2005 International Symposium on Nonlinear
Theory and its Applications (NOLTA2005)
Bruges, Belgium, October 18-21, 2005

Small-Shuffle Surrogate Method and the Application

Tomomichi Nakamura[a], Michael Small[a], Yoshito Hirata[b] and Kazuyuki Aihara[b]

[a] Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Email: entomo@eie.polyu.edu.hk, ensmall@polyu.edu.hk
[b] Institute of Industrial Science, The University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
Email: yoshito@sat.t.u-tokyo.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—We describe a new surrogate method for investigating whether there is some kind of dynamics in irregular fluctuations (short term variability), even if the data have long term trends or periodicities. This is, in other words, an investigation of whether irregular fluctuations are independently distributed random variables. These cases are theoretically incompatible with the assumption required to apply previously proposed surrogate methods. We apply the method to a variety of known test systems and actual time series with unknown dynamics.

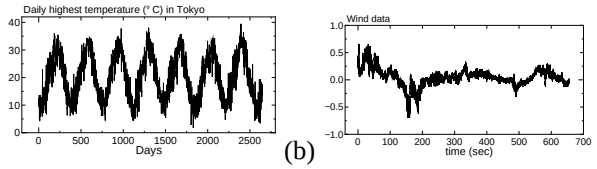
1. Introduction

The surrogate analyses have long been focused on irregular fluctuations (short term variability) and pseudo-periodic time series [2, 4, 7, 5, 6]. However, they are not only the behaviours we can see in the real world. Some data have irregular fluctuations and some trends (periodicities). See Fig. 1 (a) and (b), which we will examine later. These are more complicated than time series intended by these method, and the methods are theoretically inconclusive. In this paper, to investigate whether there is some kind of dynamics in irregular fluctuations irrespective of whether the data are with or without trends, we introduce a method, the small-shuffle surrogate (SSS) method [1] and apply the method to two actual data,

2. The Small-Shuffle Surrogate Method

The basic premise of the proposed technique is that if irregular fluctuations are not random, then there is some kind of underlying dynamical system: whatever trending is contaminating the data. In such a case, the data index (order) itself has important implications irrespective of whether time series are linear or nonlinear. Hence, whenever the index changes, the flow of information also changes and the resultant time series no longer reflects the original dynamics. We focus our attention on this point and propose a new surrogate method using this idea. The purpose of our method is to distinguish between irregular fluctuations with or without dynamics.

To investigate irregular fluctuations (especially when they are with long term trends), we want to destroy local structures or correlations in irregular fluctuations (short term variability) and preserve the global behaviours (trends). To generate such surrogate data, we shuffle the data index on a "small" scale: this is in contrast to the random-shuffle surorgate (RSS) method where the data index is shuffled on a "large" scale and any structure of the original data is destroyed [2]. We generate surrogate data as follows; Let the original data be $x(t)$, let $i(t)$ be the index of $x(t)$ (that is, $i(t) = t$, and so $x(i(t)) = x(t)$), let $g(t)$ be Gaussian random number and $s(t)$ will be the surrogate data.



(a) (b)

Figure 1: Two series examined in this paper. Both the data have irregular fluctuations and seemingly trends. (a) daily highest temperature in Tokyo from 1 January 1998 to 31 March 2005. The time series should have at least one long term periodicity, one year (365 days) periodicity. (b) wind data. The data was measured using an anemometer with 50 Hz located at 1 m above the ground in Institute of Industrial Science, The University of Tokyo in Komaba, Tokyo, from 13:08 JST on 25 October 2004 (Mon) for about 1 hour. We use the part of wind data, which corresponds to about 11 minutes.

(i) Obtain $i'(t) = i(t) + Ag(t)$, where A is an amplitude (adding Gaussian random numbers to the index of the original data). Note that the index $i(t)$ will be a sequence of integers whereas the perturbed sequences $i'(t)$ will not.

(ii) Sort $i'(t)$ by the rank-order [1] and let the index of $i'(t)$ be $\hat{i}(t)$ (rank-order the perturbed index, thereby generating a slightly perturbed index of the original data)

[1] By rank-order we mean the sequence in which the values of different relative magnitude occur. For example, the rank-order of the sequence $\{\pi, 0, e, \sqrt{2}\}$ is $\{4, 1, 3, 2\}$.

