International Journal of Information Technology & Decision Making
Vol. 12, No. 3 (2013) 491–521

DOI: 10.1142/S0219622013500193

AN INTEGRATED DATA CHARACTERISTIC TESTING SCHEME FOR COMPLEX TIME SERIES DATA EXPLORATION

LING TANG

School of Economics and Management
Beijing University of Chemical Technology
Beijing 100029, China

Centre for Energy Chemical Management
Beijing University of Chemical Technology
Beijing 100029, China
tangling@mail.buct.edu.cn

LEAN YU*

School of Economics and Management
Beijing University of Chemical Technology
Beijing 100029, China

Centre for Energy Chemical Management
Beijing University of Chemical Technology
Beijing 100029, China

Alibaba Business College
Hangzhou Normal University
Hangzhou 310036, China

College of Information Science and Technology
Beijing University of Chemical Technology
Beijing 100029, China
yulean@mail.buct.edu.cn

FANGTAO LIU

Department of Management Science
City University of Hong Kong
Hong Kong, China
msftliu@city.edu.hk

WEIXUAN XU

Institute of Policy and Management
Chinese Academy of Sciences
Beijing 100190, China
wxu@casipm.ac.cn

*Corresponding author.

In this paper, an integrated data characteristic testing scheme is proposed for complex time series data exploration so as to select the most appropriate research methodology for complex time series modeling. Based on relationships across different data characteristics, data characteristics of time series data are divided into two main categories: nature characteristics and pattern characteristics in this paper. Accordingly, two relevant tasks, nature determination and pattern measurement, are involved in the proposed testing scheme. In nature determination, dynamics system generating the time series data is analyzed via nonstationarity, nonlinearity and complexity tests. In pattern measurement, the characteristics of cyclicity (and seasonality), mutability (or saltation) and randomicity (or noise pattern) are measured in terms of pattern importance. For illustration purpose, four main Chinese economic time series data are used as testing targets, and the data characteristics hidden in these time series data are thoroughly explored by using the proposed integrated testing scheme. Empirical results reveal that the natures of all sample data demonstrate complexity in the phase of nature determination, and in the meantime the main pattern of each time series is captured based on the pattern importance, indicating that the proposed scheme can be used as an effective data characteristic testing tool for complex time series data exploration from a comprehensive perspective.

Keywords: Data characteristic; complex time series; integrated testing scheme; data exploration.

1991 Mathematics Subject Classification: 22E46, 53C35, 57S20

1. Introduction

The increasing use of time series data has initiated numerous studies within various fields,[1] involving forecasting,[2] similarity measurement,[3] classification,[4] programming[5] and other tasks. Undoubtedly, these researching models are closely related to and even based on data characteristics of the observed time series data. For example, conventional statistical methods can fit stationary data well, but may lead to spurious regression in the case of data with nonstationarity characteristic.[6] Similarly, nonlinear models, e.g., threshold time series models,[7] are implemented under the assumption that the samples are driven by nonlinear dynamics. Due to the complexity of time series data, many artificial intelligence (AI) tools and hybrid approaches are proposed instead of traditional techniques.[8] When time series data demonstrates strong seasonality, various seasonal methods, such as seasonal adjustment, seasonal decomposition and seasonal variables introduction, can improve the performance of data analysis effectively.[9] Therefore, data characteristic exploration is a necessary step in the time series data analysis.

There are abundant studies on data characteristic exploration for time series data. Usually, data characteristics of time series are mainly referred to stationarity (and nonstationarity),[6] linearity (and nonlinearity),[10] chaoticity,[11] complexity,[12] fractality,[13] memorability,[14] regularity (and irregularity),[15] cyclicity,[16] seasonality,[9] saltation (or sudden change),[17] randomicity (or noise pattern),[18] and others. Accordingly, various algorithms are proposed to explore and test each data characteristic. For example, in stationarity (and nonstationarity) testing, since the seminal work on stationarity testing proposed by Dickey and Fuller[19] and by Nelson and Plosser[6] who argued that traditional techniques would lose power in the case of nonstationary data was presented, stationarity testing has received much attention from both theoretical and empirical perspectives. As for linearity and nonlinearity, numerous nonlinear tests have been proposed since the first nonlinear statistical test was proposed by Hurst[10] in 1951. As excellent examples of nonlinear dynamics, other complex systems theories, such as chaotic system theory,[20] complex system theory[21] and fractal system theory,[13] have been developed in recent decades. Accordingly, various statistics, e.g., fractal dimensions,[22] entropies,[12] Lyapunov exponent,[23] are computed to evaluate characteristics of chaotic property, complexity, fractality, memorability, regularity (and irregularity) and randomicity in time series data. Similarly, a variety of seasonal paradigms, such as HEGY-type methods,[24] are widely used to test seasonality characteristic. For the characteristic of sudden change (or saltation), Chow test,[25] CUSUM-type test[26] and other similar tests attempt to find abrupt breaks with structural changes hidden in time series data in terms of level or tendency. In the case of randomicity, noise level is often evaluated by various forms, such as signal-to-noise ratio (SNR).[18]

However, most of these above studies focused only on one or several data characteristics. In the existing literature, there is no integrated solution to all data characteristics exploration. Furthermore, it has been proved that there are a number of coexisting data characteristics[27] in time series data. In such a situation, even though some testing approaches for single data characteristic can work out whether the time series possesses such characteristics, it is not clear which characteristic is the dominant characteristic in the time series. This often leads to the fact that the users are difficult to select a suitable analysis and forecasting tool for further processing. For this reason, an integrated data characteristic testing scheme is required to solve the issue.

