Available online at www.sciencedirect.com
ScienceDirect
PHYSICS LETTERS A
Physics Letters A 370 (2007) 265–274
www.elsevier.com/locate/pla

Testing serial dependence by Random-shuffle surrogates and the Wayland method

Yoshito Hirata [a,b,c,*], Shunsuke Horai [b,c], Hideyuki Suzuki [a,c], Kazuyuki Aihara [a,b,c]

[a] *Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*
[b] *Aihara Complexity Modelling Project, ERATO, JST, Japan*
[c] *Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan*

Received 8 September 2006; received in revised form 8 May 2007; accepted 18 May 2007

Available online 25 May 2007

Communicated by A.P. Fordy

Abstract

Given time series, a primary concern is existence of serial dependence and determinism. They are often tested with Random-shuffle surrogates, which totally break serial dependence, and the Wayland method. Since the statistic of the Wayland method fundamentally shows a smaller value for a more deterministic time series, for real-world data, we usually expect that the statistic for the original data is smaller than or equal to those of Random-shuffle surrogates. However, we show herewith an opposite result with wind data in high time resolution. We argue that this puzzling phenomenon can be produced by observational or dynamical noise, both of which may be produced by a low-dimensional deterministic system. Thus the one-sided test is dangerous.

PACS: 05.45.Tp; 05.45.-a; 05.45.Jn

1. Introduction

Surrogate data analysis is a statistical hypothesis testing, where we generate a set of time series that preserve the properties of a hypothesis and compare them with the original data using a statistic [1,2]. If there is a significant difference, we can reject the hypothesis. There are some hypotheses we can test [1]. For testing serial dependence, we often use *Random-shuffle surrogates* [2]. Random-shuffle surrogates are generated by randomly exchanging the order of data in a time series. For the statistic of comparison, we often choose an index for characterising determinism [1]. Among various methods for quantifying determinism [1,3–5], the Wayland method [4] is easy to implement, and widely used for testing the determinism in various systems such as neural networks [6], silicon crystal growth [7], local climate [8], blast furnace [9], dynamics of normal vowels [10], El Nino Southern Oscillation [11], brain activity observed using magnetoencephalograms [12], myocardial dynamics [13], and semiconductor nanowire growth [14]. The statistic of the Wayland method produces a smaller value for a more deterministic time series.[1] Therefore we expect that the statistic for the original data is smaller than or equal to those of their Random-shuffle surrogates and that the one-sided test is enough.

* Corresponding author at: Aihara Laboratory, Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan. Tel.: +81 3 5452 6697; fax: +81 3 5452 6694.

E-mail address: yoshito@sat.t.u-tokyo.ac.jp (Y. Hirata).

[1] In this Letter, we call a time series more deterministic if temporal changes of neighbouring orbits are more coherently directed in state space. Since the Wayland method is a good method for quantifying it [4], we choose the statistic generated by the Wayland method as the standard. For example, when we used the 2-dimensional delay embedding space with delay 1 for scalar time series data with 1000 points each, we got 3.1×10^{-4} as the Wayland statistic for the Hénon map, 0.085 for a linear autoregressive model defined as $x_{t+1} = 0.9x_t - 0.9x_{t-1} + n_t$ where n_t follows the Gaussian distribution with mean 0 and standard deviation 1, and 0.40 for the Gaussian noise with mean 0 and standard deviation 1. Thus, the Hénon map is more deterministic than the linear autoregressive model and the linear autoregressive model is more deterministic than the Gaussian noise in the above meaning.

0375-9601/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physleta.2007.05.061

In this Letter, we intend to give a warning for this simple-minded construction of the hypothesis testing. Namely we applied the above procedure of testing to wind data we observed with high time resolution of 50 Hz, and found a puzzling phenomenon: the test statistic of the Wayland method obtained from the original data became greater than the maximum of those obtained from its Random-shuffle surrogates, meaning that the Wayland method regards the original data as less deterministic than their Random-shuffle surrogates. We discuss that this phenomenon may be explained by observational or dynamical noise. Thus the one-sided test may cause a misleading conclusion.

2. Wayland method

First, we summarise the method proposed by Wayland et al. [4] for detecting determinism in a time series. This method uses the continuity of dynamical evolution as the measure of determinism. Suppose that a scalar time series $\{s_t\}_{t=1}^{N}$ is given. Let τ be the time lag, and d, the embedding dimension. Then we form delay coordinates x_t by $(s_t, s_{t-\tau}, \ldots, s_{t-(d-1)\tau})$. For selecting τ, we choose τ using the first minimum value of the mutual information [15] or the first zero crossing of the autocorrelation [1]. If the mutual information has a minimum and the autocorrelation crosses 0, then we used the smaller τ. If a time series does not satisfy either of these conditions, then we tested several different τ.