(iii) Obtain the surrogate data $s(t) = x(\hat{i}(t))$ (reorder the original data with the perturbed index [2])

When the amplitude A is selected appropriately, the index is shuffled only on a small scale, where the generated surrogate data loses local structures or correlations but preserve the global behaviours as much as possible. We call the method the *Small-Shuffle Surrogate* (SSS) method. The SSS data have the same probability distribution as the original data. Hence, the null hypothesis addressed by our algorithm is that irregular fluctuations (short term variability) are independently distributed (ID) random variables (in other words, there is no short term dynamics or determinism). The major difference between the RSS and SSS methods is that the SSS method removes the requirement for "identically" distributed random variates.

2.1. The discriminating statistics

For surrogate test, discriminating statistics are necessary. The SSS method changes the flow of information in the data. Hence, we choose to use the auto-correlation function (AC) and the average mutual information (AMI) as discriminating statistics. The AC; an estimate of the linear correlation in data; and AMI; a general nonlinear version of AC on a time series; can answer the question: on average how much does one learn about the future from the past. To know difference between the original and SSS data, we show the full curves for the purposes of illustration, because we want to inspect the global behaviours[3].

After calculation of these statistics, we need to inspect whether a null hypothesis shall be rejected or not. We employ the Monte Carlo hypothesis testing and check whether estimated statistics of the original data fall within or outside the statistics distribution of the surrogate data [3]. We generate 39 SSS data and then the (two tailed) significance level is 0.05.

2.2. Searching for the most appropriate value of A

The SS surrogate data are influenced primarily by the amplitude A. If A is too small, the values are better at preserving any structure and correlation in the original data, however, they are not effective at destroying the local structures. As the result, difference between estimated statistics for the original and surrogate data is not distinct even when there is a dynamics in irregular fluctuations. On the other hand, if A is too large, the values are better at destroying any structure and correlation of the original data, however, they are not effective at preserving the long term behaviours. For data with trends, large values are not appropriate, because the global behaviours of the original data

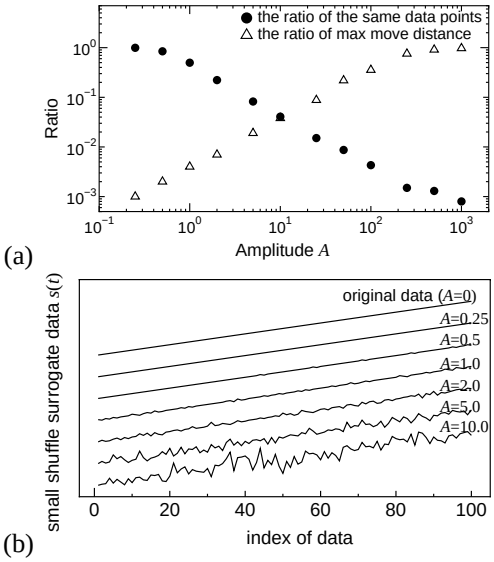


Figure 2: Relationship of the amplitude A of Gaussian random numbers and the index. The panel (a) illustrates (as a function of shift amplitude A): the proportion of points that are unperturbed by the SSS algorithm (•); and, the maximum distance that any point in the original data is perturbed in the surrogate (△, expressed as a fraction of the data length). The panel (b) illustrates the effect of different values of A. The original data is generated by $x(t) = t$, $1 \leq t \leq 100$. If the SSS and original data are identical, then the curve should be a straight line. If the SSS data is equivalent to an ordinary RSS data set, then the curve should be IID.

are lost and the influence of contaminated trends may be larger than that of irregular fluctuations. As the result, some differences appear even when there is no dynamics in irregular fluctuations. Hence, the smaller the value of A the better, if the value can destroy local structures and preserve the long term behaviours.

