Testing for linear and nonlinear Gaussian process in nonstationary time series[☆]

Ricardo Araújo Rios[a,*], Michael Small[b], Rodrigo Fernandes de Mello[a]

[a] *Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador Sãocarlense, 400, São Carlos, SP, Brazil, P.O. Box 668*
[b] *School of Mathematics and Statistics, The University of Western Australia, Crawley, WA, Australia, 6009*

Abstract

Surrogate data methods have been widely applied to produce synthetic data, while maintaining the same statistical properties as the original. By using such methods, one can analyze time series behavior or increase the data set size when few observations are available. Theiler's surrogate data methods — the most commonly considered approaches — are based on the Fourier transform. Due to the use of Fourier transform, the application of such methods is limited to stationary time series, the presence of a trend produces spurious high frequencies, and hence generates inconsistent surrogates. To solve this problem, we present two new methods that combine time series decomposition techniques and surrogate data methods. These new methods initially decompose time series into a set of monocomponents, separating the deterministic signal and trend. Afterwards, traditional surrogate methods are

[☆]This paper is based upon work supported by FAPESP (Sao Paulo Research Foundation), Brazil, under the grants 2011/02655-9 and 2009/18293-9.

[*]Corresponding author

Email addresses: rios@icmc.usp.br (Ricardo Araújo Rios), michael.small@uwa.edu.au (Michael Small), mello@icmc.usp.br (Rodrigo Fernandes de Mello)

applied on those individual monocomponents and a set of surrogates is generated. Finally, all individual surrogates plus the trend signal are combined in order to create a single surrogate series. With this method, one can investigate linear and nonlinear Gaussian processes in time series, irrespective of the presence of nonstationary behavior.
Keywords: Surrogate Data, Decomposition Method, Fourier-based Method, Nonstationary Time Series, Nonlinear Time Series

1. Introduction

Surrogate data methods [1, 2, 3] are traditionally applied on experimental data to test it against specific null hypotheses. This is achieved by algorithmically generating an ensemble of surrogate data: each surrogate data set is expected to be similar to the original data, but also consistent with the underlying null hypotheses. Specifically, properties of this underlying null hypothesis (for example linear correlation for the null of linearly filtered noise) will produce the same statistical estimate from the original data and the surrogates. However, other properties of data, unrelated to the null hypothesis, are randomized. In other words, only features consistent with the null are maintained so that statistical sampling from either surrogates or the original data will provide the same results.

A secondary application of surrogate data is ...

The most commonly considered surrogate data methods are based on the Fourier transform, such as the Fourier Transformed (FT)[1] and Amplitude

[1] In this paper, FT stands for the Theiler's Fourier Transformed method used to produce surrogates. On the other hand, the basic Fourier transform is referred to as $\mathcal{F}$.

Adjusted Fourier Transformed (AAFT) methods proposed by Theiler *et al.* [1]. These methods basically apply the Fourier transform on the original data, obtaining amplitudes and phases, and substitute the original phases by uniform random phases. Afterwards, they apply the inverse Fourier transform to obtain the surrogate data. As this new data was generated using the Fourier transform, which assumes data periodicity, nonstationary characteristics are not represented in the surrogate. The main reason behind this is that by applying the Fourier transform on nonstationary time series, the difference between the first and last observations, caused by the trend, produces spurious high frequencies, as a consequence, inconsistent surrogates are produced [4, 5].

Aiming at overcoming this problem, we extended these two Fourier-based methods by initially decomposing the time series into a set of components plus a residue. Every component contains similar behavior and the residue contains trends. Then, surrogate data is produced based on each individual component. Later on, all produced surrogates plus the original residue are added to compose the surrogate, which is indeed considered as synthetic data based on the original time series. To separate components, we consider the Empirical Mode Decomposition (EMD) method. The last component produced by EMD, i.e., the trend, contains nonstationary features. According to our experiments, we confirmed that our approach improves Theiler's methods to produce synthetic data to nonstationary time series.

The remainder of this paper is organized as follows. In Section 2, we present an overview about the surrogate data methods and we discuss important surrogate methods. The proposed approach is presented in Section 3.

In Section 4, we present an analysis of the proposed approach. Experimental results as well as a detailed discussion about the advantage of our approach are presented in Sections 5 and 6; finally, in Section 8, we draw conclusions and discuss future work.

2. Surrogate Methods

The study of surrogate data was introduced by Theiler *et al.* [1], whose the main objective was to analyze time series to confirm whether they belong to the same generation process. In general, this evaluation is performed in two straightforward steps. First, synthetic data is produced combining part of the original data properties and another specific generation process. This step is repeatedly performed to produce a set of surrogates. In the second step, discriminating statistics are computed to compare the original time series against all surrogates. Based on the computed values, one can verify the similarity among them, and, consequently, state whether or not they were created using the same process.

The discriminating statistics can be computed using different methods, such as the Grassberger-Procaccia (GP) correlation dimension [6], Autocorrelation Function (ACF) [7], Spectral Density (SD) [8], Average Mutual Information (AMI) [9] and Space-Time Separation Plot (STP) [10].

The results obtained by the discriminating statistics are then used to perform Statistical Hypothesis tests which assess the null hypothesis that the original and surrogate data are similar. For example, if we intend to test the null hypothesis that the original data is linear, we can produce surrogate data guaranteeing it will be linear. Thus, if the original and surrogate data

have the same properties, we accept the null hypothesis, which simply states we failed to find evidence that the original data is not linear. However, if the null hypothesis is rejected then we have shown with some statistical confidence that the original data is not linear.

Initially, Theiler *et al.* [1] proposed two methods to generate surrogate data which have been widely studied and employed [2, 11, 12]: the Fourier Transformed (FT) surrogate; and the Amplitude-Adjusted Fourier Transformed (AAFT) surrogate.

The FT method was designed to identify the nonlinear property in time series. This method defines as the null hypothesis that the analyzed time series is linear [1, 2]. Hence, this method produces surrogates using a linear process. Afterwards, discriminating statistics are computed on the original and surrogates. Finally, if statistics are significantly different, the null hypothesis is rejected.

In order to better understand the FT method, consider a time series $x(t) = \{x(1), x(2), \cdots, x(N)\}$ with length N. Let $X(f)$ be the coefficient produced by the Fourier transform $\mathcal{F}$ at frequency f on $x(t)$, Equation 1.

$$X(f) = \mathcal{F}(x(t)) = \int_{-\infty}^{\infty} x(t) \cdot e^{-i2\pi ft} dt \tag{1}$$

The previous equation can also be rewritten in terms of its amplitude $A(f)$ and phase $\phi(f)$ as shown in Equation 2.

$$X(f) = A(f) \cdot e^{i\phi(f)} \tag{2}$$

Then, the phase-randomized Fourier transform is obtained by rotating

the phase at each frequency f considering an independent random variable φ which is uniformly chosen within the interval $[0, 2\pi)$ [1], as shown in Equation 3.

$$\tilde{X}(f) = A(f) \cdot e^{i[\phi(f)+\varphi(f)]} \tag{3}$$

Given the inherent linearity of the uniform distribution, this randomization creates surrogates with phases varying in a linear way. Finally, the surrogate data $y(t)$ is obtained by applying the Inverse Fourier transform $\mathcal{F}^{-1}$ (Equation 4) [1].

$$y(t) = \mathcal{F}^{-1}\{\tilde{X}(f)\} = \mathcal{F}^{-1}\left\{\int_{-\infty}^{\infty} X(f) \cdot e^{i\varphi(f)} df\right\} \tag{4}$$

The second surrogate data method proposed by Theiler *et al.* is called the Amplitude Adjusted Fourier Transformed (AAFT) method. In this case, the null hypothesis assumes that besides the time series dynamics are linear, observations may be influenced by a nonlinear static transform [1]. According to authors, most conventional methods used to estimate nonlinearity indicate that a given time series is nonlinear, but they do not provide further information to conclude if the nonlinearity comes from the time series dynamics or from the amplitude distribution [1]. On the other hand, by using AAFT, surrogates are produced respecting the same amplitude distribution of the original time series and presenting similar ACF, but not equal, once there is an adjustment on the amplitude [1, 13]. Aiming at improving AAFT to produce surrogates that preserve both amplitude distribution and ACF, Schreiber *et al.* [14] proposed a new method called Iterative Ampli-

tude Adjusted Fourier Transformed (IAAFT). However, this method is not considered in our comparative study.

