Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Physica D 200 (2005) 171–184

PHYSICA D

www.elsevier.com/locate/physd

# Recurrence plot statistics and the effect of embedding

T.K. March[a,*], S.C. Chapman[a], R.O. Dendy[a,b]

[a] *Space and Astrophysics Group, Department of Physics, Warwick University, Coventry CV4 7AL, UK*
[b] *UKAEA Culham Division, Culham Science Centre, Abingdon, Oxfordshire OX14 3DB, UK*

Received 2 February 2004; received in revised form 28 October 2004; accepted 11 November 2004

Communicated by C.K.R.T. Jones

## Abstract

Recurrence plots provide a graphical representation of the recurrent patterns in a timeseries, the quantification of which is a relatively new field. Here we derive analytical expressions which relate the values of key statistics, notably determinism and entropy of line length distribution, to the correlation sum as a function of embedding dimension. These expressions are obtained by deriving the transformation which generates an embedded recurrence plot from an unembedded plot. A single unembedded recurrence plot thus provides the statistics of all possible embedded recurrence plots. If the correlation sum scales exponentially with embedding dimension, we show that these statistics are determined entirely by the exponent of the exponential. This explains the results of Iwanski and Bradley [J.S. Iwanski, E. Bradley, Recurrence plots of experimental data: to embed or not to embed? Chaos 8 (1998) 861–871] who found that certain recurrence plot statistics are apparently invariant to embedding dimension for certain low-dimensional systems. We also examine the relationship between the mutual information content of two timeseries and the common recurrent structure seen in their recurrence plots. This allows time-localized contributions to mutual information to be visualized. This technique is demonstrated using geomagnetic index data; we show that the AU and AL geomagnetic indices share half their information, and find the timescale on which mutual features appear.

*PACS:* 07.05.Rm; 07.05.Kf; 89.75.Kd

*Keywords:* Recurrence plots; Recurrence quantification analysis

## 1. Introduction

Patterns are ubiquitous in nature, where their presence may imply inherent predictability. As a result there is great interest in developing methods for detecting and quantifying patterns, leading to quantitative measures of

[*] Corresponding author: Tel.: +44 24 7657 3874; fax: +44 24 7652 3672.
*E-mail address:* tom.march@astro.warwick.ac.uk (T.K. March).

0167-2780/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2004.11.002

structure, similarity, information content, and predictability. Here we consider recurrence plots, which offer a means to quantify the pattern within a timeseries, and also the pattern shared between two timeseries.

Recurrence plots are a method for visualizing recurrent patterns within a timeseries or sequence. They were first proposed in 1981 by Maizel and Lenk [1] as a method of visualizing patterns in sequences of genetic nucleotides. They have since been introduced into the study of dynamical systems [2], where much effort has been put into building quantification schemes for the plots and for the patterns within them. There are now many quantitative recurrence plot measures available [3,4]. These have been applied with success to patterns as diverse as music [5], climate variation [6], heart rate variability [7], webpage usage [8], video recognition [9], and the patterns in written text and computer code [10].

In outline, a data series $S$ can be considered as a set of $n$ scalar measurements

$$S = \{s_1, s_2, s_3, \dots, s_n\} \tag{1}$$

from which a sequence of $N$ $d$-dimensional vectors $\mathbf{a}_k$ can be constructed using a procedure known as time-delay embedding. The vectors are defined as

$$\mathbf{a}_k = \{s_k, s_{k+\tau}, s_{k+2\tau}, \dots, s_{k+(d-1)\tau}\}, \tag{2}$$

where $\tau$ is a delay parameter and $d$ is known as the embedding dimension, [11]; these parameters are typically chosen independently of the recurrence plot technique, for example see [12]. A recurrence plot is constructed by considering whether a given pair of these coordinates are nearby in the embedding space. Typically, the maximum norm is used,

$$\|\mathbf{a}_i - \mathbf{a}_j\| \equiv \max_k\{|s_{i+k} - s_{j+k}|\} \tag{3}$$

so that the distance between two coordinates equals the maximum distance in any dimension. A recurrence plot is represented by a tensor $T_{ij}^d$ whose elements correspond to the distance between each of the $N^2$ possible pairs of coordinates $\mathbf{a}_i$, $\mathbf{a}_j$ [2]:

$$T_{ij}^A = \Theta(\epsilon - \|\mathbf{a}_i - \mathbf{a}_j\|), \tag{4}$$

where $\Theta$ is a step function (0 for negative arguments, 1 for positive arguments). For each pair of coordinates in the series whose separation is less than the threshold parameter $\epsilon$, $T_{ij}$ takes the value unity, which can be plotted as a black dot on an otherwise white graph.

A recurrence plot of independent and identically distributed (IID) data appears as a random scattering of black dots, while a regularly repeating signal (such as a sine wave, e.g. see Fig. 1 of [13]) appears as a series of equally spaced, 45° diagonal black lines. An irregularly repeating signal (such as the output of a chaotic system) typically appears as a pattern of small diagonal lines of varying length. Paling of the plot away from the main diagonal indicates that the longer one observes no repeat of a particular feature, the less likely a repeat is to occur. In this case it follows that probability depends on time, and therefore that the process which generated such data is non-stationary.