The main contribution of this paper is to formulate an integrated data characteristic testing scheme to explore the data characteristics hidden in time series data, especially in complex time series data. The remaining part of this paper is organized as follows: Sec. 2 describes the proposed integrated data characteristic testing scheme in detail. For illustration, an empirical analysis using four Chinese economic time series data as testing target is performed via the proposed testing scheme in Sec. 3. Section 4 concludes this paper and discusses the future research directions.

2. Methodology Formulation

In this section, the overall formulation process of the integrated data characteristic testing scheme for complex time series data exploration is presented. In the process, data characteristics are first discussed in Sec. 2.1. Then, an integrated data characteristic testing scheme is formulated in terms of the identified data characteristics. Accordingly, its testing methods and main steps are given in Sec. 2.2.

2.1. *Data characteristics*

Generally speaking, data characteristics are mainly referred to a kind of special properties of data itself. Typical data characteristics of time series include stationarity (and nonstationarity), linearity (and nonlinearity), chaoticity, complexity, fractality, memorability, regularity (and irregularity), cyclicity, seasonality, saltation (or sudden change), randomicity and others.

As the most basic characteristic of time series data, stationarity (or nonstationarity) is closely related to data generation process (DGP) without (or with) unit roots. In particular, stationarity or nonstationarity can be defined as the following form: if there is no unit root in DGP, the time series is termed as stationary and it can converge with time; otherwise, the time series is nonstationary, and at least one unit root exists in the data generation process where the main statistical natures are inclined to change as time goes.[6,19]

The dynamics of data generation process can be further divided into two main categories: linear system and nonlinear system. If time series is driven by such systems which are difficult to govern the dynamics using simple linear equations, it is assumed that there are nonlinear characteristics in the time series. In particular, if nonlinear tools can be statistically proved to be more effective than linear ones in modeling,[28] or if the residuals from a linear fit did not follow an independent identical distribution as assumed,[29] the time series data are commonly considered to be nonlinear.

As excellent examples of nonlinear dynamics, other complex systems theories, such as chaotic system, complex system and fractal system, have been developed in recent decades. Chaotic characteristic, i.e., Lorenz attractor, was first found by Lorenz[20] in 1963, which was well known as "Butterfly Effect". Similarly, Li and Yorke[11] defined such characteristic as the behavior that seems superficially irregular but obeys certain hidden rules, just sensitive to the initial states. Actually, a chaotic process is an intermediate between regular state and irregular state.

The term complexity or complex system provides another description for chaotic theory.[21] It is commonly accepted that a higher level of complexity indicates a more complicated time series with more irregular and less regular components.

Fractal theory as a geometry term was proposed by Mandelbrot[13] and it is used to describe the similarity between components and the whole, i.e., self-similarity. The fractal characteristic exploration for time series started with the work of Packard *et al.*[30] and they proposed phase-space reconstruction theory in their study. Based on fractal theory, various fractal dimensions, such as correlation dimension and information dimension, have been formulated to describe the dimensional complexity of time series data.[31]

Focusing on memorability, Hurst[10] gave the first nonlinear statistic-Hurst exponent. Accordingly, a long-memory process exploration was already involved as scale-unchanged characteristic via rescaled rand (R/S) analysis.[10] Particularly, characteristic of memorability reveals the power of the hidden rules governing the time series with long memory. That is, if there is a strong characteristic of memorability, the time series data is likely to obey the driven laws and demonstrate regularity.[14]

From above analysis, it can be easily concluded that the characteristics of stationarity (and nonstationarity), linearity (and nonlinearity), chaoticity, complexity, fractality, memorability and regularity (and irregularity) are strongly dependent on the dynamical system generating time series data. These characteristics mainly show the intrinsic natures of time series data. Concretely speaking, stationarity and nonstationarity characteristics reveal the nature of data generation process of time series, which is described by unit roots and time-varying process. Linearity and nonlinearity reflect the relationship among various factors (or patterns) within the dynamics systems of time series data. In addition, chaoticity, complexity, fractality, memorability, regularity and irregularity are referred to the complex state of the dynamics systems driving time series in terms of the governing power of hidden rules. These characteristics can be considered as nature characteristics, which are closely related to the essential property of the dynamics systems generating time series data.

It is worthy of noticing that though different complex systems theories, e.g., chaotic system theory, complex system theory and fractal system theory, were developed to discover respective characteristics of time series, they have in common to be able to analyze the complex state of time series dynamics effectively, where regular (also termed determinate) and irregular (also termed random or stochastic) patterns take places simultaneously. For example, a chaotic characteristic often displays an intermediate of time series between regular and irregular relationship. Complexity can be considered as even more vivid description of chaotic characteristic. Similarly, self-similarity in fractal systems implies hidden regular parts existing within time series dynamics, and fractal dimensions are referred to its complexity state. Memorability characteristic reveals the power of the hidden rules governing the time series in terms of long memory. Furthermore, existing testing approaches for these above characteristics are quite similar and even mixed with each other, which are further discussed in Sec. 2.2.1. Therefore, the characteristics of chaotic property, complexity, fractality, memorability, regularity, irregularity and other similar ones can fall into one category to explore the complexity state of dynamics driving time series, which is termed complexity characteristic in this study.

