Small-shuffle surrogate data: Testing for dynamics in fluctuating data with trends

Tomomichi Nakamura[*] and Michael Small
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

(Received 18 April 2005; published 23 November 2005)

We describe a method for identifying dynamics in irregular time series (short term variability). The method we propose focuses attention on the flow of information in the data. We can apply the method even for irregular fluctuations which exhibit long term trends (periodicities): situations in which previously proposed surrogate methods would give erroneous results. The null hypothesis addressed by our algorithm is that irregular fluctuations are independently distributed random variables (in other words, there is no short term dynamics). The method is demonstrated for numerical data generated by known systems, and applied to several actual time series.

DOI: 10.1103/PhysRevE.72.056216 PACS number(s): 05.45.Tp, 02.50.−r, 05.10.−a

There are many natural phenomena that show irregular fluctuations (short term variability). The question of whether the fluctuations are random or not is an old one and extremely important. If the fluctuations are not random, then they are due to some kind of dynamical structure and then it might be possible to build deterministic models or model systems from the time series. Clearly, such models are of immense value for both understanding and predicting the time series. To investigate whether the data can be fully described by independent and identically distributed (IID) random variables, the random-shuffle surrogate (RSS) method has been proposed [1]. Although this method is effective for time series with no trends (periodicities) like that shown in Figs. 1(a) and 1(b), the algorithm is ineffective for data exhibing slow trends or periodicities [see Figs. 1(c) and 1(d)]. Such cases are theoretically incompatible with the assumption of the RSS method as well as other linear surrogate tests [1,2]. There is currently no method which can tackle this problem. In this Communication, to investigate whether there is dynamics in data which also exhibits irregular fluctuations, we introduce such a method.

The basic premise of this 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 surrogate method using this idea. The purpose of our method is to distinguish between irregular fluctuations with or without dynamics.

After describing our technique, we will present our choice of discriminating statistic. Then, we will apply this algorithm to two cases using simulated time series data. One case is that data have no trend (this case can also be adequately addressed with the standard surrogate methods). The other case is that data have trends (this case is not consistent with existing surrogate techniques). In each case, the data we use are both noise free and subsequently contaminated by 10% Gaussian observational noise. Also, we apply the method to three actual time series: cobalt data, nuclear magnetic resonance (NMR) laser data, and daily sunspot numbers.

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 behaviors (trends). To generate such surrogate data, we shuffle the data index on a "small" scale: this is in contrast to the RSS method where the data index is shuffled on a "large" scale and any structure of the original data is destroyed. 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 the Gaussian random number and $s(t)$ will be the surrogate data.

(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 [9] 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).

(iii) Obtain the surrogate data $s(t)=x(\hat{i}(t))$ (reorder the original data with the perturbed index [10]).

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 behaviors 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.

The SSS method changes the flow of information in data. Hence, we choose to use the autocorrelation function (AC) and the average mutual information (AMI) as discriminating