In this paper we investigate the statistics of recurrence plots, and their meaning in relation to well-understood statistics from nonlinear timeseries analysis. First, we examine the meaning of two of the key statistics in recurrence quantification analysis (RQA), namely the determinism and the entropy of line length distribution [3], and the effect on them of the time-delay embedding procedure [14,11]. Iwanski and Bradley [13] found that the appearance and statistics of recurrence plots for certain low-dimensional systems are not significantly altered by a small change in the embedding dimension $d$, suggesting that these statistics may be important new invariant characteristics of a system. However, unlike traditional measures where invariance relies on the embedding dimension being sufficiently high, Iwanski and Bradley found the same statistics for an unembedded recurrence plot as for an embedded version. This was further examined by Gao and Cai [15], who suggested that many recurrence plot statistics may rely on information from a higher embedding dimension than was used to construct the recurrence plot. However this does not completely explain why these quantities appear to be invariant with respect to the embedding dimension;

nor whether these quantities are independent of each other, or of other better known measures. This is important, since independent quantities potentially yield new information about a system. In Section 2 we show that all embedded recurrence plots are present within the unembedded plot, accessible via a simple transformation. Using this transformation, we derive in Section 3 the effect of embedding on two RQA statistics: determinism, and entropy of line length distribution. For the case of exponential scaling of the correlation sum [see Eq. (9)] with embedding dimension, which might be expected for certain low-dimensional systems, we derive expressions which relate these quantities to the Kolmogorov entropy rate [14]. This is important for two reasons. First, it provides a new perspective on the physical meaning of these quantities. Second, it can be used to establish baseline values for independent and identically distributed (IID) processes, above or below which a measurement can be said to be significant.

In Section 4, we examine the converse question of how well-known statistics from nonlinear timeseries analysis relate to recurrence plots. We demonstrate that a standard algorithm for computing the mutual information between two timeseries is related to counting the number of black dots common to the recurrence plots of the two timeseries in question. This suggests the definition of a new form of cross-recurrence plot which, when drawn, allows contributions to the mutual information to be visualized. We apply this technique to a physical system in which issues of predictability and correlation are of practical interest. Earth's geomagnetic activity is monitored by a non-uniformly distributed circumpolar ring of magnetometers, which measure fluctuations in horizontal magnetic field strength due to enhancements in auroral activity. These measurements are compiled to form the AE geomagnetic indices [16], of which we consider AU (a proxy for the maximum eastward flowing polar current) and AL (a proxy for the maximum westward flowing current). In common with many other “real world" timeseries, these timeseries show both low- and high-dimensional behavior, in this case well-defined features on timescales of days (storms) which are embedded in colored noise [17].

## 2. Effect of embedding dimension

We now derive a transformation which generates an embedded recurrence plot from an unembedded recurrence plot. This result is central to the subsequent discussion of the effect of embedding on statistics derived from recurrence plots. A single recurrence on an umbedded $d = 1$ plot is represented by a single black dot, corresponding to a pair of data points closer together than $\epsilon$. If we consider Fig. 1 (left) to represent part of a $d = 1$ recurrence plot, the example illustrated relates to points numbered 2 and 8, i.e.

$$|a_2 - a_8| < \epsilon. \tag{5}$$

Fig. 1 (right) shows a line of length 2. Still taking $d = 1$, the situation represented is

$$|a_2 - a_8| < \epsilon \quad \text{and} \quad |a_3 - a_9| < \epsilon. \tag{6}$$

Consider forming coordinates in a $d = 2$, $\tau = 1$ embedding space [see Eq. (2)]. If we consider Fig. 1 (left) to represent a region of a $d = 2$ recurrence plot, the black dot now represents

$$\|\mathbf{a}_2 - \mathbf{a}_8\| < \epsilon, \tag{7}$$

where $\mathbf{a}_n$ now denotes $\{a_n, a_{n+1}\}$. Using the maximum norm, Eq. (3), this is equivalent to Eq. (6). Therefore a single dot in $d = 2$ represents a line of length 2 in $d = 1$. The transformation from $d = 1$ to $d = 2$ thus reduces the length of all diagonal lines by one dot. An isolated dot is removed entirely. Formally, we represent the transformation to arbitrary dimension as

$$T_{ij}(d) = T_{ij}(1) \times T_{i+\tau,j+\tau}(1) \times T_{i+2\tau,j+2\tau}(1) \times \cdots \times T_{i+(d-1)\tau,j+(d-1)\tau}(1). \tag{8}$$

An element on the recurrence plot with embedding dimension $d$ is thus related to a diagonal sequence of $d$ elements on the unembedded recurrence plot $T_{ij}(1)$. This transformation enables the conversion of an unembedded recurrence

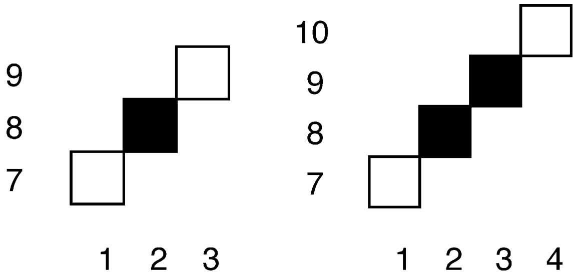


Fig. 1. Representation of diagonal lines of length 1 and 2 on a recurrence plot, corresponding to pairs of points in the original timeseries.

plot into any embedded recurrence plot with any values of $d$ or $\tau$. This suggests that embedding in the construction of recurrence plots is not strictly necessary, since all of the information is contained within the unembedded plot $T_{ij}(1)$; let us refer to this as the parent plot. Rather than performing embedding, information can be extracted directly from this parent plot. Understanding how the information is contained in the parent plot assists in consideration of how various recurrence plot statistics are affected by embedding.

## 3. Meaning of recurrence plot statistics

Given the transformation derived above, we now consider two of the key statistics of recurrence plots, namely the determinism and the entropy of the diagonal line length distribution. We show that both these statistics are related to the correlation sum, and also relate them to the probability distribution of line lengths on an unembedded plot. In the case of exponential scaling of the correlation sum with embedding dimension, we show that they do not depend on the embedding dimension $d$.

Recurrence quantification analysis (RQA) provides a set of statistical measures which have been proposed to quantify patterns based on the lines and dots visible on a recurrence plot [3]. The fraction of the plot colored black is the most fundamental statistic associated with recurrence plots. This is known as the recurrence rate in RQA, and is known elsewhere as the correlation sum $C_d(\epsilon)$ [14,18]. $C_d(\epsilon)$ is the fraction of pairs of coordinates closer together than $\epsilon$, and is defined by

$$C_d(\epsilon) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(\epsilon - \|\mathbf{a}_i - \mathbf{a}_j\|). \tag{9}$$

A recurrence plot can be considered to be a two-dimensional pictorial representation of the points that contribute to Eq. (9) for a particular value of $\epsilon$.