Figure 2 shows the relationship of the amplitude A and the data index. Panel (a) shows that as A increases, the number of data points which do not move decreases and the ratio of maximum move distance increases. To show the influence of the amplitude visually, we directly compare the original data and the SSS data at different amplitude A. Panel (b) shows that until A is about 2.0, the behaviour of $s(t)$ is almost the same as the original data ($A = 0$), as the A increases, the behaviour of $s(t)$ becomes more stochastic. This result indicates that broadly speaking, we should use up to $A = 2.0$, if we want to generate SSS data which loses the local structures or correlations of the original data and preserve the global structures or behaviours.

In preliminary tests [4], we find that $A = 1.0$ is most ap-

[2]The simple example is as follows; Let $x(t)$ be $(13, 12, 14, 11, 15)$, where $i(t)$ is $(1, 2, 3, 4, 5)$. We obtain the perturbed index $i'(t)$, where let $i'(t)$ be $(0.1, -1.3, 3.2, 4.5, 2.7)$. Sorted $i'(t)$ becomes $(-1.3, 0.1, 2.7, 3.2, 4.5)$ and hence $\hat{i}(t)$ is $(2, 1, 5, 3, 4)$. Then, $s(t)$ which is $x(\hat{i}(t)$ (that is, $x(2), x(1), x(5), x(3), x(4)$) is $(12, 13, 15, 14, 11)$.

[3]It is possible to consider only one value of AMI or AC at an arbitrary time lag for all tests, for example AC or AMI at time lag=1.

[4]We have investigated which values of A are the most appropriate using AC and AMI and some models which are described later, where $A = 0.25$, 0.5, 1.0, 2.0, 5.0 and 10.0. We find that $A = 1.0$ is sufficient. When $A < 1.0$, AC and AMI cannot give clear difference between the

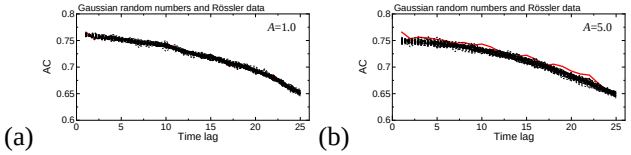


Figure 3: A plot of the AC for Gaussian random number with x component of the Rössler equations: (a) $A = 1.0$ and (b) $A = 5.0$, where the number of SSS data is 39. The solid line is the original data and dotted lines are the SSS data.

propriate and more than adequate for nearly all purposes, in this case about 50% of the data points in the SSS data is in the same index as the original data. Figure 3 shows the typical results of these calculations. Although AC of the original data fall within the distributions of SSS data in panel (a), the base line of the SSS data strays far from that of the original data in panel (b). Hence, we use $A = 1.0$ in our calculations. We note that although we expect this value is appropriate in most cases, the value of A will depend on features of the data, and smaller or larger values may be justified in some situations.

3. Numerical Examples

We now demonstrate the application of our algorithm, and confirm our theoretical arguments with several cases. In each case the number of data points used is 5,000, the data used are both noise free and contaminated by 10% Gaussian observational noise. The first application is to data with no trend. We use Gaussian random numbers as data with no dynamics. To study data with dynamics, we use the following models. The linear AR model is given by $x_t = a_1 x_{t-1} + a_6 x_{t-6} + \eta_t$, where we use $a_1 = 0.3$, $a_6 = 0.2$ and η_t is Gaussian dynamical noise with standard deviation 1.0. The Ikeda map is given by

$$f(x,y) = \left(\ 1 + \mu\,(x\cos\theta - y\sin\theta)\,, \mu\,(x\sin\theta + y\cos\theta) \ \right),$$

where $\theta = a - b/\left(1 + x^2 + y^2\right)$ with $\mu = 0.83$, $a = 0.4$ and $b = 6.0$. The Logistic map is given by $x_t = a x_{t-1}(1.0 - x_{t-1})$, where $a = 4.0$. In all cases, we use x_t as the observational data.

Figures 4 (a) and (b) show that when there is no dynamics (that is, data are Gaussian random numbers), both AC and AMI of the original data fall within the distributions of the SSS data. However, in other cases, that is, when there is dynamics, even if systems and data are contaminated stochastically, AC or AMI or both are distinct. Here, we note that some differences clearly appear when the time lag is relative small, because the information in systems is not retained for longer periods of time. When the data is

original data and the SSS data even if there is dynamics. When $A \geq 1.0$, AC and AMI can give clear difference. However, when the irregular fluctuations are with trends, $A \geq 5.0$ are too large.