[*] Electronic address: entomo@eie.polyu.edu.hk

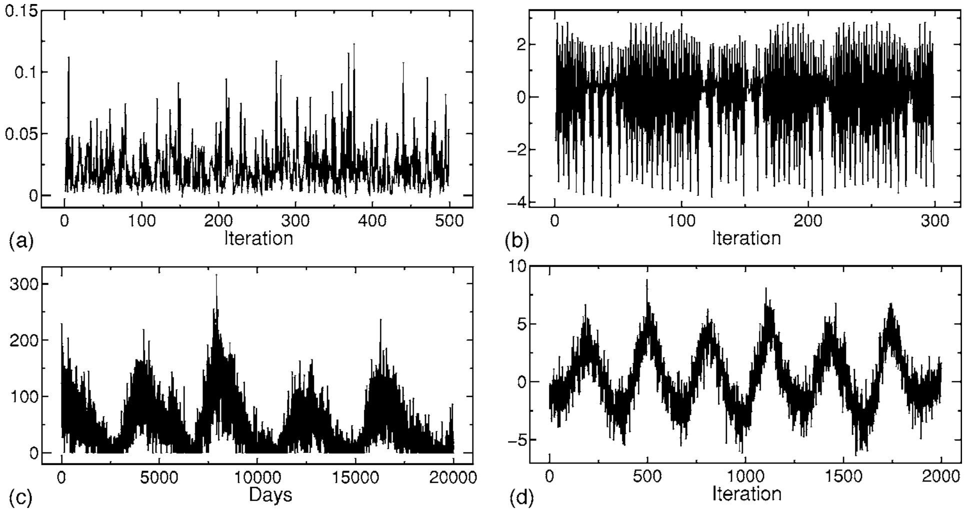

FIG. 1. Segments of four time series examined in this paper: (a) cobalt data, (b) NMR laser data, (c) daily sunspot numbers, and (d) Gaussian random numbers with x component of the Rössler equations. For (a) and (b) the important question is whether there is any intersample dynamics. For (c) and (d) the important thing is whether there is any intersample dynamics apart from the longer "periodic" fluctuations. The ordinate axis is arbitrary.

statistics. The AC, an estimate of the linear correlation in data, and the AMI, a general nonlinear version of the AC on a time series, can answer the question: on average how much does one learn about the future from the past [3]. After calculation of these statistics, we need to inspect whether a null hypothesis shall be rejected or not. We employ Monte Carlo hypothesis testing and check whether estimated statistics of the original data fall within or outside the statistics distribution of the surrogate data [4]. When the statistics fall within the distribution of the surrogate data, we consider that the original and the surrogate data may come from the same population and then the surrogate null hypothesis may not be rejected. We generate 39 SSS data and then the (two tailed) significance level is 0.05.

Clearly, the SSS data are influenced primarily by the amplitude A. If A is too small, the data are shuffled very little or not at all, and then the SSS data are almost identical to the original data. Conversely, if A is too large, data are shuffled on a large scale, and the SSS data are almost random like the RSS data. Hence, the smaller the value of A the better, if the value can destroy local structures and preserve the long term behaviors.

Figure 2 shows the relationship of the amplitude A and the data index. Figure 2(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. Figure 2(b) shows that until A is about 2.0, the behavior of $s(t)$ is almost the same as the original data ($A=0$), as the A increases, the behavior 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 behaviors. For data with no trend, larger values of A are available. However, if data have some trends, large values are not appropriate, because the global behaviors are lost and the influence of contaminated trends becomes larger than that of irregular fluctuations. In our calculations [11], we find that $A=1.0$ is most appropriate and more than adequate for nearly all purposes, in this case about 50% of the data points in the SSS data are in the same index as the original data. Figure 3 shows the typical results of these calculations. Figure 3(a) shows that AC of the original data fall within the distributions of the SSS data. According to the criterion mentioned previously, we cannot reject the hypothesis. However, Fig. 3(b) shows that the base line of the SSS data strays far from that of the original data. That is, AC of the original data fall outside the distributions of the SSS data, and then we reject the hypothesis. Therefore, we find that smaller values of A are more appropriate and moreover

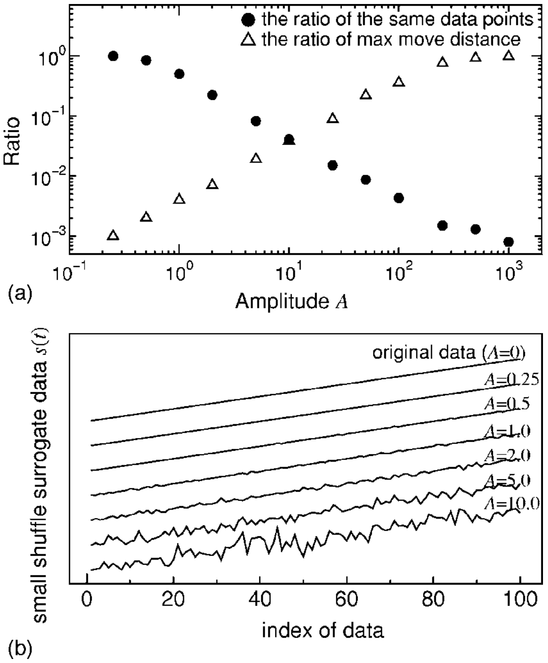

FIG. 2. The relationship of the amplitude A of Gaussian random numbers and the index. (a) illustrates (as a function of shift amplitude A) the proportion of points that are unperturbed by the SSS algorithm (●) and the maximum distance that any point in the original data are perturbed in the surrogate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