The remaining statistics in RQA are the fraction of black dots involved in diagonal lines, known as the determinism $D_d$, the entropy of the diagonal line length distribution $E_d$, the ratio of determinism to correlation sum, and the slope of the line of best fit on a graph of recurrence probability versus distance from main diagonal, known as the trend [3]. Except for the trend, these statistics can be related to the probability distribution of diagonal line lengths $P_d(L)$, which is the probability of observing a diagonal black line of length $L$ beginning from a randomly selected element of the recurrence plot. From Eq. (8), the distribution of line lengths on an embedded recurrence plot is related to the distribution on an unembedded plot by

$$P_d(L) = P_1(L + d - 1). \tag{10}$$

Hence any statistic formed from the embedded $P_d(L)$ can be constructed from the unembedded $P_1(L+d-1)$. For example, using Eq. (10), the correlation sum can be written as

$$C_d = \sum_{L=d}^{\infty}(L-d+1)P_1(L). \tag{11}$$

This relationship can be reversed to give

$$P_1(L) = C_{L+2} - 2C_{L+1} + C_L. \tag{12}$$

Hence any statistics derived from $P_1(L)$ can also be derived from the correlation sum, as we now explicitly show.

First we consider the determinism $D_d$ [3], which was observed to be invariant to embedding dimension by Iwanski and Bradley [13]. This is the ratio of black dots included in lines of length greater than unity to the total number of black dots. The determinism $D_d$ quantifies the prevalence of lines, and is believed [3] to quantify how deterministic a system is. This can be related to the probability $C_d$ of observing a black dot in a randomly selected location, and to the probability of observing an isolated black dot. The number of black dots included in lines is equal to the total number of black dots minus the number of isolated black dots (lines of length unity), so we can write

$$D_d = \frac{C_d - P_1(d)}{C_d}. \tag{13}$$

Using Eq. (12) to express $P_1(d)$ in Eq. (13), we have

$$D_d = \frac{2C_{d+1} - C_{d+2}}{C_d}. \tag{14}$$

Thus the determinism at embedding dimension $d$ can be inferred from knowledge of the correlation sum at nearby embedding dimensions $d$, $d+1$ and $d+2$.

The next statistic in the RQA is the Shannon entropy of the line length distribution [3]. This is defined as

$$E_d = -\sum_{L=1}^{\infty} Q_d(L)\ \ln\ Q_d(L), \tag{15}$$

where $Q_d(L)$ is the probability of observing a line of length $L$ given the fact that a line is observed. This can be related to the probability $P_d(L)$ of observing a line of length $L$, and the probability of observing a line of arbitrary length. Using Eqs. (10) and (12) we obtain

$$Q_d(L) = \frac{C_{L+d+1} - 2C_{L+d} + C_{L+d-1}}{C_d - C_{d+1}}. \tag{16}$$

Hence, like the determinism, the Shannon entropy of line length distribution can be obtained from the correlation sum.

### 3.1. Exponential scaling of correlation sum

Suppose we assume that the correlation sum $C_d$ can be expressed as an inverse exponential function of $d$ with exponent $K_2$. This is strictly true for data derived from an IID process, and is observed for many low-dimensional chaotic processes under certain conditions [14]; in this case $K_2$ is known as the Kolmogorov entropy rate. It has been previously shown that this can be extracted from the distribution of recurrence plot diagonal line lengths $P_d(L)$ [4]. We write the correlation sum as

$$C_d = A\,\mathrm{e}^{-K_2 d}, \tag{17}$$

where we have absorbed the dependence of $C_d$ on the threshold parameter $\epsilon$ into the constant $A$. Substitution of Eq. (17) into Eq. (12) yields

$$P_1(L) = A(1 - \mathrm{e}^{-K_2})^2\,\mathrm{e}^{-K_2 L}. \tag{18}$$

This implies that $P_1(L)$ is an exponential function of $L$ with the same exponent $K_2$ that governs the dependence of $C_d$ on $d$. This result has been derived independently by an alternative route which considers the divergence of trajectories directly [15]. From Eqs. (13) and (18), the determinism $D_d$ can be written

$$D_d = 1 - \frac{A\,\mathrm{e}^{-K_2 d}(1 - \mathrm{e}^{-K_2})^2}{A\,\mathrm{e}^{-K_2 d}}. \tag{19}$$

This simplifies to give

$$D_d = 1 - \gamma^2, \tag{20}$$

where we define $\gamma = (1 - \mathrm{e}^{-K_2})$. For exponential scaling of $C_d$, the determinism is a constant independent of the embedding dimension $d$ chosen, and is determined by the exponential scaling exponent. Where the correlation sum only exhibits exponential scaling over a limited range of embedding dimensions (such as might be expected for a low-dimensional chaotic process), this expression remains true, since Eq. (14) only relies on knowledge of adjacent (in $d$) correlation sums.

To derive the Shannon entropy of line length distribution, Eq. (15), we insert Eq. (17) into Eq. (16) to give

$$Q_d(L) = (1 - \mathrm{e}^{-K_2})\,\mathrm{e}^{-K_2(L-1)}, \tag{21}$$

which when inserted into Eq. (15) gives

$$E_d = K_2\left(\frac{1}{\gamma} - 1\right) - \ln\gamma. \tag{22}$$

As with $D_d$, this is independent of the embedding dimension $d$. However, unlike Eq. (20) this expression is only true in the case of perfect exponential scaling.

As a demonstration of these results, Fig. 2 shows the correlation sum computed as a function of embedding dimension for the logistic map, $x_{t+1} = \mu x_t(1 - x_t)$, in the chaotic regime with $\mu = 4$. This shows reasonable scaling of the correlation sum with dimension, as in Eq. (17), and yields $K_2 = 0.6349 \pm 0.0004$. By Eq. (20), this implies a value for the determinism $D_d$ of $0.7791 \pm 0.0002$ and by Eq. (22) a value for $E_d$ of $1.4709 \pm 0.0006$. These values are shown in Figs. 3 and 4 as the solid lines, while the actual values computed from recurrence plots of the data are shown as asterisks. Until statistical noise becomes important (around $d = 25$–$30$), the points lie convincingly on the lines.