The AAFT method [1] receives the original time series $x(t)$ and then computes a rank for each observation. Then, the computed ranks are used to sort the $x(t)$ observations in an increasing order, returning a new series $x_r(t)$. Next, AAFT generates a new time series $y(t)$ using a Gaussian process. This new time series is reordered so that the ranks agree with the $x_r(t)$ ranks. After that, the FT method is applied on $y(t)$ generating a new series $y'(t)$. Finally, AAFT produces the surrogate data $x_s(t)$ for $x(t)$ by reordering the observations in $x(t)$ in a way that its ranks agree with the $y'(t)$ ranks. [1].

The main problem faced by these methods is related to the stationarity restriction imposed by the Fourier transform. Theiler's FT and AAFT methods cannot create surrogate data sufficiently similar to time series characterized by nonstationary behavior. In such situation, the surrogate produced by these methods is affected by the amplitude variation in the Fourier transform, producing surrogates completely different from the original time series [4, 5].

Another surrogate method called Small Shuffle Surrogate (SSS) was proposed by Nakamura and Small[5]. To generate surrogate data, this method performs the following steps: i) the original time series $x(t)$ is analyzed and the indices of its observations are stored in a list $i(t)$; ii) a new index list is created by considering equation $i'(t) = i(t) + A \cdot g(t)$, in which A represents an amplitude and $g(t)$ is a sequence of Gaussian random numbers. In this equation, the amplitude A is responsible for defining the scale of changes in the index list $i(t)$; iii) list $i'(t)$ is sorted and stored in a new list $\hat{i}(t)$; iv)

finally, a surrogate $s(t)$ is obtained by selecting values of the original time series $x(t)$ according to new indexes $\hat{i}(t)$, i.e., $s(t) = x(\hat{i}(t))$.

According to authors, this method can be used to investigate irregular fluctuations in time series, once it destroys local structures or correlations and keeps the global behavior, such as trend. Hence, the null hypothesis addressed by this new method is that the time series consists of a underlying (slow) trend and that the fast dynamics are random.

The main limitation of the SSS method comes from the difficulty to indicate if data are linear or nonlinear, because both behaviors are characterized by some dynamics [4]. Aiming at solving this limitation, the same authors presented a new method called Truncated Fourier Transform Surrogate (TFTS), which produces surrogates by randomizing phases only in the high-frequency domain. For this, they define a threshold f_ε to determine whether phases may change or not, i.e., if phases are characterized by frequencies higher than f_ε, then they are randomized; otherwise, they remain the same. Since high-frequency phases are randomized, the nonlinearity present in irregular fluctuations is destroyed, whereas the global behavior is preserved by the untouched phases. Hence, the null hypothesis addressed by the TFTS method states that irregular fluctuations are generated by stationary linear systems [4].

The TFTS method was later considered by Lucio *et al.* [13], who presented two new techniques named AAFT_{TD} and IAAFT_{TD}. Similarly to our approach, the authors designed these techniques to preserve the global nonstationarity present in time series. In summary, the techniques consist of detrending and retrending the time series, applying the TFTS and AAFT (or

IAAFT) method to generate the surrogate data. These last three methods emphasize the importance of producing surrogate data considering nonstationary behavior in time series as approached in this work.

The main problem with the TFTS method, and consequently with the methods based on it, is the need for setting a value for parameter f_ε, which determines the frequencies to be randomized. In summary, when this parameter assumes low values, most of the phases will be randomized, producing surrogates very similar to the traditional FT method. On the other hand, high values mean the surrogate will be very close to the original data.

Concluding, the nonstationary problem faced by the traditional Theiler FT and AAFT methods and the restriction imposed by parameter f_ε to the TFTS method have motivated the development of two new methods presented in the following section.

3. Improving surrogate methods by decomposing time series

The surrogate methods presented in this paper were designed to solve the problem faced by Theiler's Fourier Transformed (FT) and Amplitude Adjusted Fourier Transformed (AAFT) methods when analyzing nonlinear time series. As previously presented, by applying these Fourier-based methods on time series with trends, the produced observations are influenced by spurious high frequencies, which affect the general behavior of surrogates. In order to overcome this drawback, we initially decompose time series into a set of monocomponents plus a residue, which represents the time series trend. Afterwards, we apply traditional surrogate methods on every monocomponent, producing a set of monocomponent surrogates. Those individual

surrogates are combined to produce a single surrogate data, which is finally retrended by adding the residue obtained in the first step. These detrending and retrending steps allow to preserve the global nonstationarity in the surrogate series.

This decomposition step was performed by using of the Empirical Mode Decomposition (EMD) method [15], which reduces the time series into monocomponents, also called Intrinsic Mode Functions (IMFs), revealing important information embedded in the original series [15]. IMFs are also referred to as monocomponent due to its characteristic of representing only one frequency at any given time instant, supporting the study of instantaneous frequencies and amplitudes using the Hilbert Spectral Analysis (HSA) [15]. The most important advantage of using EMD is the possibility of decomposing time series irrespective of its generation processes, i.e., the decomposition proccess is not affected by the nonlinearity, nonstationarity, and/or stochasticity present in time series.

The key point to perform this decomposition is the sifting process, which initially identifies local minima and maxima values for observations along time. Afterwards, these extrema are connected through the cubic spline method and, thus, the upper and lower envelopes are defined, which must cover all values [15]. Then, mean $m_1(t)$ of these envelopes is calculated and the first monocomponent candidate $c_1(t)$ is obtained by using Equation 5, in which $x(t)$ represents the analyzed time series.

$$c_1(t) = x(t) - m_1(t) \tag{5}$$

Then, the first monocomponent candidate $c_1(t)$ is used in place of the

original data and the sifting process is repeated k times, producing $c_{1k}(t) = c_{1(k-1)}(t) - m_{1k}(t)$. This process continues until $c_{1k}(t)$ satisfies the IMF monocomponent definition [15]: i) the number of extrema and the number of zero-crossings must be either equal or differ at most by one; or ii) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. After retrieving the first monocomponent that satisfies the IMF definition, i.e., $c_{1k}(t)$, the first IMF monocomponent is obtained according to Equation 6.

$$h_1(t) = c_{1k}(t) \tag{6}$$

Then, this first IMF is removed from data, i.e., $x(t) - h_1(t)$, and the resultant series is again analyzed by the whole process, producing further IMFs until reaching a stop criterion. This criterion is defined when the last component $h_n(t)$ becomes a monotonic function, avoiding the extraction of further components. Hence, this last component is referred to as final residue $r_M(t)$ [15]. In summary, according to EMD, a time series $x(t)$ is composed of a set of monocomponents plus a residue as presented in Equation 7, in which M represents the number of monocomponents obtained from time series $x(t)$.

$$x(t) = \sum_{m=1}^{M-1} h_m(t) + r_M(t) \tag{7}$$

The decomposition step in our method permits detrending the time series, by the residue extraction, before applying any surrogate method. Hence, as next step, our method executes the Theiler's FT method on all decomposed monocomponents $\{h_m(t)\}$, except on residue $r_M(t)$, producing a set of

monocomponent surrogates.