Traditionally, time series data comprises of a series of coexisting components, i.e., trend, cyclical, seasonal, mutable (sudden change) and noise patterns[32]:

$$y_t = f(t_t, c_t, s_t, m_t, n_t), \tag{2.1}$$

where t_t, c_t, s_t, m_t and n_t denote the trend, cyclical, seasonal, mutable and noise patterns hidden in time series data y_t at time t. Accordingly, the characteristics of cyclicity, seasonality, mutability (or saltation) and randomicity display the respective impacts of their related patterns on the observed time series, in terms of pattern importance.

Cyclical patterns, which return to the beginning and repeat itself in the same sequence with peaks and troughs, are important components hidden in time series data. It can show the main governing rules, e.g., the business cycles in economic time series.[16] It is worth noticing that seasonal pattern can be seen as a special case of cycle pattern within one year, which is mainly caused by environmental factors (e.g., temperature, weather and climate) and socioeconomic ones (e.g., calendar and festival),[33] which attracts even more attentions in extant literatures. Accordingly, characteristics of cyclicity and seasonality measure the influence of cyclical and seasonal patterns on original time series data, respectively.

Mutable patterns are generally related to structural changes in time series data, including level (or intercept) and slope of tendency.[17] Accordingly, whether a structural break exists and how it impacts time series can be expressed by saltation (or mutability) characteristic. Besides, time series data is always corrupted by noise to some different degrees. Thus randomicity characteristic is used to measure how much noise has polluted the time series. These issues are extremely crucial in the field of data analysis and signal processing, especially in data quality estimating and noise reduction techniques.[18]

Based on above discussions, it is not difficult to conclude the main relationships across data characteristics of time series data, as shown in Fig. 1. Generally, data characteristics of time series can be divided into two main categories, i.e., nature and pattern characteristics. The nature characteristics are closely related to the intrinsic feature of dynamics systems driving time series data. They include stationarity and nonstationarity, which reveal the data generation processes whether to remain or vary as time goes; linearity and nonlinearity, which display relationship between various factors or components within dynamics linear or nonlinear; and complexity, which reflects the complex states of dynamics systems in terms of the governing power of inner laws. On the other hand, time series data usually involves diverse hidden patterns at the same time series. They are cyclical, seasonal, mutable and noise components. Accordingly the impacts of these hidden patterns on the original time series can be estimated by characteristics of cyclicity, seasonality, mutability (or saltation) and randomicity, respectively.

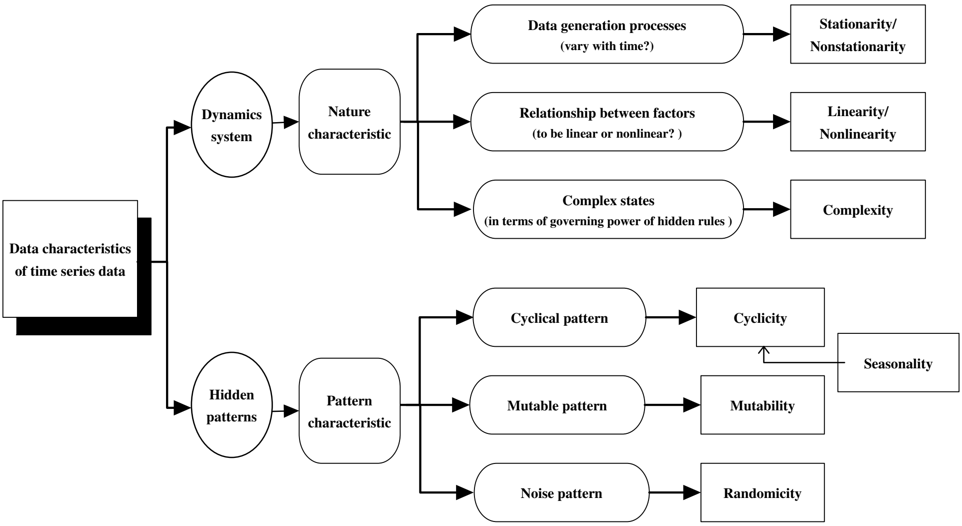

Fig. 1. Data characteristics of time series data.

It is worth noticing that though nature and pattern characteristics are different from each other and explore time series data from different angles of view, they are closely related to and even dependent on each other. On the one hand, nature characteristics take the time series data as a whole to determine inner structural rules in its dynamics system, while pattern characteristics are referred to hidden components and analyze time series data from a specific perspective. On the other hand, since the dynamics system comprises of the hidden patterns, nature characteristics are actually dependent on the features of all patterns together with their relationships. Similarly, patterns constitute time series data in such a way determined by nature characteristics.