Choose l integers between $(d-1)\tau+1$ and $N-m$. Call the set of these integers T. For x_t of $t \in T$, find the k nearest neighbours $x_{n_i(t)}$ such that $|n_i(t)-t| > \tau$ and $|n_i(t)-n_j(t)| > \tau$ for $i, j = 1, 2, \ldots, k$ with $i \neq j$. For notational convenience, we define that $n_0(t) = t$.

For each $x_{n_i(t)}$, we look at its image $x_{n_i(t)+m}$ and take the translation vector,

$$v_i(t) = x_{n_i(t)+m} - x_{n_i(t)}. \tag{1}$$

In this Letter we set $m = \tau$ if we do not mention it for the sake of simplicity. To quantify this notion, let us define the average translation vectors $v(t)$ as follows:

$$v(t) = \frac{1}{k+1}\sum_{i=0}^{k} v_i(t). \tag{2}$$

Using these vectors, we define the translation error $e(t)$ as

$$e(t) = \frac{1}{k+1}\sum_{i=0}^{k} \frac{\|v_i(t)-v(t)\|^2}{\|v(t)\|^2}. \tag{3}$$

The translation error $e(t)$ yields the fractional spread in the displacements relative to the average displacement $v(t)$.

We find the median of $e(t)$ over T, and declare it to be a test statistic for determinism. Wayland et al. demonstrated that the test statistic is close to 0 if a time series is deterministic and the test statistic is about 1 if a time series is random. Our rough explanation for this phenomenon is that if a time series is deterministic, the quantity $\|v_i(t)-v(t)\|$ is small relative to $\|v(t)\|$ and we find small $e(t)$. On the other hand, if a time series is random, the order of $\|v_i(t)-v(t)\|$ is similar to that of $\|v(t)\|$ and thus we have $e(t) \approx 1$. However, since what we discussed so far are two extreme cases, we are not sure what will happen in between. Here, we set $k = 4$ and $l = 100$ if they are not stated.

If we use the one-sided test, we should reject the null-hypothesis of no serial dependence when there are more than a certain number of embedding dimensions where the Wayland statistic obtained from the original data is smaller than the minimum of those obtained from the Random-shuffle surrogates. Instead, if we use the two-sided test, we should reject the null-hypothesis when there are more than a certain number of embedding dimensions where the Wayland statistic obtained from the original data is out of the interval defined by the Wayland statistics obtained from the Random-shuffle surrogates. In this Letter, we set the "certain number" to five so that when we use the two-sided test with 39 surrogates, the total significance level becomes 0.01.

3. Application to wind data

We applied surrogate data analysis with the Wayland statistic to a time series of wind. The wind was measured by a ultrasonic anemometer at 50 Hz on the top of the building of Institute of Industrial Science, The University of Tokyo in Komaba, Tokyo, at 14:00 JST, on 18 August 2004, for 1 hour. A part of this time series is shown in Fig. 1(a). We confirmed using the method of Kennel [16] that this part of the time series is stationary.

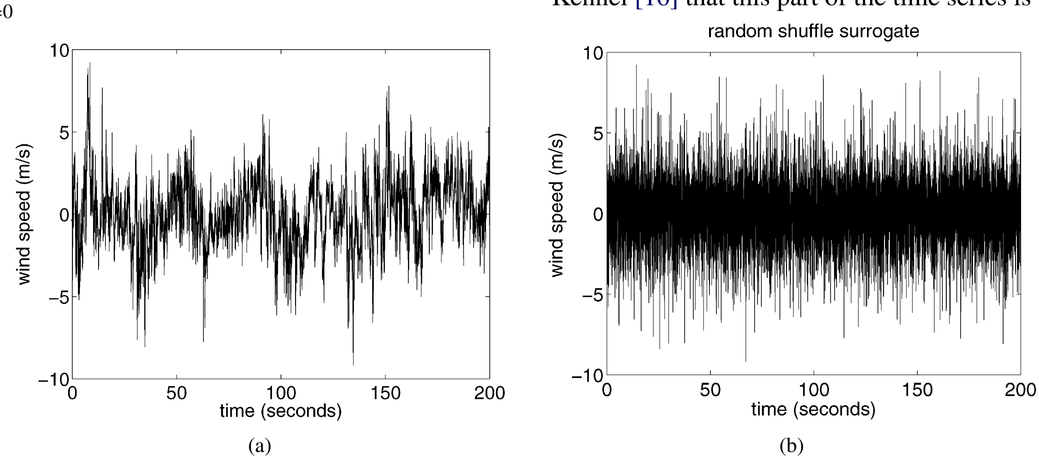

Fig. 1. (a) An example of the wind data and (b) its Random-shuffle surrogate.