In the last step, all monocomponent surrogates are summed to form a single surrogate data. Finally, the global trend of the original time series is combined to this single surrogate by adding the residue. In summary, this new adapted method, called EMD-FT, is defined by Equation 8, in which $y(t)$ is the surrogate data, $X_{m,k}(f)$ represents the coefficients obtained applying the Discrete Fourier transform on the *m-th* monocomponent, and $\varphi(f)$ represents the values obtained with the phase randomization.

$$y(t) = \sum_{m=1}^{M-1} \left(\frac{1}{N} \sum_{k=1}^{N} X_{m,k}(f) \cdot e^{i\varphi(f)} \right) + r_M(t) \tag{8}$$

On the other hand, the adapted EMD-AAFT method was created using exactly the same steps previously presented, however the AAFT method was adopted to produce surrogate data for every monocomponent instead.

One of the most important contributions of the proposed methods is the possibility of removing nonstationary influences during the decomposition of the original time series. After decomposition, every monocomponent contains simpler behavior which is better represented using sinusoidal functions. Thus, when we apply Theiler's surrogate data methods on every monocomponent to produce surrogates, except the residue which represents the time series trend. Consequently, by using FT and AAFT in our methods, one can test the linearity in stationary or nonstationary time series, what is not possible using Theiler's methods directly on the original time series.

In order to evaluate the proposed methods, in the next section, we analytically demonstrate that surrogates produced by the original FT and AAFT methods are similar to the proposed EMD-FT and EMD-AAFT methods,

even when there is a trend embedded in the time series.

4. Analyzing the proposed surrogate methods

In order to investigate the efficiency of the proposed methods, we analyzed the surrogate data produced by FT and EMD-FT methods. The main objective of this analysis is to assure that the adapted method produces surrogates with the same behavior as the traditional one and, consequently, the same null hypothesis can be used by both methods, even when there is nonstationary behavior in data.

All analysis presented in this section is based on constant phase randomization, *i.e.*, the same phase randomization is applied on every IMF, in order to simplify the stated theorem and proof. However, this is not mandatory, because one can employ our methods to apply different phase randomizations to IMFs what is discussed in Section 7.

It is important to highlight that we used the EMD-FT method because it is simpler and more intuitive, but the same results can be extended to the EMD-AAFT method. In order to proceed with this analysis, we first present a theorem which states that the surrogate data produced by both methods are exactly the same, even when there is trend embedded in the time series. By proving this theorem, we confirm the same hypothesis test used by the Theiler's FT method can be adopted for the EMD-FT method. In this sense, we can formally define our hypothesis:

Theorem: *If the trend can be separated from nonlinear time series, then Theiler's FT and the proposed EMD-FT method produce the same surrogate*

data.

Proof: To validate this theorem, we need to prove that the surrogate data produced by both methods are exactly the same to linear and nonlinear time series. However, before proceeding with this analysis, we postulate the trend can somehow be detached from time series.

Therefore, considering this postulate, we can rewrite a nonlinear time series as $x(t) = z(t) + r_M(t)$, in which $z(t)$ represents the time series observations and $r_M(t)$ is the trend. We used $r_M(t)$ to represent the trend just to keep the same pattern used to describe the Empirical Mode Decomposition (EMD) method. Hence, in the first part of our proof, after applying Theiler's FT surrogate method on the observations of $z(t)$, we obtain surrogate $y(t)^{FT}$.

$$y(t)^{FT} = FT(z(t)) + r_M(t) \tag{9}$$

Considering the definition of Fourier transform (Equation 1), we can rewrite the previous equation as:

$$y(t)^{\mathrm{FT}} = \frac{1}{N}\sum_{k=1}^{N} X_k(f) \cdot e^{i\varphi(f)} + r_M(t) \tag{10}$$

The second part of our proof is, initially, obtained by applying the EMD method on the same nonlinear time series $x(t)$, which returns a set of mono-components $\{h_m(t)\}$ and a residue $r_M(t)$. By definition, residue $r_M(t)$ represents the time series trend, once the EMD method was previously proved to be nonlinear and useful to retrieve the trend from time series [16]. Afterwards, Theiler's FT method is also used to generate surrogates, but Equation

1 may be individually applied on every monocomponent $h_m(t)$, resulting in surrogate $y(t)^{EMD-FT}$.

$$\begin{aligned} EMD(x(t)) &= h_1(t) + \cdots + h_m(t) + r_M(t) \\ y(t)^{\text{EMD-FT}} &= FT(h_1(t)) + \cdots + FT(h_m(t)) + r_M(t) \\ &= \left(\frac{1}{N}\sum_{k=1}^{N} X_{1,k}(f) \cdot e^{i\varphi(f)}\right) + \cdots + \left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t) \cdot e^{i\varphi(f)}\right) + r_M(t) \end{aligned} \tag{11}$$

An important step of our proof is stated by assuming the phase randomization is performed only once for all monocomponents $h_m(t)$, i.e., all monocomponents were randomized considering the same sequence of values. Hence, surrogate $y(t)^{EMD-FT}$ can be rewritten evidencing the randomized phase according to Equation 12.

$$\begin{aligned} y(t)^{\text{EMD-FT}} &= \left[\left(\frac{1}{N}\sum_{k=1}^{N} X_{1,k}(t)\right) + \cdots + \left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t)\right)\right] \cdot e^{i\varphi(f)} + r_M(t) \\ y(t)^{\text{EMD-FT}} &= \left[\sum_{m=1}^{M-1}\left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t)\right)\right] \cdot e^{i\varphi(f)} + r_M(t) \end{aligned} \tag{12}$$

Finally, in order to prove the theorem stated in this section, we need to evaluate the relation $y(t)^{FT} = y(t)^{EMD-FT}$, i.e., observations generated by both methods are equal.

$$
\begin{aligned}
y(t)^{FT} &= y(t)^{EMD-FT} \\
\left(\frac{1}{N}\sum_{k=1}^{N} X_k(f)\cdot e^{i\varphi(f)}\right) + r_M(t) &= \left[\sum_{m=1}^{M-1}\left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t)\right)\right]\cdot e^{i\varphi(f)} + r_M(t)
\end{aligned}
\tag{13}
$$

By subtracting $r_M(t)$ from both sides of Equation 12, we obtain the equality shown in Equation 14.

$$
\left(\frac{1}{N}\sum_{k=1}^{N} X_k(f)\cdot e^{i\varphi(f)}\right) = \left[\sum_{m=1}^{M-1}\left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t)\right)\right]\cdot e^{i\varphi(f)}
\tag{14}
$$

Finally, we divide both sides of the equation by $e^{i\varphi(f)}$:

$$
\begin{aligned}
\left[\left(\frac{1}{N}\sum_{k=1}^{N} X_k(f)\right)\right]\cdot e^{i\varphi(f)} &= \left[\sum_{m=1}^{M-1}\left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t)\right)\right]\cdot e^{i\varphi(f)} \\
\left(\frac{1}{N}\sum_{k=1}^{N} X_k(f)\right) &= \sum_{m=1}^{M-1}\left(\frac{1}{N}\sum_{k=1}^{N} X_{m,k}(t)\right)
\end{aligned}
\tag{15}
$$

This equality proves the sum of amplitudes, obtained by applying Theiler's FT on EMD decomposed monocomponents, is equal to the amplitude obtained using Theiler's FT surrogate method directly on the time series. □

Therefore, we confirm the new method supports the same null hypothesis

as Theiler's FT and AAFT, but without any interference of the nonstationary behavior, once the nonlinear EMD method permits treating the trend as a separated component.