2.2. *An integrated data characteristic testing scheme*

Based on relationships across data characteristics, an integrated data characteristic testing scheme is proposed for complex time series data exploration in this study. Figure 2 illustrated the main framework of the proposed integrated testing scheme. Accordingly, two main steps are involved, i.e., nature determination and pattern measurement. In the first step, unit root test, nonlinear test and complexity test are respectively performed to determine nature characteristics of the time series data. In the second step, characteristics of cyclicity (and seasonality), mutability and

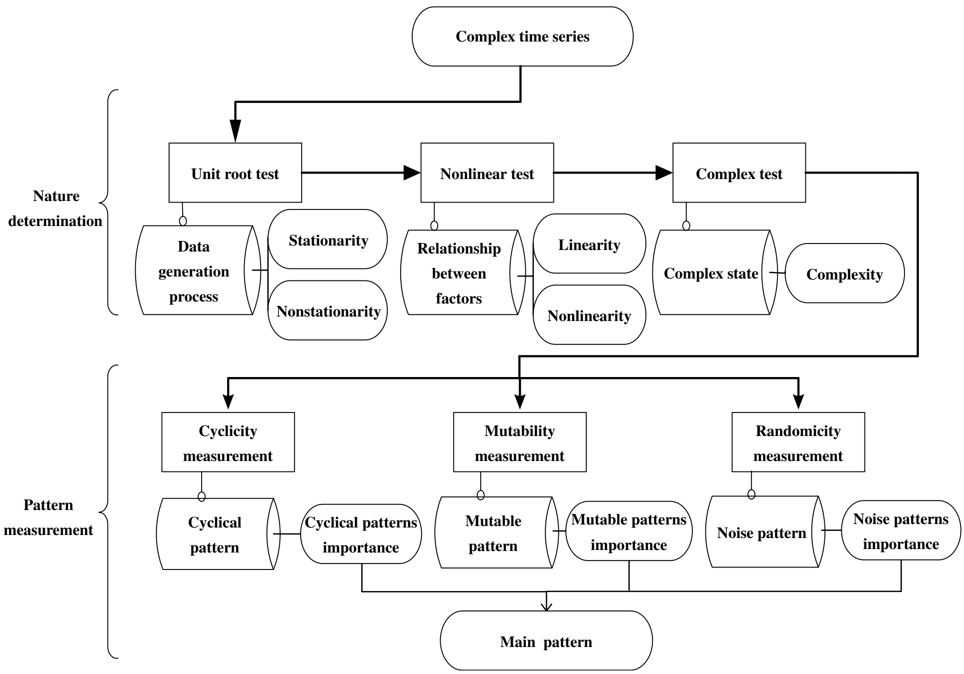

Fig. 2. Framework of integrated data characteristic testing scheme for complex time series data exploration.

randomicity are tested not only to make sure whether all patterns, i.e., cyclical (and seasonal), mutable and noise patterns, exist in time series dynamics, but also to measure the relative importance of different patterns and determine the main pattern characteristic hidden in the observed time series data.

Basically, the generic reasons for formulating the proposed integrated testing scheme reflect the following three-fold. First of all, since nature characteristics reveal the intrinsic features of the data generating dynamics and pattern characteristics show the hidden components, the proposed scheme will first explore data properties from the general point of view in terms of nature characteristics and then further analyze hidden patterns from a specific perspective. Second, amongst the nature characteristics, stationarity (or nonstationarity) may be the most essential characteristic which should not be ignored in time series studies, thus unit root test should be performed first in the proposed scheme, and then nonlinearity and complexity are considered. Finally, in pattern measurement, seeing that time series data always possess coexisting patterns simultaneously, diverse pattern characteristics, i.e., cyclicity, mutability and randomicity, will be estimated in a parallel way, which is different from the nature characteristic determination in a sequential manner.

For further illustration, the following two subsections discuss existing testing technologies for each data characteristic, which can be employed in the proposed testing scheme. Furthermore, the main procedure of the proposed integrated data characteristic testing scheme is described in Sec. 2.2.3.

2.2.1. *Nature determination*

As previously mentioned, nature characteristics include stationarity (and nonstationarity), linearity (and nonlinearity) and complexity. Accordingly, unit root test, nonlinear test and complexity test are respectively implemented to explore the corresponding data characteristics and various existing relevant tests for each characteristic will be described in this subsection.

(1) Unit root test

As one of the most essential characteristics, testing for stationarity and nonstationarity has attracted much attention from both theoretical and empirical perspectives since the seminal work of Dickey–Fuller (DF) test proposed by Dickey and Fuller[19] was published. In these studies, Nelson and Plosser[6] pointed out that traditional techniques would fail in the case of nonstationary data and thus the data should be tested before modeling and see whether the nonlinearity exists. Generally speaking, data generation process and testing method are two main factors in formulating the unit root testing. Accordingly, with different data generation process functions and the related testing models, various stationary test methods can be formulated, amongst which DF test,[19] Augmented DF (ADF) test,[34] Phillips–Perron (PP) test,[35] KPSS test,[36] DF-GLS test[37] and NP test[38] may be the most conventional and popular methods.

