Random versus Chaotic data: Identification using surrogate method

Pritha Das
Dept. of Mathematics Email:prithadas01@yahoo.com
Bengal Engg. and Science Univ., Shibpur, Howrah 711103, India

Atin Das Email:dasatin@yahoo.co.in
Head, NH School, Calcutta, India

Lewis L. Smith
Consultant Economist, US Email:mmbtupr@aol.com

Abstract: Distinguishing between chaotic and random data are quite difficult. Many chaotic processes generate outcomes, which are random while random processes may generate data that satisfy tests for chaos. Here, we show with example of both random and chaotic data sets that they can be characterized by applying surrogate data method and then comparing values of correlation dimension calculated from the original data set and its surrogate counter-part. Obviously, we comment on the data generated by the process- not about the process as a whole.
Key Words: random, chaotic, time series, surrogate, correlation dimension

---

1. Introduction

There are not only several definitions of randomness but in significant number of cases they can produce conflicting classifications of a given time series or a line segment thereof. One of the most popular one is as follows. A run is random if there is no theory for it, and no description of the data is more concise than the data itself. [1]
Any process is random when it is so designed and operated so that each outcome is generated independently of all others, past, present and future, regardless of what any give run looks like or how it tests. For example — the output of a fairly designed, well manufactured, well maintained and fairly operated roulette wheel. Details of randomness and various examples are given in earlier work [2].
On a more technical level concept, random data are not only the product of a random process but in addition, meet the following criteria by test, when the outcomes are arranged in order of generation —

1. Each outcome is independent of the other.
2. Each outcome represents one of all possible outcomes of the operation of a process.
3. The outcomes are mutually exclusive.
4. Collectively the outcomes are exhaustive.
5. Each outcome has a probability of less than one.
6. The probabilities of all the outcomes add to one.

The concept of randomness used to classify a specific run should be based on the run's characteristics, not on the characteristics of the generating process. An important lesson is that a simple chaotic system can produce a time series that passes most tests for randomness. Conversely, a pure random system with a nonuniform power spectrum (correlated noise) can masquerade for chaos. [3] As Mandelbrot (2004) noted "…a

financial market is especially prone to …statistical mirages. My mathematical models can generate charts that — purely by the operation of random processes — appear to trend and cycle". [4]

In this paper we attempt to test whether a given set of data in the form of a time series - is chaotic and random. As noted earlier that it is quite difficult to predict as a whole about the process generating that data as chaotic or random one. So we here confine the prediction about the time series only.
Our proposed scheme is as follows: applying the method of surrogate data. While shuffling data of the given series- it preserves the probability distribution function, it does not preserve the power spectrum and correlation function. The surrogate data will be white and uncorrelated even when the original time series is not. [3] Now suppose one developed some statistic that purports to distinguish chaos from noise. In our case, it will be correlation dimension (D2). Now one compute the statistic for the original and the surrogate data, and the values are inevitably different. One needs to decide if the difference is statistically significant.
In section2, we give the description of data and in section3 we briefly outline the surrogate method and correlation dimension concepts. In section4, we give the results in tabular from and draw some conclusions. The reference section is given at the end of the paper.

2. Data

Here we have tested a number of time series. Some of them (A, B & C) are generated with random numbers produced by a computer program. Some other (D & E) series are chaotic, they are formed by recording of human EEG data in various conditions (for example, EEG1 has its subject's eyes open, while in EEG2- eyes were closed. Another data series (that is, F) was formed by numerical simulation of low dimensional Artificial Neural Network models. All these (D, E & F) three chaotic time series were under study in some different context but the dynamics of EEGs & Theoretical data were confirmed to be chaotic [5]. Another data set, G is well known chaotic Lorentz system.

Analyzing a time series with a nonlinear approach is definitely a complicated problem. Simple answers have been repeatedly offered in the literature, but researchers like Kantz et al. [6] are against such simple answers. We have calculated the correlation dimension (D2) of both the original and surrogate of each dataset and then compared the values to test the existence as well as the nature of chaos.

3. Mathematical tools

As underlined in [3] to find a statistic which is to be compared between the original and the surrogate data sets, we pick up correlation dimension as that statistic. Before analysis, we give below a simple outline of the D2 as well as the surrogate process.

3.1 Correlation Dimension (D2)

Being one of the characteristic invariant of nonlinear system dynamics, the correlation dimension gives a measure of complexity for the underlying attractor of

the system [7]. To detect the saturation value of the correlation dimension, the function plots the computed correlation dimension as a function of the embedding dimension. Mathematically, the correlation dimension is a special case of the generalized dimension, and it is given by

$$D2 = \lim_{r \to 0} \left[\sum_{i=1}^{M(r)} P_i^{\,2} / \log r \right] \quad (1)$$

with P_i being the probability to find a point of the attractor within the i^{th} subcube of phase space when phase space is subdivided into disjunctive cubes of side length r. The number $M(r)$ of cubes that contain attractor points, is related to the dimension D of the attractor :

$$M(r) \sim (1/r)^D \quad (2)$$

Parameters involved are

i) delay τ ii) minimum (m_{min}) and maximum (m_{max}) embedding dimension (m)

iii) Lower relative radius r_0 and upper relative radius r_1, between which the correlation dimension is calculated as the derivative of the log C(r)/log(r) plot. Thus r_0 and r_1 should both lie within the scaling region of the attractor. Here we draw for each data set figures by plotting D2 versus m in the range of m_{min} and m_{max} and by choosing appropriate value of m for that data set as given in table2, estimate the value of D2, as shown for a data set in Fig. 1. Details of such process have been discussed in our earlier work [5].