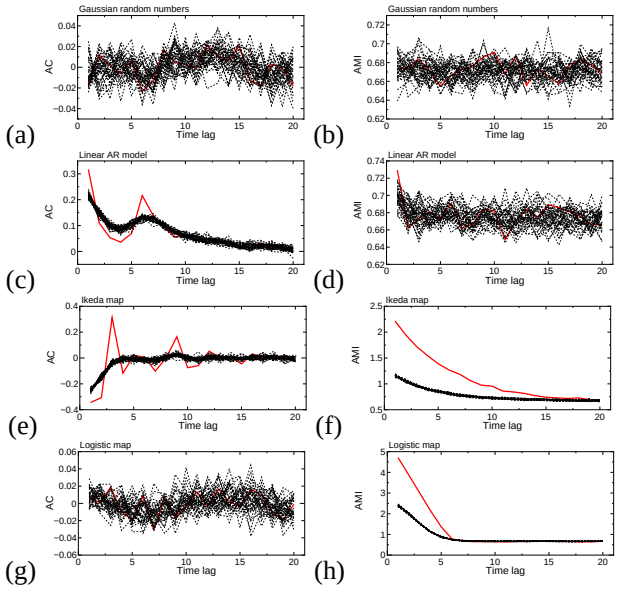


Figure 4: A plot of the AC and AMI: (a) and (b) Gaussian random number, (c) and (d) a linear AR model, (e) and (f) the Ikeda map, and (g) and (h) the Logistic map where we use $A = 1.0$ and 39 SSS data. The solid line is the original data and dotted lines are the SSS data.

contaminated by 10% observational noise, and also when the amplitude A is larger than 1.0, the results obtained are essentially the same.

The second application is to data with trends, where Rössler equations are used to generate a slow trend. The equations are given by

$$\dot{x} = -(y+z), \dot{y} = x + ay, \dot{z} = b + z(x-c),$$

where $a = 0.3909$, $b = 2.0$, $c = 4.0$, when calculated using the fourth order Runge-Kutta method with sampling interval 0.02. The equations when using these parameters exhibits period 6 behaviour [5]. Data generated using the same models as above are added to the x component of the equations, where both the systems are independent and the level of additional data to the data is equivalent to 56.2% (5dB) observational noise at each case. The behaviour is similar to that in Fig. 1 (a).

Figure 5 shows the results for these data. Panels (a) and (b) again show that when there is no dynamics in the irregular fluctuations, AC and AMI of the original data fall within the distributions of SSS data, however, AC or AMI or both are distinct when there is dynamics. In all cases, especially when the time lag is larger, behaviours of AC and AMI of the SSS data are very similar to that of the original data. This indicates that the local structures are destroyed and the global structures are preserved in the SSS data. When the data is contaminated by 10% observational noise, the results are essentially the same.

Figures 4 and 5 show that when irregular fluctuations are Gaussian random numbers (that is, there is no dynamics),

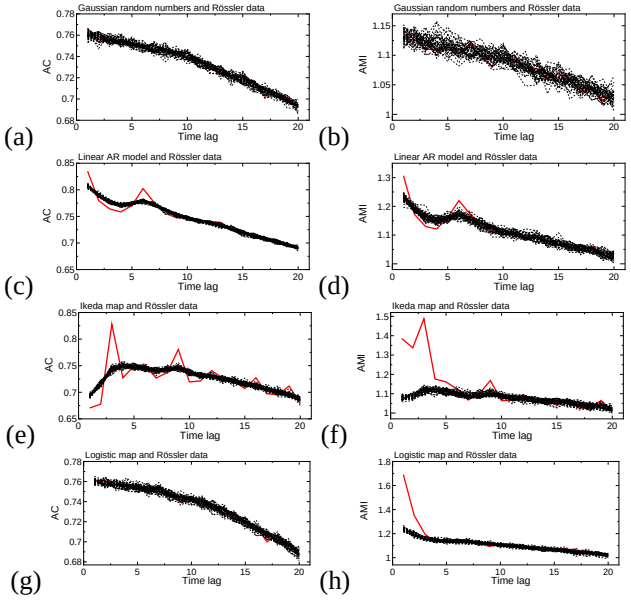


Figure 5: A plot of the AC and AMI: (a) and (b) Gaussian random number, (c) and (d) linear AR model, (e) and (f) Ikeda map, and (g) and (h) the Logistic map, with x component of the Rössler equations, where we use $A = 1.0$ and 39 SSS data. The solid line is the original data and dotted lines are the SSS data.