DF test[19] can be seen as the original form of unit root test and other tests are usually extended from DF test. In DF test, the data generation process of time series $y_t (t = 1, \ldots, N)$ is described by simple first-order autoregressive model, i.e., AR(1) process, as shown in the following form.

$$y_t = u + \rho y_{t-1} + \varepsilon_t, \tag{2.2}$$

where $y_0 = 0$, u is a constant and $\{\varepsilon_t\}$ is a sequence of independent normal random variables with mean zero and variance σ^2 (i.e., $\varepsilon_t \sim i.i.d(0, \sigma^2)$). It is assumed that the time series y_t can converge to a stationary time series if $|\rho| < 1$, which is termed as stationary; when $|\rho| = 1$, the time series is defined nonstationary with an unit root and its variance is $t\sigma^2$ at time t; and if $|\rho| > 1$, the time series demonstrates strongly nonstationary with an exponentially growing variance as t increases. Based on this data generation process, estimator $\widehat{\rho}$ for ρ are estimated via ordinary least squares (OLS) regression and the corresponding t-statistic or F-statistic are computed to test the null hypothesis $\rho = 1$ with the alternative hypothesis $\rho < 1$ via one-tailed tests.[19]

By improving data generation process, some researchers could extend DF test to formulate many novel stationary tests. For example, Dickey and Fuller[34] extended data generation model to pth-order autoregressive process, i.e., AR(p) and accordingly t-value will be modified. This approach is called as augmented DF (ADF) test. Similarly, other autoregressive models have been also introduced, such as ARIMA(0, 1, q) and ARIMA(p, 1, q),[39] to formulate some new testing statistics to measure the stationarity of time series data. In addition, an important stationary test approach, KPSS test, assuming that time series data follow level stationary or trend stationary process, is given to reverse the null hypothesis and the opposing hypothesis.[36]

Recently, increasing nonstationary tests hold the data generation process in the contest of structural changes, seasonal patterns, nonlinear dynamics and stochastic processes. In a pioneering work by Perron,[40] structural changes, i.e., breaks in intercept and/or slope, are considered in data generation process of unit root tests.[41] A similar concern arose in the seasonal unit root from the work by Hylleberg *et al.*,[24] and accordingly seasonal dummies are added into the data generation process functions for stationarity testing.[42] According to nonlinear system theory, diverse nonlinear models, such as smooth transition autoregressive (STAR), self-exciting threshold autoregressive (SETAR) and threshold autoregressive (TAR) models,[7] are also applied to describe data generation process in stationary tests.[43] Furthermore, unlike autoregressive models with fixed coefficients, stochastic unit root processes are also tested.[44]

Meantime, testing method is another important factor in stationary tests formulation and abundant researches attempted to propose some novel tests by utilizing different testing methods or modifying the original tests. Generally speaking, the testing method in stationary tests can be further partitioned into two sub-factors, i.e., estimation criteria and testing statistics. Focusing on the criterion used to estimate the parameters on deterministic variables, i.e., the method of de-trending, the DF and ADF tests use ordinary least squares (OLS), generalized least square

(GLS),[37] first-difference (FD) de-trending,[45] maximum likelihood,[46] the weighted symmetric estimator[47] and other similar estimating approaches. As far as testing statistics, t-value, F-value, Lagrange multiplier (LM) statistic, likelihood ratio statistic, Wald statistic and others are usually utilized in stationary tests.[48] Furthermore, Phillips and Perron proposed PP test which implements DF test nonparametrically under a more general assumption of accepting dependency of an error term on time and heteroscedasticity of variance by modifying the t-statistic.[35]

(2) Nonlinear test

Nonlinear testing studies started with the work by Hurst[10] who proposed the first nonlinear statistic. Some nonlinear tests were developed in frequency domain such as bispectral density analysis.[49] As mentioned above, when nonlinear system theory was incorporated into unit roots tests, nonlinear unit roots were considered.[43] Besides the two types, there are three other major types of nonlinear testing techniques in much literature, i.e., surrogate data method, residuals of linear fit method and parametric method. These methods have different testing principles, which will be elaborated below.

(a) Surrogate data based method

An important way of testing nonlinearity characteristic is to employ surrogate data. Motivated by statistical hypothesis testing, this kind of techniques presents an indirect way to detect nonlinearity. The null hypothesis of the tests is that the observed time series data is generated by a linear dynamics. In surrogate data-based method, the first step is to generate surrogate data both with linear characteristic as the null hypothesis assumes and with the main properties of the original data. Second, by comparing the original and surrogate data in terms of some given discriminative statistics, it is possible to decide whether a linear process is sufficient for describing the observed time series. Therefore, such kind of nonlinear tests involve two main steps, i.e., surrogate data generation and data comparison in terms of testing statistics.

As far as surrogate data generation, there existed many approaches, each of which aims to generate surrogate data not only retaining dynamic properties of the original data but also subject to the null hypothesis of linearity. The most popular methods are related to the use of the Fourier transform (FT), which was applied in the first surrogate data-based nonlinear test proposed by Theiler *et al.*[50] In Theiler's model, the linear properties are specified by the squared amplitudes of the (discrete) FT $S_k (k = 0, \ldots, N-1)$ from the original data $y_t (t = 0, \ldots, N-1)$, i.e., the periodogram estimator of the power spectrum:

$$|S_k|^2 = \left| \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} y_t \mathrm{e}^{j2\pi kt/N} \right| \quad (k = 0, \ldots, N-1). \tag{2.3}$$

Surrogate time series data $\bar{y}_t (t = 0, \ldots, N-1)$ are created by multiplying the FT of the data by random phases $\phi_k \in [-\pi, \pi]$, i.e., $\mathrm{e}^{j\phi_k}|S_k|$, and then transforming back to the time domain data:

$$\bar{y}_t = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} (\mathrm{e}^{j\phi_k}|S_k|)\mathrm{e}^{-j2\pi kt/N} \quad (t = 0, \ldots, N-1). \tag{2.4}$$

Besides, other algorithms, such as simply order randomizing,[51] fast FT (FFT),[52] amplitude adjustment FT (AAFT),[53] wavelet transform,[54] statistically transformed autoregressive process (STAP)[55] and hidden Markov models (HMM),[54] can be also introduced into nonlinear tests to generate surrogate data.