An initial exponential distribution of diagonal line lengths remains exponential after embedding, explaining the apparent invariance with respect to $d$ of these statistics for low-dimensional chaotic systems [13]. The determinism $D_d$ and the entropy $E_d$ are in this case governed by the exponential scaling exponent of the correlation sum, $K_2$.

A corollary is provided by the results of Zbilut et al. [19], who applied the techniques of recurrence quantification analysis to short sequences of random integers, as well as to the logistic map. There were three sequences considered: (a) consecutive digits of $\pi$; (b) pseudo-random integers generated with MATLAB; (c) experimentally derived random integers, produced by tuning a radio antenna to an empty part of the spectrum [20]. All three were considered with sequence lengths of $N = 1000$, 3000 and 5000, and only exact matches were considered to constitute recurrences. This corresponds to $\epsilon = 0$ in Eq. (4), which is only possible when working with integer sequences; for real-valued sequences, $\epsilon$ is limited by numerical precision. It was found that for (a) and (b) the determinism was slightly below 20%, and was defined up to $d_0 = 4$ for $N = 1000$, $d_0 = 5$ for $N = 3000$ and $d_0 = 6$ for $N = 5000$. However, (c) had a determinism only slightly above 0%, which was defined only up to $d_0 = 2$ regardless of $N$. The authors suggested

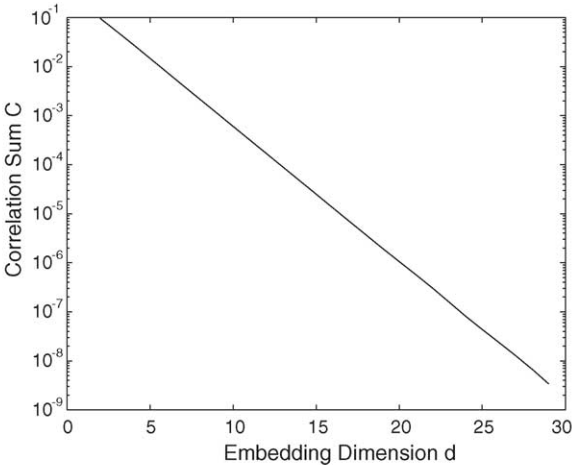


Fig. 2. Correlation sum $C_d$ computed as a function of embedding dimension $d$ for $10^5$ samples of the logistic map with $\epsilon = 0.1$. Applying Eq. (17) to the measured straight line slope gives $K_2 = 0.6349 \pm 0.0004$.

that this was possibly due to some innate randomness that sequence (c) possessed, and suggested the RQA as a test to distinguish between physical and pseudo-random numbers.

For data drawn from an IID process, the probability of a particular dot being black on the $d = 1$ plot is a constant $C_1$, the correlation sum. Referring back to the definition of Eq. (17), we can write

$$\gamma = 1 - C_1, \tag{23}$$

$$D_d = C_1(2 - C_1), \tag{24}$$

$$E_d = \frac{1}{C_1} \ln C_1 - \ln(1 - C_1). \tag{25}$$

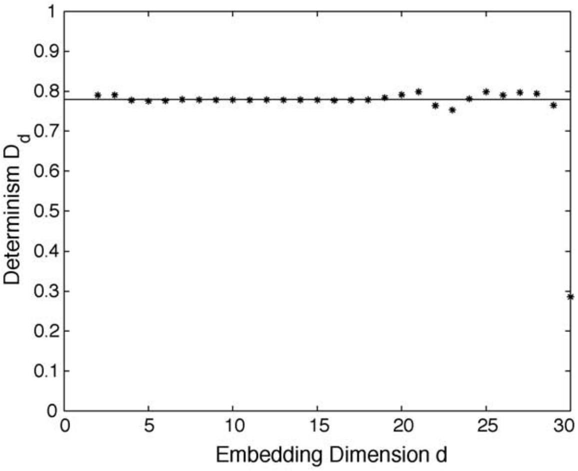


Fig. 3. Determinism $D_d$ computed as a function of embedding dimension $d$ for $10^5$ samples of the logistic map with $\epsilon = 0.1$, shown as asterisks. Solid line shows theoretical prediction of 0.7791 obtained from Eq. (20) using the measured value of $K_2$ from Fig. 2.

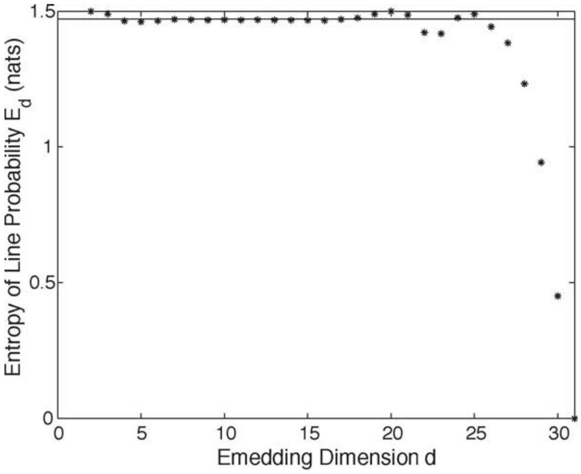


Fig. 4. Shannon entropy of line probability distribution $E_d$ computed as a function of embedding dimension $d$ for $10^5$ samples of the logistic map with $\epsilon = 0.1$, shown as asterisks. Solid line shows theoretical prediction of 1.4709 obtained from Eq. (22) using the measured value of $K_2$ from Fig. 2.

These quantities are finite for IID data because a small number of lines are created by chance. To conclude that observed data are non-random, the values measured must be compared with Eqs. (24) and (25) to establish the significance of the result.