3. 2 Surrogate data method

We follow the approach of Theiler et al. and Schreiber et al. [8, 9]. Surrogate signal is produced by phase randomizing the given data. It has similar spectral properties as of the given data, that is, the surrogate data sequence has the same mean, the same variance, the same autocorrelation function and therefore the same power spectrum as the original sequence, but (nonlinear) phase relations are destroyed. In the case of data shuffling the histograms of the surrogate sequence and the reference sequence are identical, too (Refer to Fig. 2).

We have used the first way to produce the surrogate data set of all the data sets under consideration. Next we have calculated fractal dimension of them and compared with respective original data sets as given in Table 1. For a chaotic dataset, there will be considerable difference in values as the "surrogating" process destroys the nonlinearity. For a random data set, there will be little effect of surrogate process and thus there will be little difference in D2 before and after surrogating it.

We took help of following software, apart from our own programs written in C,

i) Dataplore 2.0-6 (c) 1995-2000: from DATAN GmbH, Germany [10]

ii) TISEAN 2.1 (2000)-Nonlinear Time Series Analysis: by R. Hegger, H. Kantz, and T. Schreiber. [6]

4. Results

Table 1: Details of results with data sets

Data	D2	D2(Surrogate)	% change	Data Description
Chaotic Series				
A Rand4000	4.2	4.7	11.94	Random series of 4000 points
B Rnd2sgn	4.9	4.7	4.08	Signed random number
C Rnd4k2sr	4.6	4.4	4.35	Subtraction of 2 random series
Random Series				
D. EEG1	2.98	1.26	57.72	Elecroencelophalogram data
E. EEG2	2.08	2.74	31.73	Do
F. Theoretical	1.48	1.99	68.64	Mathematical simulation
G. Lorenz	1.95	3.94	102.05	Well known Lorenz series

Table 2: Showing values of parameters taken as discussed in Sec. 3

m_{max}, m_{min}	τ	Reference pts	r_0	r_1	Skip
20, 5	1	100	.2	1	20
Values of m	for data sets A,B,C m=5		for D,E,G m=8		For G m=6

We find from the above table that for known series consisting random data has change of less than 12% in their D2 value due to surrogate process. But the same difference is much higher for a truly chaotic data (more than 30%). Obviously, there are differences in dimension of chaos. But in any case, even for low dimensional chaotic series [4] E or a high dimensional G series, the change is much higher.

So effectively, we can conclude that for a set of data (signed or unsigned, integer or fractional etc.), we may apply the surrogate process and D2 calculation to conclude whether they are random data or chaotic data.

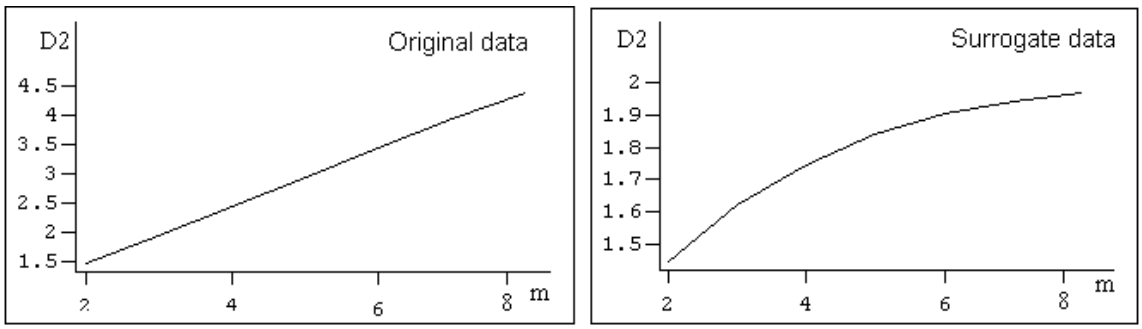

Fig. 1: Estimation of value of D2 from figure- for Lorenz data (m = 6).