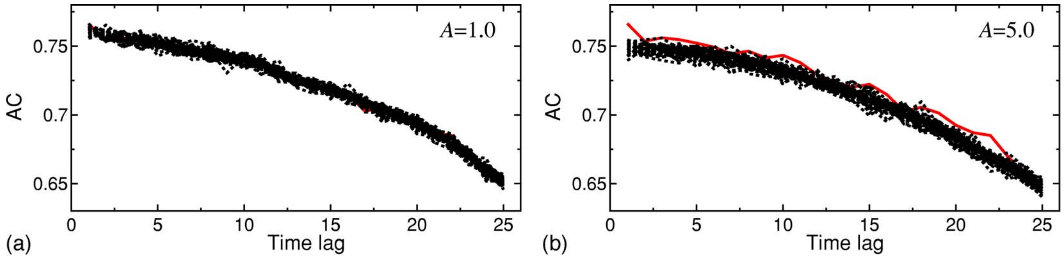

FIG. 3. (Color online) A plot of the AC for Gaussian random number with x component of the Rössler equations: (a) $A=1.0$ and (b) $A=5.0$, where the number of SSS data is 39. The solid line is the original data and dotted lines are the SSS data.

$A=1.0$ is large enough. We note that although we expect this value is appropriate in most cases, the value of A will depend on features of data, and smaller or larger values may be justified in some situations.

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 5000; 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.

(1) The linear autoregressive (AR) model given by $x_t = a_1 x_{t-1} + a_6 x_{t-6} + \eta_t$ (Ref. [5]), where we use $a_1=0.3$, $a_6=0.2$, and η_t is Gaussian dynamical noise with standard deviation (SD) 1.0.

(2) The Ikeda map given by

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

where $\theta = a - b/(1+x^2+y^2)$ with $\mu=0.83$, $a=0.4$, and $b=6.0$ (Ref. [6]). Also, to investigate when systems are contaminated stochastically, we add Gaussian dynamical noise with SD 0.04 to both the x and y components.

(3) The logistic map given by $x_t = a x_{t-1}(1.0 - x_{t-1})$ where $a=4.0$ (Ref. [7]).

In all cases, we use x_t as the observational data. We plot results for the cases of Gaussian random numbers, the Ikeda map, and the Ikeda map perturbed by dynamical noise with 10% observational noise.

Figures 4(a) and 4(b) show that when there is no dynamics (that is, data are Gaussian random numbers), the AC and