Integer sequences can be represented as a string of symbols from an alphabet of size $m$. The probability of observing a black dot in a randomly selected location on the unembedded $d = 1$ plot is given by

$$C_1 = \sum_{i=1}^{m} p_i^2, \tag{26}$$

where $p_i$ is the probability of observing symbol $i$ from the alphabet. The sequences (a)–(c) were all uniformly distributed so we can write $p_i$ as

$$p_i = \frac{1}{m}, \tag{27}$$

and

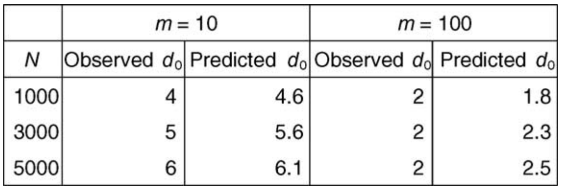
$$C_1 = \sum_{i=1}^{m} \frac{1}{m^2} = \frac{1}{m}. \tag{28}$$

The measured determinism values [19] died out above a particular value of $d$, when no diagonal black lines were seen on a finite recurrence plot. To estimate the embedding dimension at which this should occur, we examine the expected number of lines $\langle n \rangle$ on an embedded recurrence plot. This is given by the total number of elements on the plot multiplied by the probability of observing, in a randomly selected location, one white dot diagonally followed by $d + 1$ black dots on the unembedded recurrence plot:

$$\langle n \rangle = \tfrac{1}{2} N(N-1)(1-C_1)C_1^{d+1}. \tag{29}$$

Setting $\langle n \rangle$ equal to unity gives an estimate for $d_0$, the dimension where the determinism should die out:

$$d_0 \approx \frac{\log 2 - \log N(N-1) - \log C_1(1-C_1)}{\log C_1}. \tag{30}$$



| | $m = 10$ | | $m = 100$ | |
|---|---|---|---|---|
| $N$ | Observed $d_0$ | Predicted $d_0$ | Observed $d_0$ | Predicted $d_0$ |
| 1000 | 4 | 4.6 | 2 | 1.8 |
| 3000 | 5 | 5.6 | 2 | 2.3 |
| 5000 | 6 | 6.1 | 2 | 2.5 |

Fig. 5. Observed and predicted [from Eq. (30)] embedding dimension $d_0$ at which determinism $D_d$ drops to zero, as a result of finite sample size $N$, for sequences of symbols from an alphabet of size $m = 10$ symbols (columns 2 and 3) and $m = 100$ symbols (columns 4 and 5). Observed values from [19]: for $m = 10$, consecutive digits of $\pi$ and pseudo-random integers generated with MATLAB; for $m = 100$, experimentally derived random integer sequence from http://www.random.org.

Using Eqs. (24) and (30) with $m = 10$ symbols we obtain $C_1 = 0.1$ and $D_d = 19\%$. From Eq. (30), this should be measurable a priori up to $d_0 \approx 4.6$ for $N = 1000$, $d_0 \approx 5.6$ for $N = 3000$ and $d_0 \approx 6.1$ for $N = 5000$, see Fig. 5. Comparing these values with the measured results [19], we infer that (a) and (b) behave exactly as would be expected for an IID process with no additional distinguishing properties.

To explain the results for the experimentally derived random integers (c), we consider sequences of random integers from the same source [20]. The sequences supplied default to the range 1–100, an alphabet of $m = 100$ symbols. For this value of $m$ we obtain $C_1 = 0.01$ and from Eq. (24) we predict a value of determinism $D_d = 1.99\%$. This should persist up to $d_0 \approx 1.8$ for $N = 1000$, $d_0 \approx 2.3$ for $N = 3000$ and $d_0 \approx 2.5$ for $N = 5000$; this information is summarized in Fig. 5. This agrees with the result reported in [19], so that there is no reason to infer any additional randomness property for (c); the results of recurrence quantification analysis can be explained as a consequence of the different number of symbols in the sequence.

## 4. Mutual information

A recurrence plot can be considered as a visualization of the double summation in the definition of the correlation sum, Eq. (9). It is therefore reasonable to expect that a proportion of the statistics derived from recurrence plots would be related to $C_d$. Conversely, it would also be reasonable to expect that existing statistics related to $C_d$ could be derivable from recurrence plots. A recurrence plot would then provide a visualization of any such statistic. As an

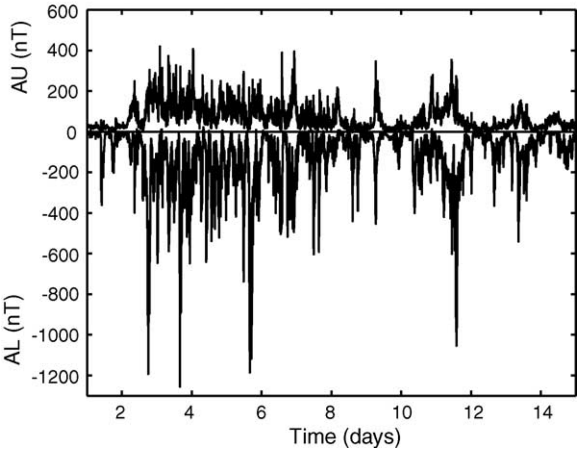


Fig. 6. Days 1–14 of the AU and AL timeseries for the year 1995. AU, being the maximum reading from a network of magnetometer stations, is mostly positive, while AL is mostly negative.

example we consider the mutual information, which is a nonlinear measure of correlation between two (or more) discrete timeseries. The mutual information $I^{AB}$ between timeseries **A** and **B** is defined by

$$I^{AB} = H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A}, \mathbf{B}), \tag{31}$$

where $H(\mathbf{A})$ is the entropy measured for timeseries **A** and $H(\mathbf{A}, \mathbf{B})$ is the joint entropy, measured from a joint histogram. For a discrete timeseries, the Shannon entropy is defined by [21]

$$H = -\sum_i p_i \log_2 p_i, \tag{32}$$

where $p_i$ is again the probability of observing symbol $i$ and the summation is taken over all $i$.