In the following section, we present the experimental setup to evaluate the proposed methods.

5. Experimental setup

In order to evaluate the proposed methods, we analyzed two sets of time series[2]. The first one was composed of three synthetic time series created by adding a trend to a sine function (Figure 1.a), a white noise process (Figure 1.b), and an autoregressive process (Figure 1.c). The autoregressive (AR) process used in these experiments was generated considering a first-order model and a Normal distribution $N(0, \sigma^2)$. This Normal distribution was also used to create the white noise time series presented in Figure 1.b.

The second set was composed of three real-world time series. The first series is illustrated in Figure 1.d, which corresponds to a collection of yearly average global temperatures [17]. Figure 1.e presents the second time series that corresponds to atmospheric concentrations of CO2 [18]. Finally, Figure 1.f represents the Dow Jones Utilities Index commonly considered in stock market analysis [8].

[Figure 1 about here.]

The evaluation process was performed by analyzing every time series and producing 99 surrogates using the methods FT, AAFT, EMD-FT, and EMD-

[2]The time series used in this experiments, as well as the source code, are available in: `http://www.icmc.usp.br/~rios/surrogate/`

AAFT. Then, the original and surrogate time series were evaluated considering three types of analyses: i) visual inspection of time series plots; ii) visual inspection of plots produced by the Autocorrelation Function (ACF) [10] and the Average Mutual Information (AMI) [9, 5, 4]; iii) hypothesis test considering the discriminating statistics obtained using AMI.

The Autocorrelation Function allows to identify temporal correlations present in series at different lags in time, depicting the difference among the methods. Formally, the Autocorrelation function $\hat{\rho}(h)$ of a time series $x(t)$ is obtained by computing the autocovariances of $x(t)$ and its time-shifted version $x(t+h)$ as defined in Equation 16, in which $E[\cdot]$ is the expected value of the expression, and μ and σ^2 are the variance and mean of $x(t)$, respectively.

$$\hat{\rho}(h) = \frac{E[(x(t) - \mu)(x(t+h) - \mu)]}{\sigma^2} \tag{16}$$

In summary, ACF allows to analyze the similarity among time series observations. By considering it as discriminating function, we can evaluate whether the similarity among the original time series observations agrees with the similarities of surrogates or not.

The Average Mutual Information (AMI) can be considered a nonlinear version of ACF, which helps to determine the dependence between past and future observations [9, 5, 4]. Equation 17 defines AMI, in which $p(x(t))$ represents the marginal probability distribution function of $x(t)$ and $p(x(t), x(t+h))$ is the joint probability distribution function of $x(t)$ and $x(t+h)$, having h as the time lag. In all experiments, the time lag as varied within interval $[1, 20]$ and sixteen bins were used to discretize data and estimate probabili-

ties.

$$I(h) = \sum_{x} p(x(t), x(t+h)) \log \left(\frac{p(x(t), x(t+h))}{p(x(t))\, p((t+h))} \right) \tag{17}$$

6. Experiments

This section presents the experimental results in two subsections. First, we analyzed the synthetic time series, then the real ones.

6.1. Synthetic Time Series

In the first synthetic experiment, we analyzed a time series created by the combination of a sine function and a linear sequence of observations, added to simulate a trend behavior. This time series (red continuous line) and the surrogates (dashed lines) generated by FT, AAFT, EMD-FT, and EMD-AAFT are illustrated in Figures 2.a, Figures 2.d, 2.g, and 2.j, respectively. By visually inspecting these plots, we observe EMD-FT produced surrogates whose behavior is very similar to the original series. Although the surrogates produced by the EMD-AAFT method are also close to the original time series, we notice the presence of small noise changing the expected behavior.

These conclusions were also drawn by analyzing the discriminating statistics. According to ACF (Figures 2.b, Figures 2.e, 2.h, and 2.k) and AMI (Figures 2.c, Figures 2.f, 2.i, and 2.l) plots, the original data (red continuous line) only falls within the surrogate distribution (dashed lines) produced by the EMD-FT method.

[Figure 2 about here.]

In the second synthetic experiment, we analyzed a time series (Figure 3.a) created by the combination of a white noise process and a linear sequence of observations used to simulate a trend behavior.

[Figure 3 about here.]

By analyzing the plots in Figure 3, we confirm the proposed methods produce surrogates, whose behavior is closer to the original time series than the Theiler's FT and AAFT methods. In case of EMD-FT and EMD-AAFT, we observe no significant differences.

The last synthetic time series was created combining an autoregressive process to a linear sequence of observations. Observing Figure 4, the proposed methods generated surrogates more consistent with the original time series. We highlight there is no significant difference between EMD-FT and EMD-AAFT when visually inspecting time series plots.

[Figure 4 about here.]

In the next section, we present the results obtained when considering real-world time series.

6.2. Real-World Time Series

The first time series analyzed in this section is presented in Figure 5.a. This time series was studied in [17] and it is composed of yearly average values of global temperatures.

[Figure 5 about here.]

By analyzing this series, we realized observations follow a trend, i.e., the mean temperature increases over time, characterizing the time series as nonstationary. Nevertheless, the surrogates created by Theiler's FT and

AAFT cannot represent such trend. On the other hand, the EMD-FT and EMD-AAFT methods produced surrogates considering the nonstationarity of the original time series. Analyzing only the AMI plots, we conclude all methods produced surrogates, which are compatible with the original time series, once they are within the surrogate distribution. However, in this situation, the ACF plot clearly indicates Theiler's FT and AAFT surrogates are different from the original time series, what is evident by plots on the left side of Figure 5.

The second real-world time series considered in this study is composed of atmospheric concentrations of CO2 and has a similar behavior to the synthetic series presented in Figure 2. This series is characterized by some trend and cyclical behavior. In this scenario, the best surrogates were generated using EMD-FT and EMD-AAFT, as expected due to presence of trend. This is also confirmed by the ACF and AMI plots. In this situation, there is no significant difference between the EMD-FT and EMD-AAFT surrogates.

[Figure 6 about here.]

Finally, the last experiment was performed on the Dow Jones Utilities Index, which was recorded from August 28^{th} to December 18^{th}, 1972[3] [8]. This time series (Figure 7) has a trend behavior as well, benefiting the EMD-FT and EMD-AAFT methods. This conclusion is also evident in the ACF plots. However, according to the AMI plots, the only ineffective method was Theiler's AAFT. In such situation, EMD-FT and EMD-AAFT surrogates have similar behavior.

[3]Although there are most current observations for this dataset, we used this period due to its adoption in several papers and textbooks on time series analysis.

[Figure 7 about here.]