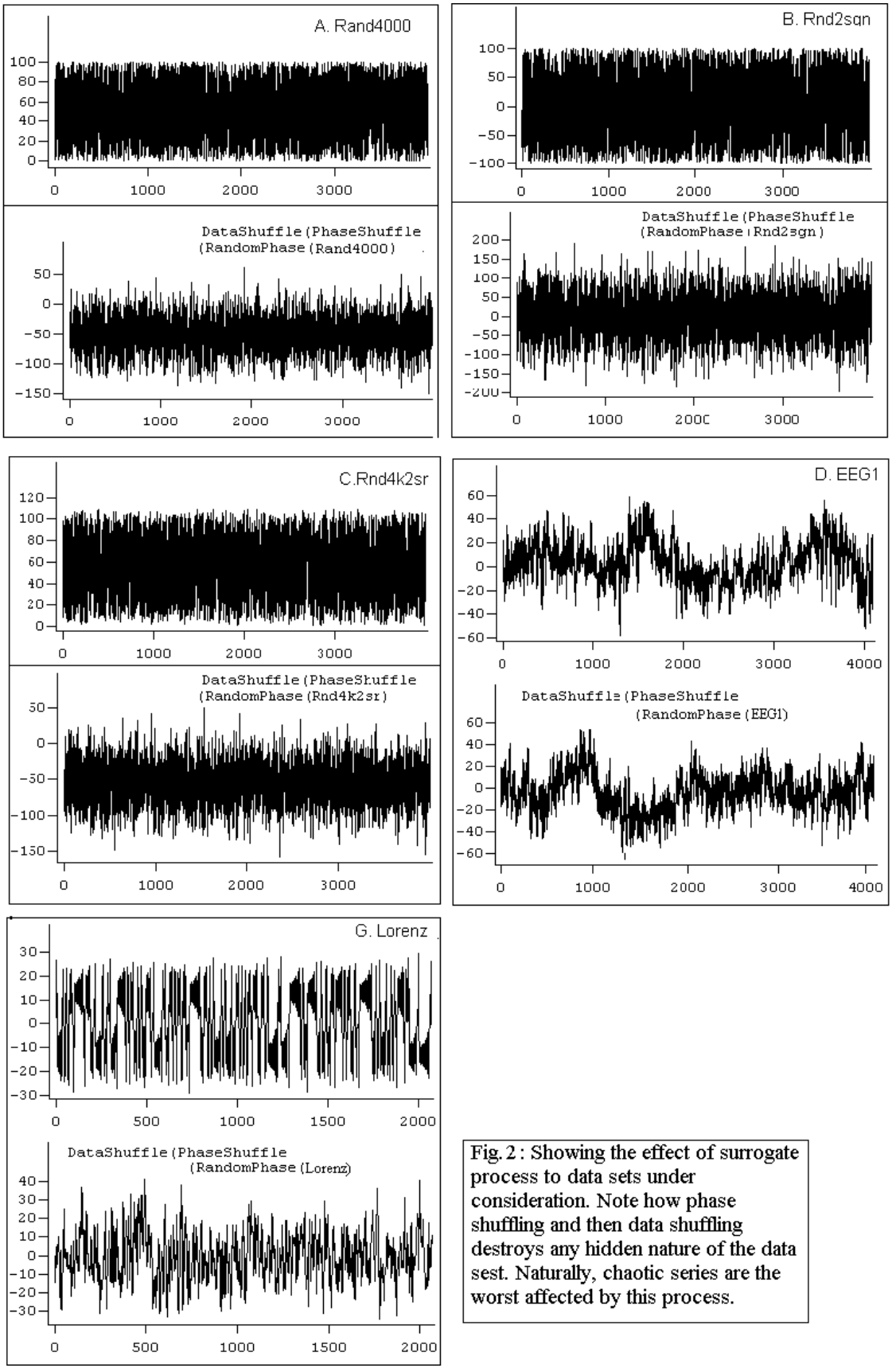

Fig. 2 : Showing the effect of surrogate process to data sets under consideration. Note how phase shuffling and then data shuffling destroys any hidden nature of the data sest. Naturally, chaotic series are the worst affected by this process.

References:

1. Gregory J. Chaitin, (2002) Conversations with a mathematician, Spriniger-Verlag, London, p. 136.

2. Smith, L. L. (2002) Economics and markets as complex system, Business Economics, Jan. 2002, p. 46-52.

3. Sprrot, J. C., (2003), Chaos and time series analysis. Oxford University Press, New York, 232-6.

4. B. Mandelbrot (2004) The (Mis)behavior of Markets, Basic Books, NYC, p. 22

5. A. Das, P. Das and A. B. Roy (2001) Nonlinear Data Analysis: A Comparison Between Experimental [EEG] Data and Theoretical [ANN] Data, Complexity, Vol. 7, No. 3 ,30-40

6. Hegger, R.; Kantz, H.; Schreiber, T. (1999) Practical implementation of nonlinear time series methods: The TISEAN package. CHAOS 1999, 9, 413–435.

7. Grassberger P. and Procaccia I. (1983) Measuring the strangeness of strange attractos. Physica D 9, pp.189-208.

8. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., & Farmer, J. D., (1992). Testing for nonlinearity in time series: the method op surrogate data. Physica D 58, 77-94.

9. Schreiber, T. & Schmitz, A. (1996) Improved surrogate data for nonlinearity test. Physical Review Letters 77, 635-8.

10. Dataplore 2.0-6, 1995-2000: from DATAN GmbH, Germany. Downloadable at http://www.datan.de