There are two standard algorithms for computing the entropy $H$. The first, [12], discretizes the data using a hierarchy of partitions which become finer in regions of the joint histogram that contain more points. The second approach, [22], uses the second Renyi entropy [23] which is approximated by the logarithm of the correlation sum. Hence we can write the second Renyi mutual information as

$$I_2^{AB} \approx \log_2 C^{AB} - \log_2 C^A - \log_2 C^B, \tag{33}$$

where $C^{AB}$ is the joint correlation sum, which is the recurrence rate of the following type of cross-recurrence plot

$$T_{ij}^{AB} = T_{ij}^{A} T_{ij}^{B}. \tag{34}$$

This definition of a cross-recurrence plot differs from the standard definition [24], but has been recently proposed by Romano et al. [25] as a visualization of recurrent structure common to two timeseries. Thus we can obtain a standard mutual information estimate from three recurrence plots: $T_{ij}^A$, $T_{ij}^B$ and $T_{ij}^{AB}$.

The mutual information depends on the values of $C^A$ and $C^B$, which in turn are conditioned by $\epsilon^A$ and $\epsilon^B$, the threshold parameters used to produce the two auto-recurrence plots. These two parameters must be chosen in some fashion, and this choice must be justified. One solution is to choose the thresholds such that the resulting auto-recurrence plots have the same correlation sum. That is

$$C^A(\epsilon^A) = C^B(\epsilon^B) = C_0. \tag{35}$$

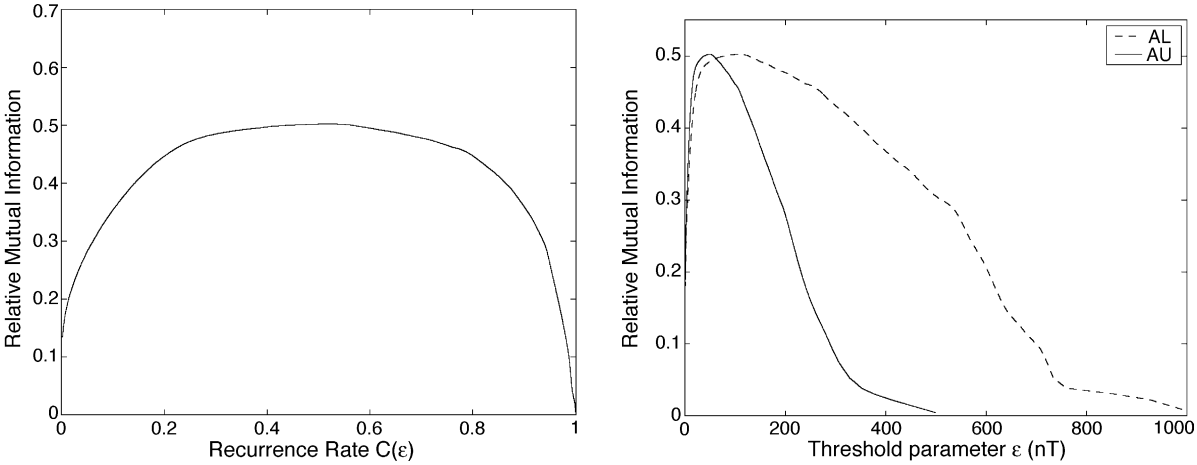


Fig. 7. Mutual information $I$ for AU and AL geomagnetic timeseries, normalized to entropy of AU and AL separately. Left: as a function of correlation sum $C_0$, see Eq. (39); right: as a function of the recurrence threshold parameter $\epsilon$ necessary to create the corresponding underlying thresholded recurrence plots for each measurement.

This choice can be simplified by defining an unthresholded recurrence plot in terms of the measured correlation sum of the timeseries

$$U_{ij}^{A} = C_d^{A}(\|\mathbf{a}_i - \mathbf{a}_j\|). \tag{36}$$

This recurrence plot has the property that if it is thresholded, then the resulting thresholded plot will have a recurrence rate (correlation sum) equal to the thresholding parameter. The corresponding unthresholded cross-recurrence plot will now be given by

$$U_{ij}^{AB} = \max\{U_{ij}^{A}, U_{ij}^{B}\} \tag{37}$$

since the definition of a thresholded recurrence plot uses the maximum norm Eq. (3). This allows us to write the joint correlation sum as a function of the elements of the joint recurrence plot

$$C^{AB}(C_0) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(C_0 - U_{ij}^{AB}). \tag{38}$$

Thus the joint correlation sum is equal to the recurrence rate of the unthresholded joint recurrence plot after it has been thresholded with a threshold parameter equal to $C_0$. Following Eq. (33) we then write the mutual information as

$$I^{AB}(C_0) = \log_2 C^{AB}(C_0) - 2\log_2 C_0. \tag{39}$$

To demonstrate the quantitative practical use of this technique, we now apply it to the geomagnetic AU and AL timeseries for the year 1995. AU reflects the activity of eastward flowing polar currents, induced in the atmosphere by activity deeper in earth's magnetosphere. AL reflects the activity of westward currents, and is typically negative. Fig. 6 shows these timeseries for the first 2 weeks of 1995. AU and AL typically come from opposite sides of the polar current system; they are therefore expected to share a certain amount of information due to large scale phenomena (storms) which are seen in both AU and AL, but to have differences due to smaller fluctuations arising from local phenomena. We use data for the entire year in order to get good statistics. Statistical noise acts to decrease

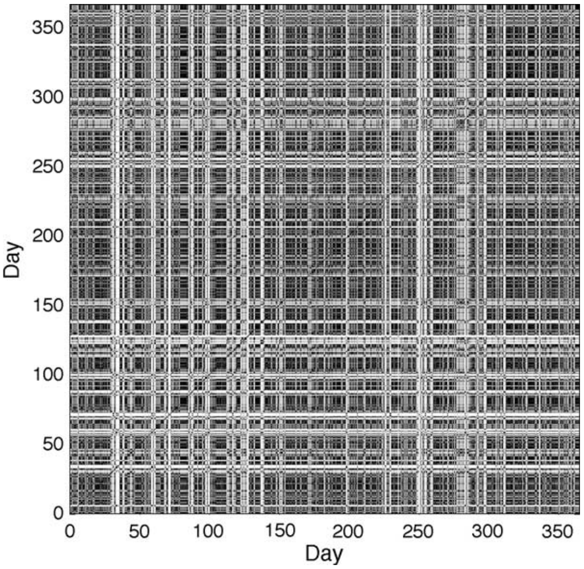


Fig. 8. Unthresholded recurrence plot of geomagnetic AU timeseries.