Finally, we also applied a hypothesis test on the discriminating statistics produced by the Average Mutual Information. Thus, we applied hypothesis tests to compare the original time series against every surrogate produced by all methods. Then, we computed the average p-value $\mu_{p-value}$ for every method. At last, the following hypothesis test was applied to compare the methods, which considers a significance level of 0.01 for the one tailed test.

$$\begin{cases} H_0: & \mu_{\texttt{p-value}} \geq 0.01 \\ H_a: & \mu_{\texttt{p-value}} < 0.01 \end{cases} \tag{18}$$

This test accepts the null hypothesis when the average p-value is greater than 0.01, otherwise, we accept the alternative hypothesis, which allows us to infer the surrogate and original time series were not produced using the same generation process. The obtained results were summarized in Table 1, in which letters A and R mean the null hypothesis was accepted or rejected, respectively.

[Table 1 about here.]

According to Table 1 the surrogates generated by Theiler's FT and AAFT were significantly different from the original time series in most situations, hence the null hypothesis was rejected for most surrogates. On the other hand, the proposed EMD-FT and EMD-AAFT methods provided greater p-values, showing their surrogates are more similar to the original time series.

By analyzing the results obtained with the synthetic and real-world time series, we conclude the proposed EMD-FT and EMD-AAFT methods provide effective surrogates, which respect the behavior of original time series. This

is evident in the presence of a trend, which affects Theiler's methods but not EMD-FT and EMD-AAFT.

7. Discussion on the phase randomization

In our proof, we applied the same phase randomization for all monocomponents, *i.e.*, after extracting a set of IMFs for a nonlinear time series, the same variable φ (Equation 3) used to rotate the phase of the first IMF is again used for the remaining IMFs. This assumption was used to simplify the analysis of our methods. By using a constant value for the phase, we were able to prove that our methods produce similar surrogates to Theiler's FT and AAFT methods, but without any stationary influence.

This assumption is not mandatory, what means we can produce the final surrogate by combining IMFs at different phase randomizations. Using this process, we can use our methods to: i) filter only deterministic IMFs [19] and apply phase randomization to produce more representative surrogates, once the stochastic behavior may be out of scope for some application domains, such as signal and image processing; ii) filter time series trends out and produce surrogates only considering the relevant behavior which is represented by IMFs. At last, we only add trends to compose the final surrogate, maintaining the nonstationary characteristic of the original time series (as approached in this work) what is not fulfilled by Theiler's FT and AAFT methods; iii) filter IMFs according to amplitudes to produce surrogates at different randomization levels. For example, consider the time series shown in Figure 1.d, which corresponds to a collection of yearly average global temperatures. By applying the EMD method on this time series, a set of IMFs

is obtained as shown in Figure 8. We notice the amplitudes of IMFs vary significantly. Hence, the phase randomization of a low-amplitude IMF adds no significant information to the final surrogate.

[Figure 8 about here.]

Finally, even using constant phase randomization, the analysis on the null hypothesis for our methods (see Section 4) remains consistent.

8. Concluding remarks

In this paper, we discussed the problem faced by Theiler's FT and AAFT methods when time series present nonstationary behavior. In such situation, surrogates produced by these methods are very different from the original data. By applying statistical methods or even performing a visual inspection on the original and surrogate time series, we cannot state whether they were created from the same process or not.

In order to address this drawback, we proposed two new methods based on Theiler's techniques. The new methods initially decompose the time series into monocomponents that are, in a second step, transformed by either Theiler's FT or AAFT method. As a result, a set of monocomponent surrogates is produced, which are combined with the original time series trend to create the surrogate time series.

Experimental results on synthetic and real-world time series confirmed the proposed methods produced surrogates in accordance to the original data in presence of nonstationarity. This is made possible due to the extraction of the series trend, which adds spurious high frequencies. As consequence, the proposed methods support the linear/nonlinear test for stationary and

nonstationary time series, what is not possible when directly using Theiler's methods.

References

[1] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J. D. Farmer, Testing for nonlinearity in time series: the method of surrogate data, Physica D: Nonlinear Phenomena 58 (1992) 77–94.

[2] T. Maiwald, E. Mammen, S. Nandi, J. Timmer, Surrogate data - a qualitative and quantitative analysis, in: R. Dahlhaus, J. Kurths, P. Maass, J. Timmer (Eds.), Mathematical Methods in Signal Processing and Digital Image Analysis, volume 27 of *Understanding Complex Systems*, Springer Berlin / Heidelberg, 2008, pp. 41–74.

[3] T. Schreiber, A. Schmitz, Surrogate time series, Physica D: Nonlinear Phenomena 142 (2000) 346 – 382.

[4] T. Nakamura, M. Small, Y. Hirata, Testing for nonlinearity in irregular fluctuations with long-term trends, Physical Review E 74 (2006) 026205.1 – 026205.8.

[5] T. Nakamura, M. Small, Small-shuffle surrogate data: Testing for dynamics in fluctuating data with trends, Physical Review E 72 (2005) 056216.1 – 056216.6.

[6] P. Grassberger, I. Procaccia, Characterization of strange attractors, Phys. Rev. Lett. 50 (1983) 346–349.

[7] G. Box, G. M. Jenkins, G. Reinsel, Time Series Analysis: Forecasting & Control, Prentice Hall, 3^a edition, 1994.

[8] P. J. Brockwell, R. A. Davis, Introduction to Time Series and Forecasting, Springer, 2002.

[9] H. Abarbanel, Analysis of Observed Chaotic Data, Institute for Nonlinear Science, Springer, 1996.

[10] A. Provanzale, L. A. Smith, R. Vio, G. Murante, Distinguishing between low-dimensional dynamics and randomness in measured time series, Physica D: Nonlinear Phenomena 58 (1992) 31–49.

[11] M. Small, C. Tse, Detecting determinism in time series: the method of surrogate data, IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 50 (2003) 663 – 672.

[12] I. Tohru, A. Kazuyuki, On dimension estimates with surrogate data sets, IEICE transactions on fundamentals of electronics, communications and computer sciences 80 (1997-05-25) 859–868.

[13] J. H. Lucio, R. Valdes, L. R. Rodriguez, Improvements to surrogate data methods for nonstationary time series, Physical review. E, Statistical, nonlinear, and soft matter physics 85 (2012) 056202–1–19.

[14] T. Schreiber, A. Schmitz, Improved surrogate data for nonlinearity tests, Phys. Rev. Lett. 77 (1996) 635–638.

[15] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, H. H. Liu, The empirical mode decomposition and the

Hilbert spectrum for nonlinear and non-stationary time series analysis, Royal Society of London Proceedings Series A 454 (1998) 903–995.

[16] N. Tsakalozos, K. Drakakis, S. Rickard, A formal study of the non-linearity and consistency of the empirical mode decomposition, Signal Processing 92 (2012) 1961 – 1969.

[17] R. H. Shumway, D. S. Stoffer, Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics), Springer, 2^a edition, 2006.

[18] W. S. Cleveland, Visualizing Data, Hobart Press, Summit, New Jersey, U.S.A., 1993.

[19] R. A. Rios, R. F. Mello, Improving time series modeling by decomposing and analyzing stochastic and deterministic influences, Signal Processing 93 (2013) 3001–3013.