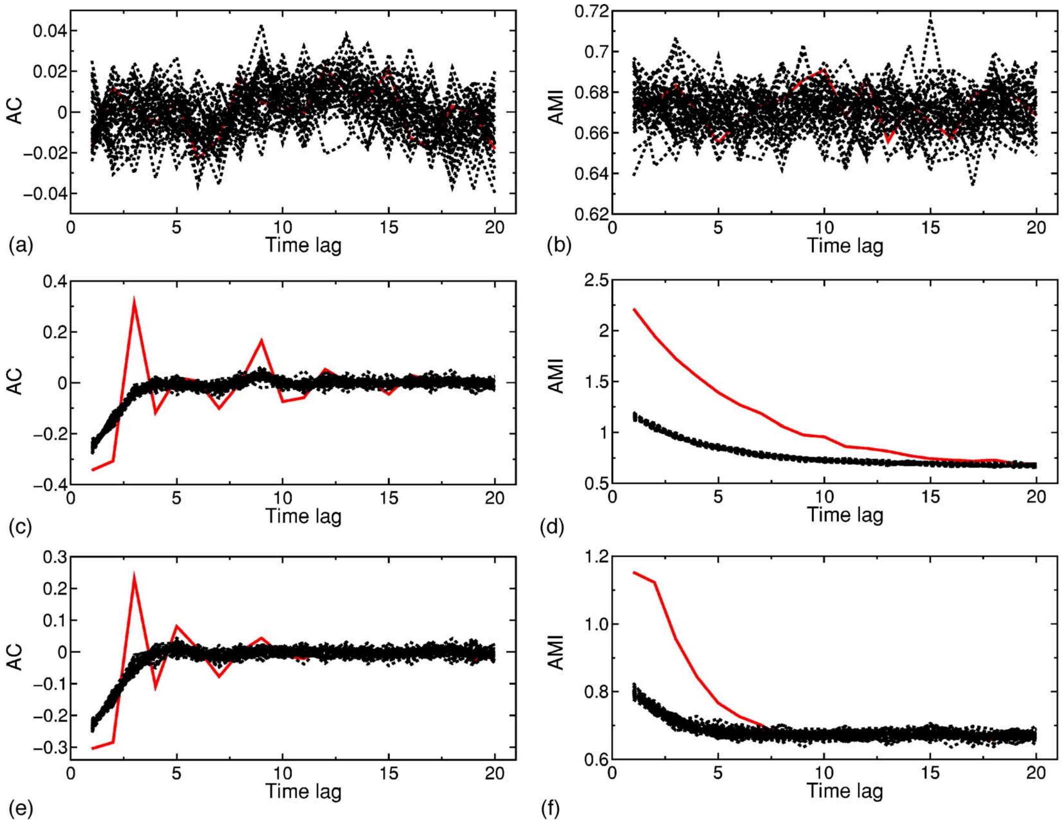

FIG. 4. (Color online) A plot of the AC and the AMI: (a) and (b) Gaussian random number, (c) and (d) the Ikeda map, and (e) and (f) the Ikeda map with dynamical and 10% observational noise, where we use $A=1.0$ and 39 SSS data. The solid line is the original data and dotted lines are the SSS data.