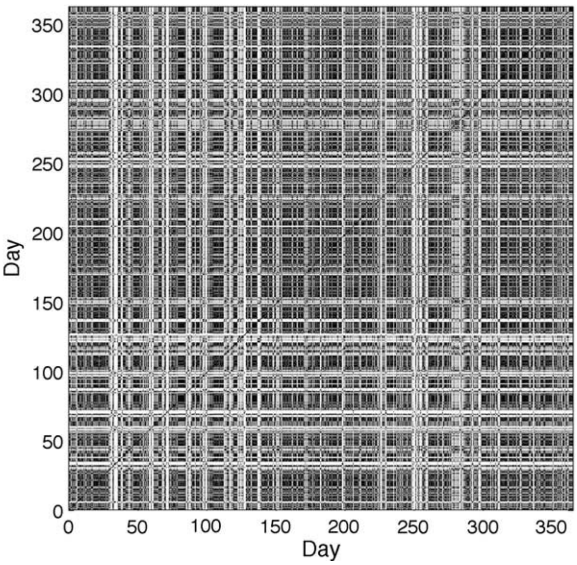


Fig. 9. Unthresholded recurrence plot of geomagnetic AL timeseries.

the measured mutual information. The variance, due to noise, of mutual information measurements has been shown to scale with $1/N$ [26], where $N$ is the number of data points and here we have $N = 5 \times 10^5$.

Within the AU and AL timeseries, three distinct classes of behavior are recognized phenomenologically: quiet time, storms and substorms. During quiet time, measurements of the order of a few nT to a few tens of nT are seen. The other extreme is seen during a magnetic storm, with measurements of hundreds of nT persisting for times of the order of several days. These events correlate strongly with features on the Sun facing the Earth [27], and thus tend

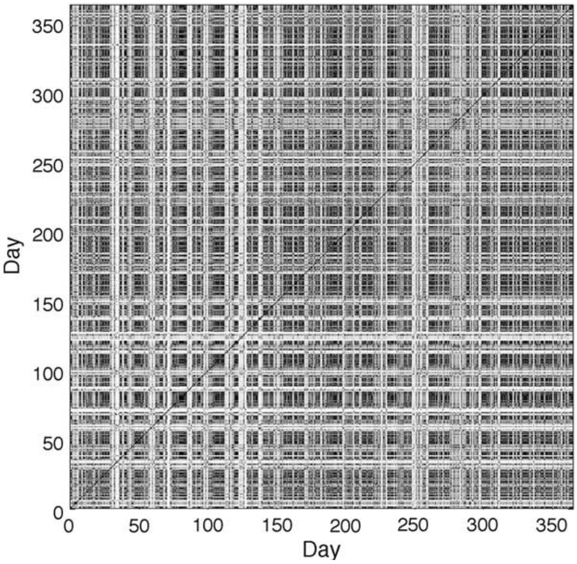


Fig. 10. Unthresholded cross-recurrence plot formed from those shown in Figs. 8 and 9.

to recur on a 27–28 days timescale (the synodic rotation period of the Sun). The intermediate event is a substorm [28], during which variations on the scale of tens to hundreds of nT persist for a few hours. Substorms are believed to result from the sudden release of stored energy built up in the magnetotail by the solar wind.

Fig. 7 shows on the left the functional form of $I(C_0)$, the mutual information as a function of correlation sum of the underlying recurrence plots, obtained for the AU and AL timeseries. On the right are $I(\epsilon^A)$ and $I(\epsilon^B)$, the mutual information as a function of the underlying threshold parameters, constructed using Eq. (35). Both figures are normalized to the entropy of AU and AL considered individually. The maximum fractional mutual information measured is 50% and corresponds to an underlying correlation sum of $C_0 = 0.52$. To obtain this value of $C_0$, the two underlying thresholded recurrence plots require thresholds of $\epsilon$ = 49 nT for AU and $\epsilon$ = 103 nT for AL. On the right of Fig. 7 is the functional form of $I(\epsilon)$, the mutual information as a function of the thresholds applied to the two underlying thresholded recurrence plots, again normalized to the entropy of AU and AL. The solid line shows the relative mutual information as a function of the threshold applied to the AU recurrence plot, while the dotted line shows the same for AL.

Figs. 8 and 9 show unthresholded recurrence plots, as defined by Eq. (36) for the AU and AL geomagnetic timeseries, respectively. The cross-recurrence plot formed from these using Eq. (37) is shown in Fig. 10. These plots show which positions in the original timeseries contribute the most to the mutual information—in this case the dark areas on the cross-recurrence plot correspond to the gaps between magnetic storms. We conclude that the mutual information being measured between AU and AL results from magnetic storms appearing in both timeseries.

## 5. Conclusions

Recurrence plots are extremely versatile: they analyse a stream of data by comparing segments of it to other segments taken at earlier and later times. The data stream itself is thus used as an analysis tool, without any assumptions about the nature of the process that produced it. There are many statistical measures associated with recurrence plots, some of which are unique to recurrence plot analysis. Here we have described two of the most common statistics, and have demonstrated that they are related to better known measures from nonlinear timeseries analysis. In the case of exponential scaling of the correlation sum with embedding dimension, the determinism and entropy of line length distribution have been shown to be determined by $K_2$. This explains the results of [13,19].

We have also shown that all recurrence plots are contained within a single parent plot which contains all of the statistics of its children. It is not strictly necessary to construct recurrence plots for a variety of embedding parameters, because the key statistics that we have considered are all contained within this parent plot, and many of these are directly derivable from the distribution of diagonal line lengths. This demonstrates clearly the effect of embedding on recurrence plots.

