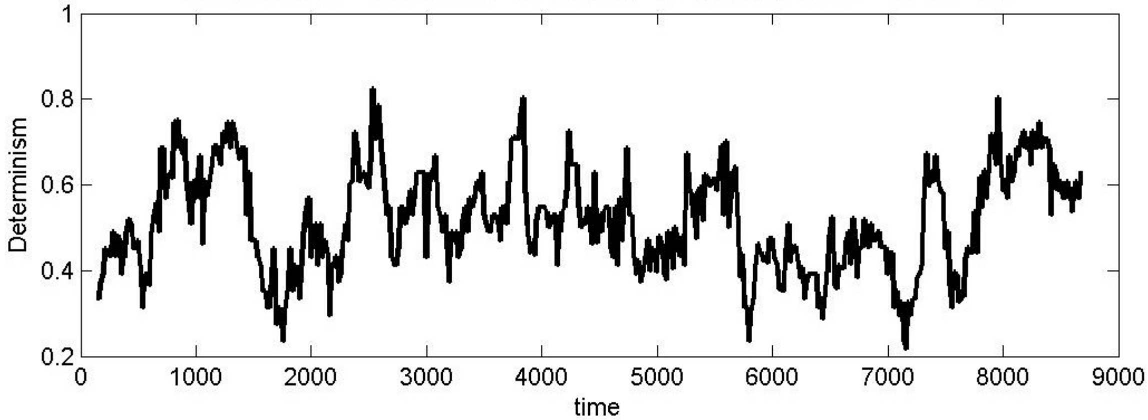

Figure 4: Plot of determinism for 'distance' time series between KNI-51 and Dongge datasets.

6 Conclusion

This method of analysing ISTSs is still in development but early results seem promising, particularly its application to paleoclimate related datasets. The creation of a proxy regularly sample time series allows for some of the dynamics of the original underlying dynamical system to be analysed. Further modification of this method may be useful in examining relationships or interactions between dynamical systems.

References

[1] Victor, JD & Purpura, KP, 1997, 'Metric-space analysis of spike trains: theory, algorithms and application', *Network: Computation in Neural Systems*, Vol.8, pp. 127-164

[2] Marwan, N, Romano, MC, Thiel, M, Kurths, J, 2007, 'Recurrence Plots for the Analysis of Complex Systems', *Physics Reports* 438 (5-6): pp. 237-329

[3] Kantz, H & Schreiber, T, 2004, *Nonlinear Time Series Analysis*, 1st Edition, Cambridge University Press, London.