According to the null hypothesis of the surrogate data-based method, if the observed data does not have nonlinear characteristic but linear, its estimated properties are almost the same as its surrogates. In particular, the properties can be quantified in terms of various discriminative statistics. Generally speaking, the statistic needs not be designed specifically for detecting nonlinearity; and any numerical measure which can be drawn from time series data could be selected as the discriminative statistic. For example, some traditional statistics can be employed, such as autocorrelations[55] and distributions via Kolmogorov−Smirnoff (KS) test,[54] while other indexes estimating complexity of time series have also been introduced as discriminative statistics, e.g., Lyapunov exponents, fractal dimensions and entropies.[56]

(b) Residuals of linear fit-based method

Another important type of nonlinear tests is by using residuals of linear fit. The null hypothesis in this kind of tests is that the observed data is linear and its residuals of linear regression are independent. Based on such principle, residuals of linear fit are analyzed to test nonlinear characteristic of the original data. Some excellent studies of residuals-based tests can refer to Keenan[57] and Tsay[58] who argued if the observed data are generated by nonlinear dynamics, the residuals $\widehat{\varepsilon}_t = y_t - \widehat{y}_t$ of a linear fit, e.g., AR(p) regression $\widehat{y}_t = u + \sum_{i=1}^{p} a_i y_{t-i}$, are related to a proxy variable of the nonlinear behavior in the time series, e.g., $x_t = \widehat{y}_t^2 - \widehat{y}_t$, and the test is implemented based on the regression of the residuals $\widehat{\varepsilon}_t$ on the proxy variables x_t by using F-statistic; Brock *et al.*[59] who proposed BDS test to explore the independence of residuals based on phase-space reconstruction; Pena and Rodriguez,[29] where the residuals of linear fit are modeled and the appropriate order of the autoregressive process is determined via certain model selection criteria, such as Bayesian information criterion (BIC) and Akaike's information criterion (AIC) criteria, and if it is obtained zero, the residuals are assumed independent and the original data is of linearity; and McLeod and Li[60] and Peña and Rodríguez[61] who attempted to formulate testing statistics using residuals, and test the proposed statistics whether follow given distributions which they should be under the hypothesis of linearity of the observed time series data.

(c) Parametric method

Parametric method as the third nonlinear testing approach is implemented under the assumption that nonlinear models can perform much more effectively than linear ones in dynamic systems when the observed time series data is of nonlinear characteristic. In parametrical nonlinear tests, various linear models (such as AR(p) model) and nonlinear models (such as exponential autoregressive (ExpAR), smooth transition autoregressive (STAR), logistic STAR, exponential STAR (ESTAR) and threshold type models) are employed to model the observed time series, respectively. Then the goodness of fit is statistically compared across the linear and nonlinear models in terms of information criteria, such as residual variance and AIC, in order to verify if nonlinear models can effectively improve modeling performance by tailed test.[28,62]

(3) Complexity test

Complexity characteristic of time series data reflects the complex state between regular and irregular relationships, and a lower level of complexity indicates a stronger governing power of the inner rules over the observed time series data. Complexity tests for time series data are usually based on phase-space reconstruction theory by Packard *et al.*[30] An m embedding dimensional phase-space is constructed from the original time series $\{y_t\}(t = 1, \ldots, N)$ with a time delay or lag τ:

$$X_n = \{y_n, y_{n+\tau}, \ldots, y_{n+(m-1)\tau}\} \quad (n = 1, 2, \ldots, N - (m-1)\tau). \tag{2.5}$$

Given a threshold distance r, the correlation integral $C(m, r)$ can be defined as:

$$C(m, r) = \frac{2}{[N - (m-1)\tau][N - (m-1)\tau - 1]} \sum_{i<j} H(r - \|X_i - X_j\|), \tag{2.6}$$

where $H(z)$ is Heaviside function, and $H(z) = 0$ if $z \leq 0$, while $H(z) = 1$ if $z > 0$. $\|X_i - X_j\|$ is the distance between the two vectors.

As discussed above, complexity characteristic is based on characteristics of chaotic property, fractality, memorability, regularity (and irregularity) and other similar properties in different complex systems theories. Actually, existing testing approaches for these above characteristics are quite similar and usually mixed with each other, with one same aim at estimating the governing power of inner regular laws over the observed time series data. Related approaches refer to the use of fractal dimensions, entropies, Lyapunov exponent and recurrence plots.