A further result is that the mutual information between two timeseries can be obtained from their recurrence plots, and is related to counting the number of shared black dots. Similar comparisons of unthresholded recurrence plots yield the mutual information as a function of the threshold parameter $\epsilon$. This allows time-localized contributions to the mutual information to be assessed and quantified, as we have shown for the example of geomagnetic indices.

Comparisons between repeated patterns in signals from nonlinear systems are particularly valuable when the systems in question are spatially extended and evolve in a non-stationary fashion. Macroscopic plasmas, whether naturally occurring or created in fusion experiments, often fall into this category, which presents a substantial challenge to the techniques of statistical and time series analysis; see, for example, the discussions in recent studies of astrophysical[29], solar[30], and fusion[31] plasma observations. The successful application of recurrence plots and the concepts of information theory to the geomagnetic plasma timeseries studied in the present paper is, therefore, encouraging.

## Acknowledgements

This work was funded in part by the United Kingdom Engineering and Physical Sciences Research Council, and by the Particle Physics and Astronomy Research Council.

## References

[1] J.V. Maizel Jr., R.P. Lenk, Enhanced graphic matrix analysis of nucleic acid and protein sequences, Proc. Natl. Acad. Sci. U.S.A. 78 (1981) 7665–7669.

[2] J.P. Eckmann, S.O. Kamphorst, D. Ruelle, Recurrence plots of dynamical systems, Europhys. Lett. 4 (1987) 973–977.

[3] C.L. Webber Jr., J.P. Zbilut, Dynamical assessment of physiological systems and states using recurrence plot strategies, J. Appl. Physiol. 76 (1994) 965–973.

[4] P. Faure, H. Korn, A new method to estimate the Kolmogorov entropy from recurrence plots: its application to neuronal signals, Physica D 122 (1998) 265–279.

[5] J. Foote, M. Cooper, Visualizing musical structure and rhythm via self-similarity in: Proceedings of the International Conference on Computer Music, Havana, Cuba, 2001.

[6] N. Marwan, J. Kurths, Nonlinear analysis of bivariate data with cross-recurrence plots, Phys. Lett. A 302 (2002) 299–307.

[7] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, J. Kurths, Recurrence-plot-based measures of complexity and their application to heart-rate-variability data, Phys. Rev. E 66 (2002) 026702.

[8] M. Bernstein, J.D. Bolter, M. Joyce, E. Mylonas, Architectures for volatile hypertext, in: Proceedings of the Third Annual ACM Conference on Hypertext, San Antonio, TX, USA, 1991, pp. 243–260.

[9] R. Cutler, L. Davis, Robust periodic motion and motion symmetry detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, South Carolina, USA, 2000.

[10] K.W. Church, J.I. Helfman, Dotplot: a program for exploring self-similarity in millions of lines of text and code, J. Am. Statist. Assoc. 2 (1993) 153–174.

[11] F. Takens, Detecting Strange Attractors in Turbulence, vol. 898 of Lecture Notes in Mathematics, Springer, New York, 1981.

[12] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134–1140.

[13] J.S. Iwanski, E. Bradley, Recurrence plots of experimental data: to embed or not to embed? Chaos 8 (1998) 861–871.

[14] H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, Cambridge University Press, 1997.

[15] J. Gao, H. Cai, On the structures and quantification of recurrence plots, Phys. Lett. A 270 (2000) 75–87.

[16] T.N. Davis, M. Sugiura, Auroral electrojet activity index AE and its universal time variations, J. Geophys. Res. 71 (1966) 785–801.

[17] B. Hnat, S.C. Chapman, G. Rowlands, N.W. Watkins, M.P. Freeman, Scaling in solar wind epsilon and the AE, AL and AU indices as seen by WIND, Geophys. Res. Lett. 10 (2002) 1029.

[18] M.C. Casdagli, Recurrence plots revisited, Physica D 108 (1997) 12–44.

[19] J.P. Zbilut, A. Giuliani, C.L. Webber Jr., Recurrence quantification analysis as an empirical test to distinguish relatively short deterministic versus random number series, Phys. Lett. A 267 (2000) 174–178.

[20] M. Haahr, http://www.random.org.

[21] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, 1949.

[22] D. Prichard, J. Theiler, Generalized redundancies for time series analysis, Physica D 84 (1995) 476–493.

[23] A. Renyi, Probability Theory, North-Holland, Amsterdam, 1970.

[24] J.P. Zbilut, A. Giuliani, C.L. Webber Jr., Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification, Phys. Lett. A 246 (1998) 122–128.

[25] M. Romano, M. Thiel, J. Kurths, W. von Bloh, Multivariate recurrence plots, Phys. Lett A 330 (2004) 214–223.

[26] M.S. Roulston, Estimating the errors on measured entropy and mutual information, Physica D 125 (1999) 285–294.

[27] M.G. Kivelson, C.T. Russell (Eds.), Introduction to Space Physics, Cambridge University Press, 1995 (Chapter 13).

[28] L.R. Lyons, Substorms: fundamental observational features, distinction from other disturbances, and external triggering, J. Geophys. Res. 101 (1996) 13011–13025.

[29] J. Greenhough, S.C. Chapman, S. Chaty, R.O. Dendy, G. Rowlands, Characterising anomalous transport in accretion disks from X-ray observations, Astron. Astrophys. 385 (2002) 693–700.

[30] J. Greenhough, S.C. Chapman, R.O. Dendy, V.M. Nakariakov, G. Rowlands, Statistical characterisation of full-disk EUV/XUV solar irradiance and correlation with solar activity, Astron. Astrophys. 409 (2003) L17–L20.

[31] J. Greenhough, S.C. Chapman, R.O. Dendy, D.J. Ward, Probability distribution functions for ELM bursts in a series of JET tokamak discharges, Plasma Phys. Controll. Fusion 45 (2003) 747–758.