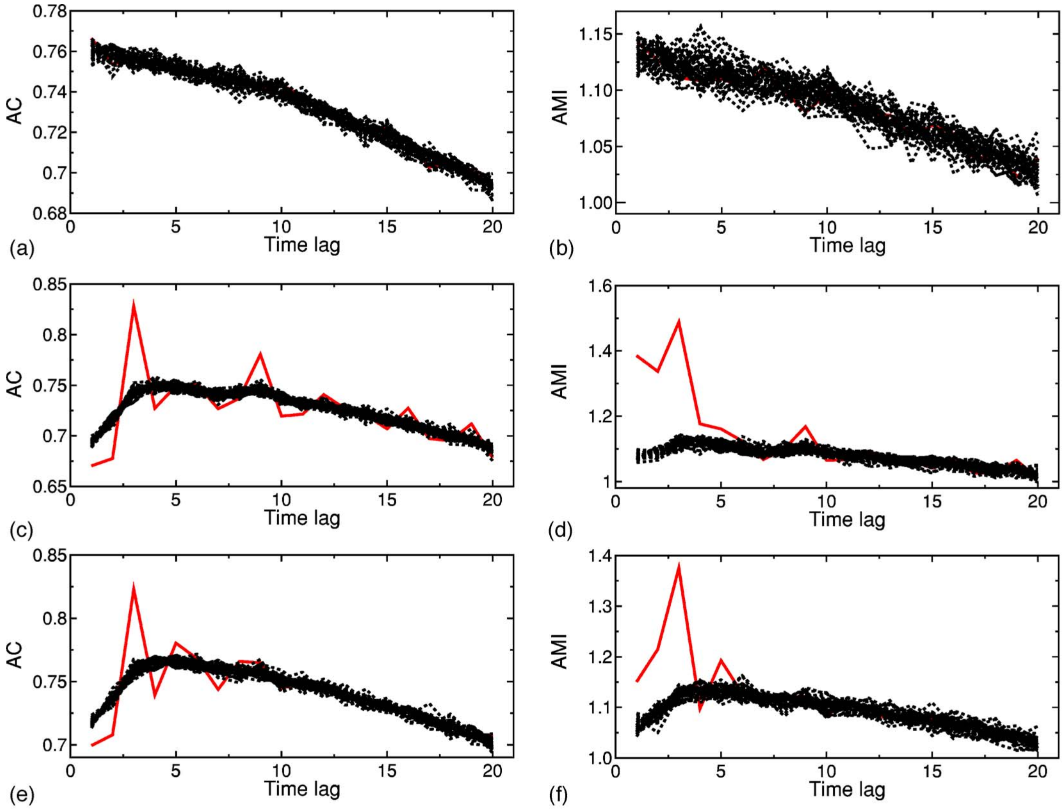

FIG. 5. (Color online) A plot of the AC and AMI: (a) and (b) Gaussian random number, (c) and (d) the Ikeda map, and (e) and (f) the Ikeda map with dynamical and 10% observational noise, with x component of the Rössler equations, where we use $A=1.0$ and 39 SSS data. The solid line is the original data and the dotted lines are the SSS data.

the 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 relatively small, because the information in the systems is not retained for longer periods of time. When the data are 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 the Rössler equations are used to generate a slow trend. The equations are given by

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

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 behavior [8]. 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 are equivalent to 56.2% (5 dB) observational noise at each case. See the behaviors in Fig. 1(d).

Figure 5 shows the results for these data. Figures 5(a) and 5(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 the SSS data, however, AC, or AMI, or both are distinct when there is dynamics. In all cases, especially when the time lag is larger, behaviors 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 are 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), both the AC and the 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.

Based on the result of these computational studies, we apply the proposed method to three experimental systems: (1) time intervals of γ-ray emissions of cobalt, which has been recognized as random; (2) a NMR laser data, which has been known to be nonlinear [2]; and (3) daily sunspot numbers from 1 January 1849 to 31 December 2004, which seem to have trends. See Figs. 1(a)–1(c), respectively. We use 10 000 data points for the former two examples, and 56 978 data points for the daily sunspot numbers.

Figure 6 shows segments of the SSS data and the results. Figures 6(a), 6(c), and 6(e) show segments of the SSS data.