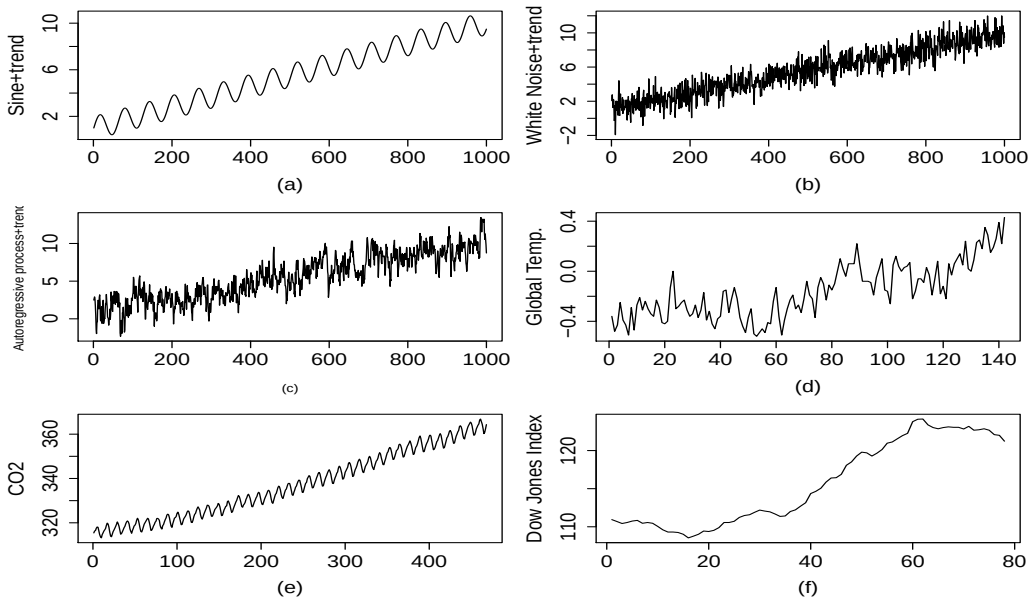

Figure 1: Time series used to evaluate the proposed methods: (a) A sine function combined with a trend; (b) A white noise process combined with a trend; (c) Autoregressive process combined with a trend; (d) Global Temperature [17]; (e) Atmospheric concentrations of CO2 [18]; (f) Dow Jones Utilities Index [8].

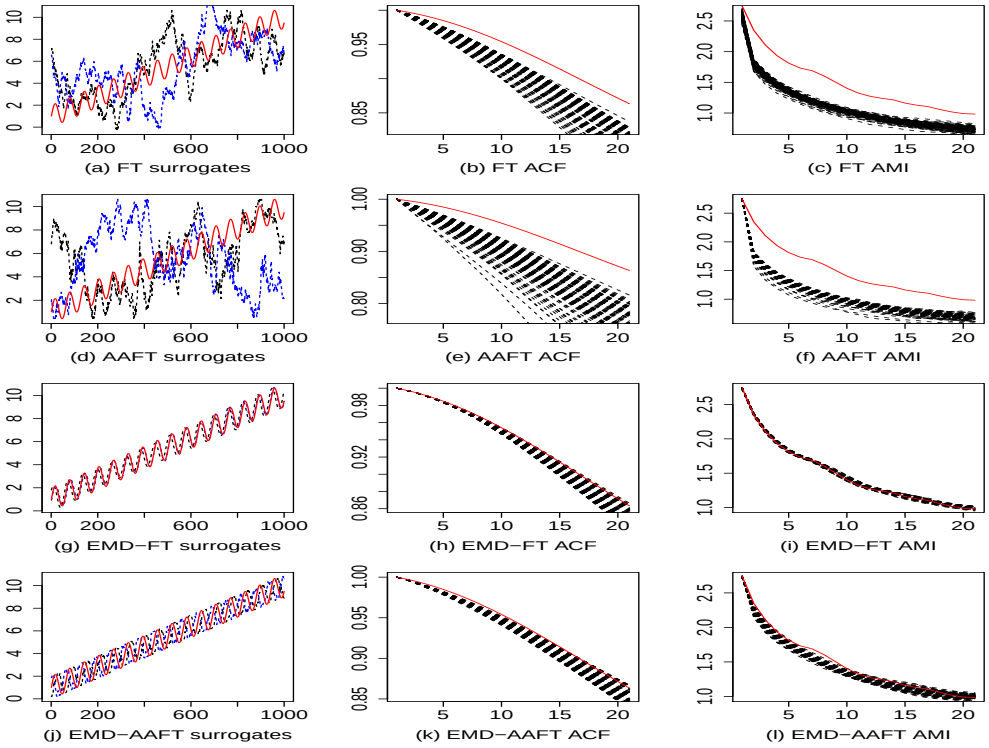

Figure 2: Surrogates generated from the synthetic time series created by the combination of trend and sine function. At the left side, the original time series (red continuous line) and its surrogates (dashed lines) are presented. In the middle and right side, the ACF and AMI plots are shown.

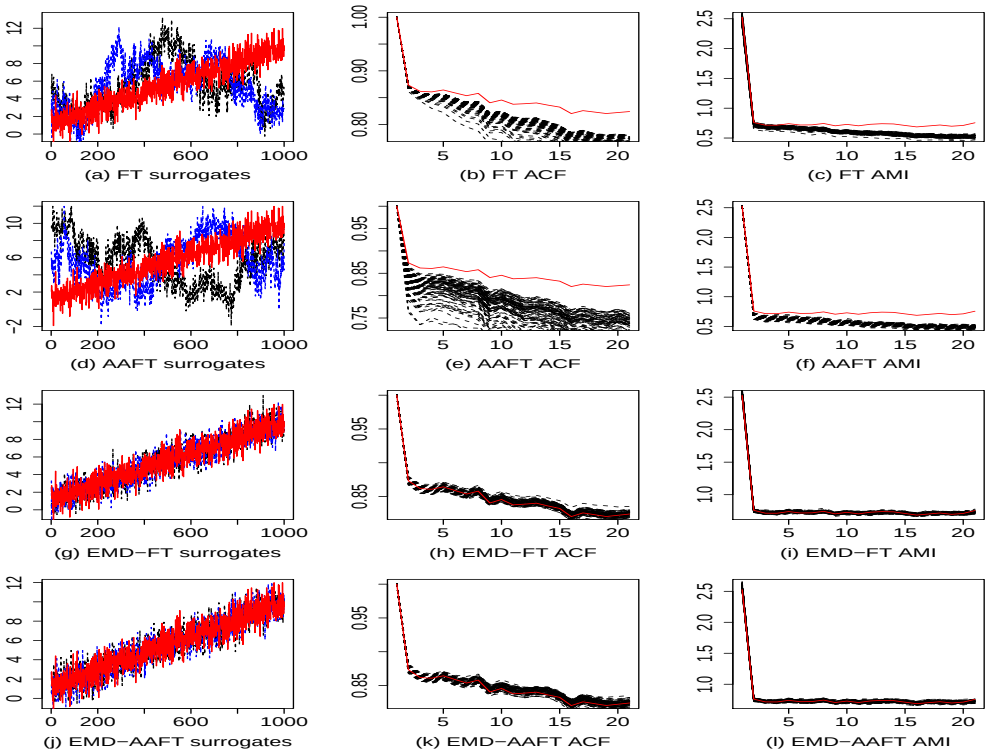

Figure 3: Surrogates generated from the synthetic time series created by the combination of trend and white noise process. At the left side, the original time series (red continuous line) and its surrogates (dashed lines) are presented. In the middle and right side, the ACF and AMI plots are shown.