Fractal dimensions are the main statistics in fractal system theory, which is widely used to describe the dimensional complexity of time series data.[31] For example, correlation dimension,[22] the most popular fractal dimension, can be defined as:

$$D_m = \lim_{r \to 0} \left[\frac{\ln(C(m, r))}{\ln(r)} \right]. \tag{2.7}$$

In practice, correlation dimension can be gained by regressing the curve of $\ln(r) - \ln(c(m, r))$ via an ordinary least square estimation. Besides, a simple way to obtain the fractal dimension D is based on Hurst exponent H, i.e., $D = 2 - H$.[63]

Even though Hurst exponent is the first nonlinear statistic, it can be still considered as an effective tool for fractal analysis and its R/S analysis can effectively estimate the governing power of regularity with long memory.[10] Thus, both Hurst exponent and R/S analysis have been proved to be useful for identifying complexity, fractal structure and chaotic property of time series data.[64]

Likewise, diverse entropies have been constructed to evaluate the complexity state of time series data. Entropy is actually a thermodynamic quantity describing disorder in dynamics system. A large value of entropy commonly refers to a high level of complexity of time series data. Entropy estimators can be generally classified into two categories, i.e., spectral entropies and embedding entropies. Usually the spectral entropies focus on amplitude components of power spectrum of time series data as the probabilities via Fourier transformation (FT), power spectral density and other similar measurements,[65] while embedding entropies are estimated based on phase-space reconstruction theory. Embedding entropies are determined by finding points on the trajectory that are closer together in phase-space but occurring at different time points, including approximate entropy, sample entropy, fuzzy entropy and Kolmogorov−Sinai (KS) entropy.[12,66] Typically, the KS entropy can be defined as $K = \lim_{r\to 0}\lim_{m\to\infty} \frac{1}{\tau}\frac{C(m,r)}{C(m+1,r)}$.[67]

Lyapunov exponent λ_i which quantifies the rate of attractor expansion or contraction (i.e., the growth rate of subspaces in phase-space), is another important statistic in complexity determination.[23] Specifically, λ_1 presents the average rate at which linear distances grow between two points on the attractor, and such two points with an initial distance δ may have their separation $\delta e^{\lambda_1 t}$ at time t. The summation $(\lambda_1 + \lambda_2)$ relates the growth rate in two dimensional space, and similarly the average growth rate in d dimensional areas is calculated by the sum of the first d exponents $\Lambda_d = \sum_{i=1}^{d} \lambda_i$. In terms of this rate, the existence of positive exponents reveals uncertainty due to the exponential divergence of points or sensitive dependence on initial conditions, while the negative ones are associated with attractor controlling and subspaces contract. Thus, if the cumulative exponent $\Lambda < 0$, the attractor is globally stable; $\Lambda > 0$ is related to be globally unbound; and $\Lambda = 0$ indicates a balance between expansive and contractive dynamics.

Besides various statistics, some other useful tools can be also utilized to analyze the complexity of time series data, such as diagrams of recurrence plots,[68] power spectral density,[69] and others, which efficiently visualize complexity of time series data. For example, recurrence plots diagram are drawn by the matrix $R_{i,j}(r) = H(r - \|X_i - X_j\|)$.

2.2.2. *Pattern measurement*

In the previous subsection, all characteristics are related to nature of dynamics system generating time series data, while pattern characteristics, i.e., cyclicity, seasonality, mutability and randomicity display impact of their related patterns (i.e., cyclical, seasonal, mutable, and noise patterns) on original time series data. Thus, the measurement of pattern characteristics should not only test presence of these patterns but also extract the main ones through evaluating pattern importance. Accordingly, both testing and modeling approaches for these pattern characteristics are analyzed.

(1) Cyclicity measurement
Cyclicity is the most important hidden pattern hidden in time series data. Amongst various cyclical patterns, seasonal pattern can be considered as a special case with one-year time scale caused mainly by environmental and socioeconomic factors.

To discover the main time scale of cyclical pattern in time series data, some useful exploration tools are used. For example, power spectral analysis aims to extract dominant cyclicity from the observed data through time-frequency transition, such as Fourier transfer.[70] Similarly, autocorrelation analysis can also help to determine existence of a hidden cyclical pattern, by showing how one point relates to the others in the interval of a given time scale.[71] Besides, some complex system techniques, such as recurrence plots[72] and R/S analysis,[73] have also been utilized to discover cyclictiy and seasonality in time series data.

Testing for cyclicity or seasonality prefers to introduce the relative pattern variables into original models and then test their significance, including determinate pattern tests and pattern unit root tests.[24,74] In determinate pattern tests, the null hypothesis is generally designed that the main frequency of time series data $w^2 = 0$ with the alternative $w^2 \neq 0$. Accordingly some typical studies may be referred to Canova–Hansen (CH) test,[75] Caner test[76] and Tam–Reinsel test.[77] In the pattern unit root tests, they assumes cyclical or seasonal patterns are varied with time. Thus it is necessary to extend cyclicity or seasonality tests into nonstationary tests. Similarly some typical techniques include Dickey–Hasza–Fuller test,[78] HEGY test[24] and others. Taking quarterly time series data for example, the seasonal process is formed as the sum of a purely deterministic component u_t and a stochastic process z_t in HEGY type tests:

$$y_t = u_t + z_t, \quad \alpha(L)z_t = v_t, \tag{2.8}$$

where $\alpha(L) \equiv 1 - \sum_{i=1}^{4} \alpha_i^* L^i$ is an AR(4) in the conventional lag operator L. The disturbance of $\{v_t\}$ has a zero mean, and the long-run variance is bounded at both zero frequency, $w_0 \equiv 0$, and at the seasonal spectral frequencies, $w_k \equiv 2\pi k/4(k = 1, 2)$. The seasonal unit roots are in $\alpha(L)$; and the null hypothesis is $\alpha(L) = 1 - L^4 \equiv \Delta_4$.