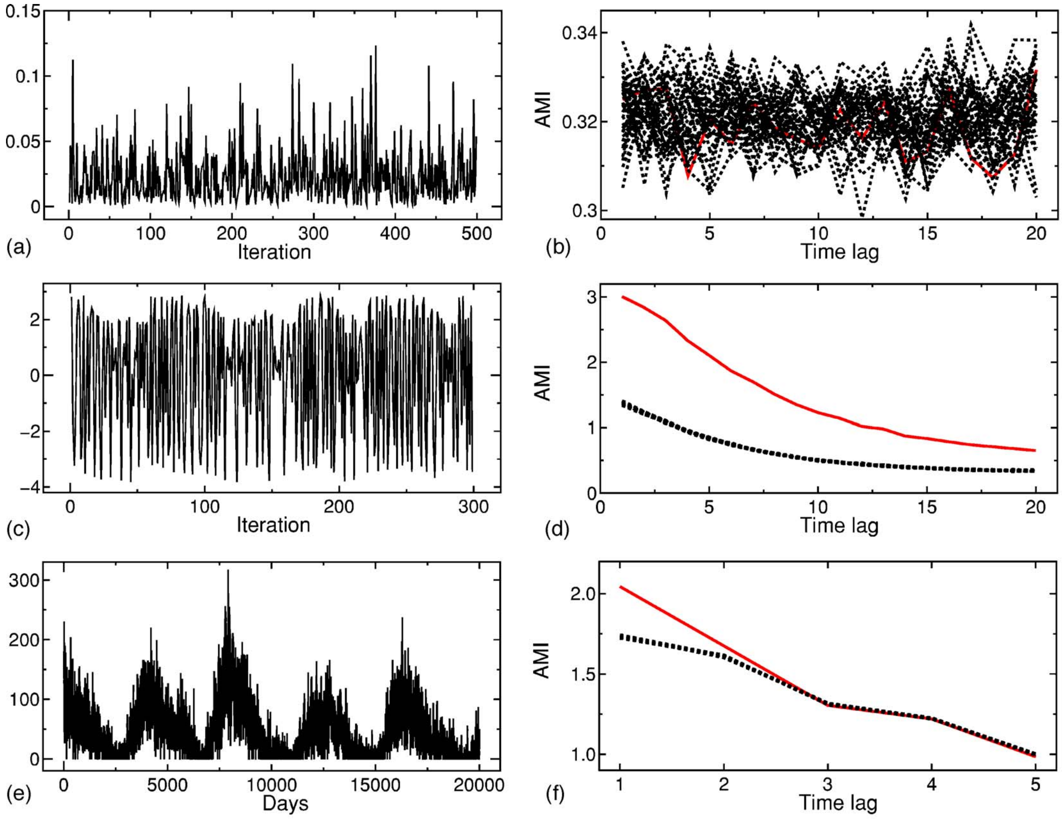

FIG. 6. (Color online) Segments of the SSS data and a plot of the AMI: (a) and (b) cobalt data, (c) and (d) the NMR laser data, and (e) and (f) daily sunspot numbers, where we use $A=1.0$ and 39 SSS data. The solid line is the original data and the dotted lines are the SSS data in (b), (d), and (f). The ordinate axis in (a), (c), and (e) is arbitrary.

Figure 6 do not show significant difference between the original data and the SSS data. Although we do not show the results of the AC in Fig. 6, the results are essentially the same as those of the AMI in all case. Figure 6(b) shows that the AMI of the cobalt data fall within the distributions of the SSS data. Hence, we consider that the cobalt data have no dynamics. Figure 6(d) shows that the AMI of the NMR laser data fall outside the distributions of the SSS data. Hence, we consider that the NMR laser data have dynamics. These results are in agreement with the previously obtained understanding of these data. Figure 6(f) also shows that the AMI of the daily sunspot numbers fall outside the distributions of the SSS data. Hence, we conclude that irregular fluctuations in the daily sunspot numbers have some kind of short term (interday) dynamics.

We have described an algorithm for investigating whether there is dynamics in irregular fluctuations. This method may be applied (and indeed *should* be applied) even if systems and data are contaminated stochastically and data have long term trends.

This research is supported by a Hong Kong University Grants Council Competitive Earmarked Research Grant (CERG) No. PolyU 5216/04E.

[1] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).

[2] H. Kantz and T. Schreiber, *Nonlinear Time-Series Analysis* (Cambridge University Press, Cambridge, 1997).

[3] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996).

[4] J. Theiler and D. Prichard, Physica D **94**, 221 (1996).

[5] M. Small and K. Judd, Phys. Rev. E **59**, 1379 (1998).

[6] K. Ikeda, Opt. Commun. **30**, 257 (1979).

[7] R. M. May, Nature (London) **261**, 459 (1976).

[8] M. Small, D. Yu, and R. G. Harrison, Phys. Rev. Lett. **87**, 188101 (2001).

[9] 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}.

[10] 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 we let $i'(t)$ be $(0.1,-1.3,3.2,4.5,2.7)$. Sorted $i'(t)$ be-

comes $(-1.3,0.1,2.7,3.2,4.5)$ 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).

[11] We have investigated which values of A are the most appropriate using the AC and the 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 a clear difference between the original data and the SSS data even if there is dynamics. When $A \geqslant 1.0$, AC and AMI can give a clear difference. However, when the irregular fluctuations are with trends, $A \geqslant 5.0$ is too large.