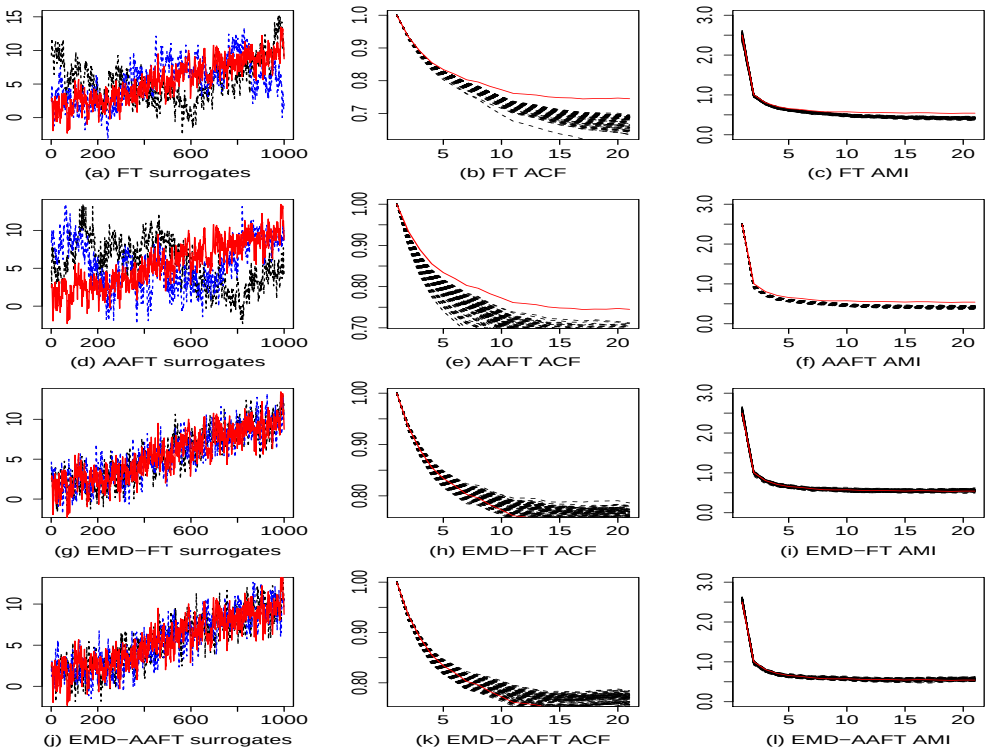

Figure 4: Surrogates generated from the synthetic time series created by an autoregressive noise. At the left side, the original time series (red continuous line) and its surrogates (dashed lines) are presented. In the middle and right side, the ACF and AMI plots are shown.

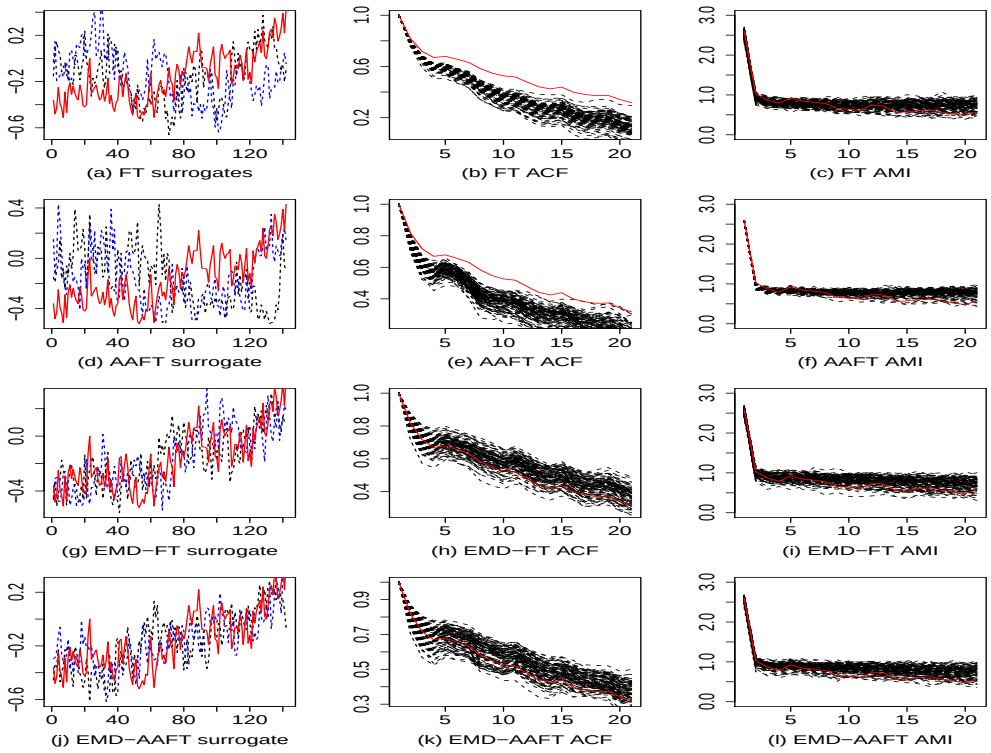

Figure 5: Surrogates generated from the real-world time series composed of average values, yearly, collected of global temperatures. At the left side, the original time series (red continuous line) and its surrogates (dashed lines) are presented. In the middle and right side, the ACF and AMI plots are shown.

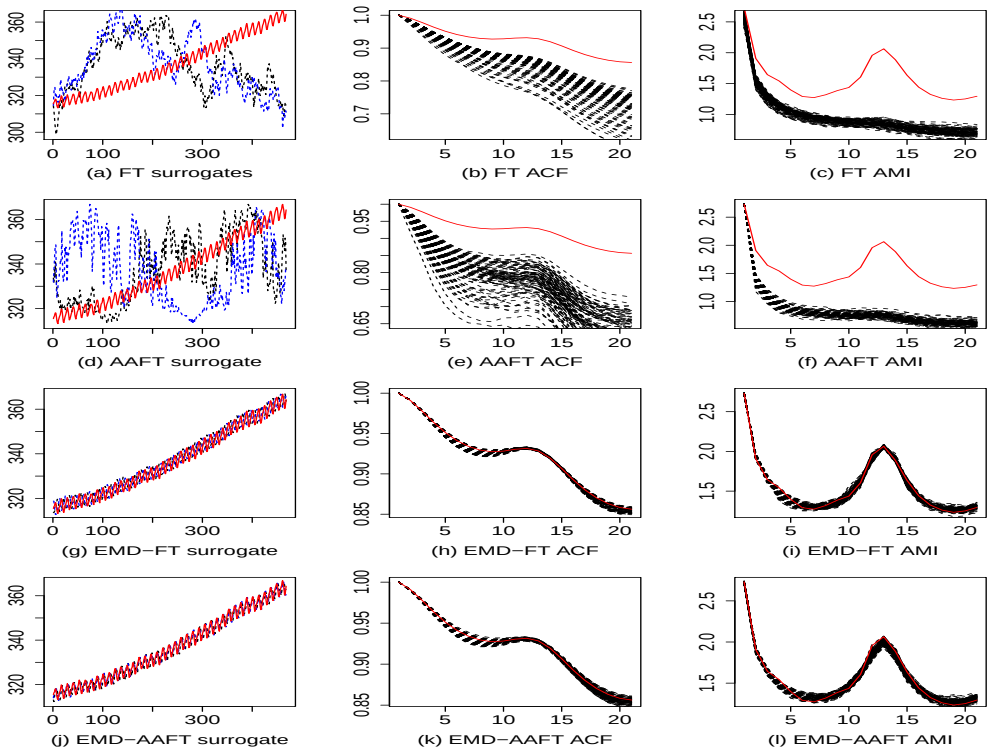

Figure 6: Surrogates generated from the real-world time series composed of atmospheric concentrations of CO2. At the left side, the original time series (red continuous line) and its surrogates (dashed lines) are presented. In the middle and right side, the ACF and AMI plots are shown.