both AC and AMI of the original data fall within the distributions of the SSS data, but when there is dynamics, the AC or AMI or both fall outside, even if systems and data are contaminated stochastically. Therefore, applying the SSS method can detect whether there is dynamics or not using AC and AMI.

4. Application

Based on the result of these computational studies, we apply the proposed method to three experimental systems: (1) daily highest temperature in Tokyo, and (2) wind data, which seem to have trends. See Figs. 1 (a) and (b). We use 2,647 data points for the daily highest temperature in Tokyo, and 32,768 data points (about 11 minutes) for the wind data.

Figure 6 shows the SSS data and the results. Panels (a) and (c) show that the SSS data are similar to the original data. Panel (b) shows that AMI of the daily highest temperature in Tokyo fall outside the distributions of the SSS data. Hence, we consider that the temperature data have some kind of dynamics behind the data. When we use the daily lowest temperature in Tokyo, the result obtained is essentially the same. Panel (d) shows that AMI of the wind data fall outside the distributions of the SSS data. Hence, we consider that the wind data have dynamics. Although we do not show the results of AC in Fig. 6, the results are essentially the same as those of AMI in all case.

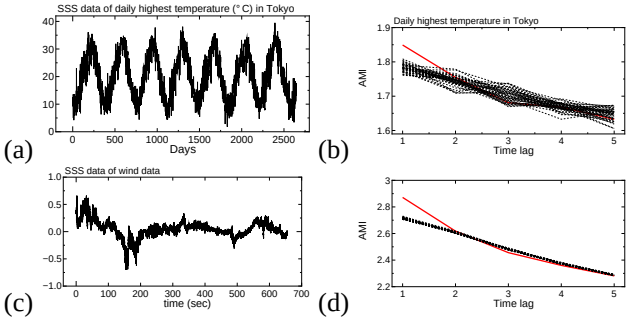


Figure 6: SSS data and a plot of the AMI: (a) and (b) daily highest temperature in Tokyo, and (c) and (d) wind data where we use $A = 1.0$ and 39 SSS data. The solid line is the original data and dotted lines are the SSS data in panels (b) and (d).

Acknowledgments

This research is supported by a Hong Kong University Grants Council Competitive Earmarked Research Grant (CERG) number PolyU 5216/04E. Also, this study was partly supported by the Industrial Technology Research Grant Program in 2003, from the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

References

[1] T. Nakamura and M. Small, "Small-Shuffle Surrogate Data: Testing for Dynamics in Fluctuating Data with Trends", submitted, 2005.

[2] J. Theiler, S. LuDanK, A. Longtin, B. Galdrikian and J.D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D* 58 (1992) 77.

[3] J. Theiler and D. Prichard, "Constrained-realization Monte-Carlo method for hypothesis testing", *Physica D* 94 (1996) 221.

[4] T. Schreiber and A. Schmitz, "Improved Surrogate Data for Nonlinearity Tests", *Phys. Rev. Lett*, 77 (1996) 635.

[5] M. Small, D. Yu and R.G. Harrison, "Surrogate Test for Pseudoperiodic Time Series Data", *Phys. Rev. Lett*, 87 (2001) 188101.

[6] X. Luo, T. Nakamura and M. Small, "Surrogate test to distinguish between chaotic and pseudoperiodic time series", *Phys. Rev.* **E** 71 (2005) 026230.

[7] J. Theiler, "On the evidence for low-dimensional chaos in an epileptic electroencephalogram", *Phys. Lett.* **A** 196 (1995) 335.