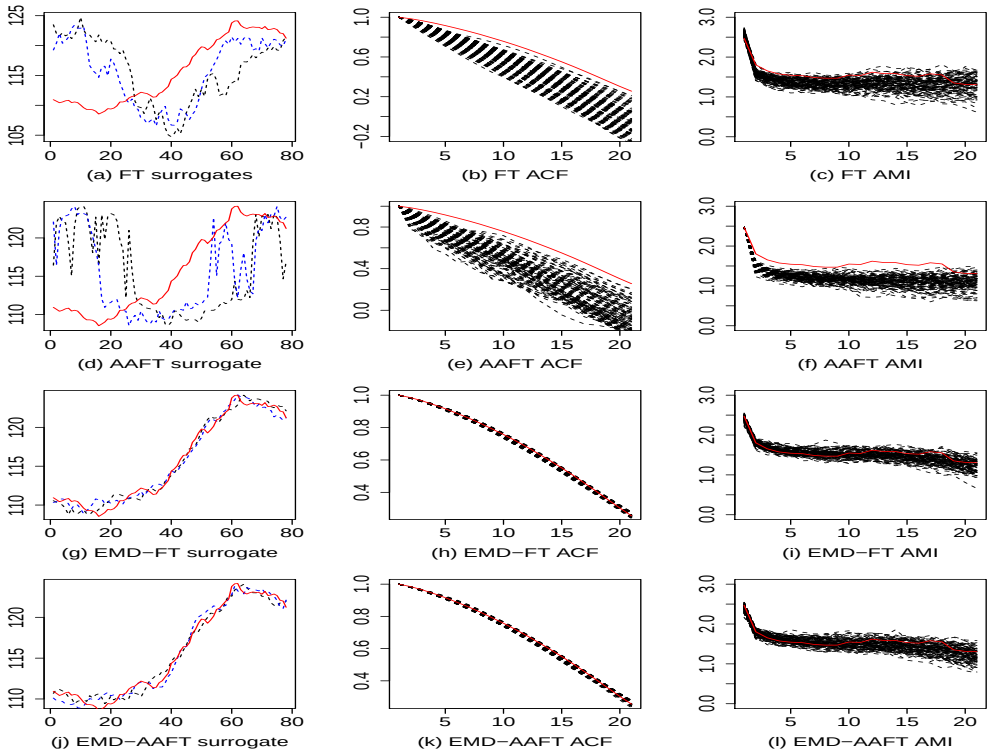

Figure 7: Surrogates generated from the real-world time series of Dow Jones Utilities Index. At the left side, the original time series (red continuous line) and its surrogates (dashed lines) are presented. In the middle and right side, the ACF and AMI plots are shown.

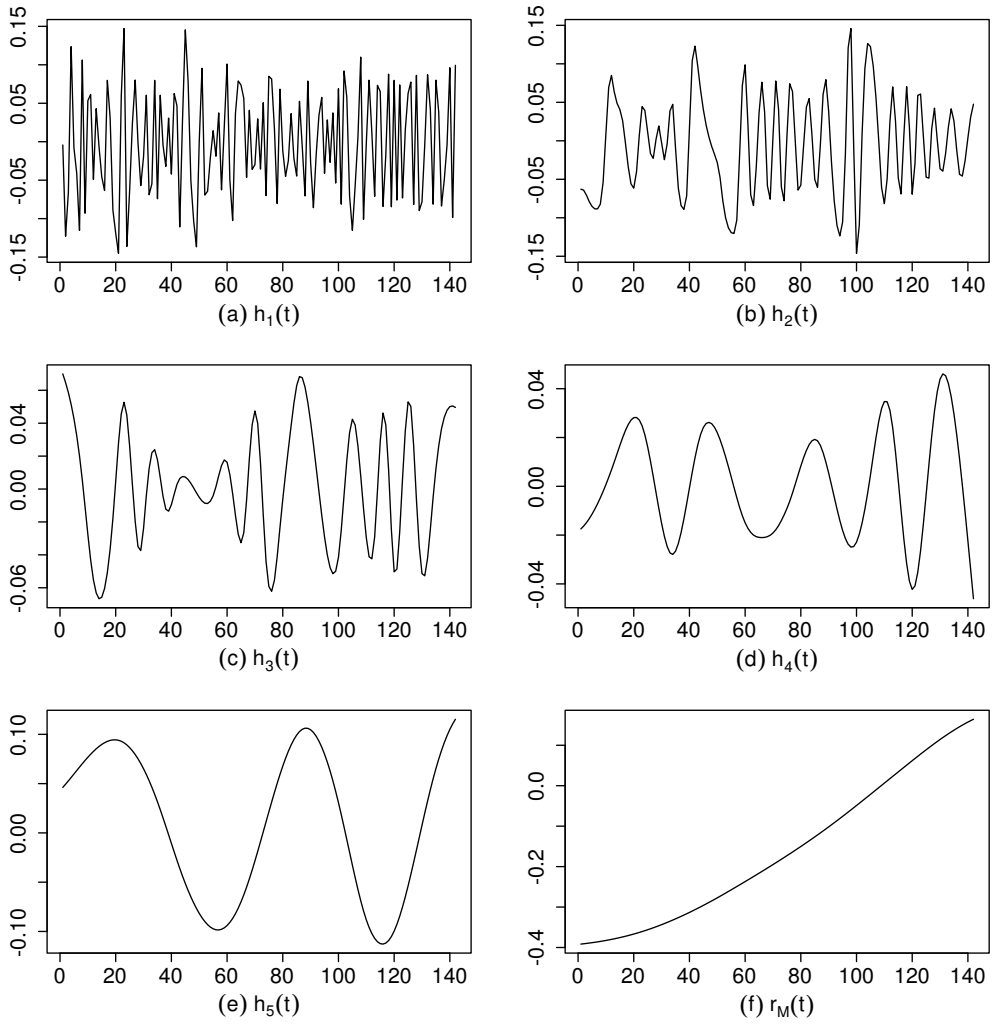

Figure 8: The plots from (a) to (f) show the IMFs and the residue extracted from the time series presented in Figure 1.d.

Table 1: Hypothesis test using Average Mutual Information (AMI): FT and AAFT rejected the null hypothesis in 9 out of 12 scenarios, confirming the surrogate series produced are not representative enough; EMD-FT and EMD-AAFT accepted the null hypothesis in all scenarios, confirming they produce more significant surrogate data.

Time Series	FT	AAFT	EMD-FT	EMD-AAFT
Sine + Trend	0 (R)	0 (R)	0.753 (A)	0.013 (A)
White noise + Trend	0 (R)	0 (R)	0.601 (A)	0.508 (A)
AR(1) + Trend	0 (R)	0 (R)	0.565 (A)	0.571 (A)
Global Temperature	0.727 (A)	0.942 (A)	0.829 (A)	0.886 (A)
CO2	0 (R)	0 (R)	0.483 (A)	0.412 (A)
Dow Jones	0.073 (A)	0 (R)	0.344 (A)	0.321 (A)