To model cyclicity or seasonality, the standard approaches include two main types, i.e., regression with pattern dummies and extraction by decomposition. The first-type models checked the observed time series with variables of possible cyclicity or seasonality via deterministic function. For example, since peaks and troughs are usually involved, trigonometric forms, such as sine curve, can be employed to describe cyclical and seasonal patterns.[79] Besides, Auto-Regressive Integrated

Moving Average (ARIMA) model can be extended to Seasonal ARIMA (SARIMA) with seasonal variables[80]:

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)e_t, \tag{2.9}$$

where p, d, q, P, D, Q are integers, and s is the season length. $\phi_p(B)$ is the autoregressive operator of order p, $\theta_q(B)$ is a moving average operator of order q, and $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are polynomials in B^s of order P and Q. B denotes the backward shift operator, i.e., $Bx_t = x_{t-1}$, $B^2x_t = x_{t-2}$ and so on, d is the number of regular difference, and D is the order of seasonal differences.

The other type of modeling attempts to find cyclical or seasonal patterns from time series by decomposing original data into a set of major modes including the relative patterns, and the popular decomposition tools mainly include smoothing-based techniques (e.g., exponential smoothing), moving average, filters and others.[81,82] Take seasonality for example, the most commonly used method is the X12 type methods proposed by Census Bureau.[9] Based on series of moving average, X12 type methods can decompose the time series data y_t into three main components, namely trend-cycle component tc_t, seasonal factor sf_t and irregular component ir_t. Multiple forms can be used to combine these components into the original data. Typically, the most widely used forms are the additive and multiplicative forms, i.e., $y_t = tc_t + sf_t + ir_t$ and $y_t = tc_t \times sf_t \times ir_t$, respectively. Besides, the Census-X12-ARIMA program, the latest X12 type model, provides some tests for the significance of seasonality and the stability across years of seasonal factors.

(2) Mutability measurement

Mutability is related to whether and to what extent any mutable break (or structural change) exists in time series data. According to existing studies, there are two main types for structural changes testing.

The first type is based on model estimation with the null hypothesis that if structural change is not present, then the parameters estimating for global time series data are close to the values for at least one subset of the sample. With such assumption, a most popular example may be Chow test[25] where a known single break point is tested by using F-test:

$$F = \frac{\frac{\mathrm{SSE}-\mathrm{SSE}_1-\mathrm{SSE}_2}{m+1}}{\frac{\mathrm{SSE}_1+\mathrm{SSE}_2}{N_1+N_2-2m-2}} \sim F(m+1, N_1+N_2-2m-2), \tag{2.10}$$

where SSE is the sum of residual square errors in modeling the whole time series data with $N_1 + N_2 - m - 1$ degrees of freedom, and SSE_1 and SSE_2 are those for the data subsets before and after the break point with degrees of freedom $N_1 - m - 1$ and $N_2 - m - 1$, respectively. Since the work by Chow, diverse extensions have been proposed. For example, Quandt[83] augmented Chow test to each possible point. Furthermore, the statistics of largest Wald, Lagrange multiplier and likelihood ratio-like, Bayesian information criterion, Schwarz' criterion and others can also be introduced to test numbers of possible breaks.[84]

Another main kind of mutability tests is the CUSUM type approach first proposed by Brown *et al.*[85] based on cumulated sum of recursive residuals. An important extension of CUSUM is the iterative cumulative sums of squares (ICSS) test.[26] In ICSS algorithm, the time series is supposed to have a stationary unconditional variance σ_1^2, over an initial time period until a sudden change takes place at the time point K_1. The new unconditional variance σ_2^2 is stationary until the next breakpoint K_2, and this process repeats. To test a sudden break, a cumulative sum of square residuals is calculated by

$$C_k = \sum_{t=1}^{k} c_t^2 \quad (k = 1, 2, \ldots, N), \tag{2.11}$$

where c_t is uncorrelated random variables with zero mean and unconditional variance, computed by

$$c_t = \frac{y_t - \left(\sum_{i=0}^{t-1} y_i / t\right)}{\sqrt{\left(\frac{t+1}{t}\right) S_Y^2}} \quad (t = 1, 2, \ldots, N), \tag{2.12}$$

where S_Y^2 is the variance of total samples of time series. The ICSS statistic can be defined as:

$$D_k = \frac{C_k}{C_N} - \frac{k}{N} \quad (k = 1, 2, \ldots, N), \quad D_0 = D_N = 0. \tag{2.13}$$

If there are no breakpoint along the time series data, D_k will oscillate around zero and otherwise departure from zero.

In modeling mutable patterns, introducing dummy variables of structural change is a popular way. For example,

$$\Delta y_t = u + \beta t + \theta DU_t(t_b) + \gamma DT_t(t_b) + \alpha y_{t-1} + \varepsilon_t, \tag{2.14}$$

where DU_t and DT_t are the dummy variables of structural break in the level and slope occurring as t_b, respectively, i.e., $DU_t = 1$ if $t > t_b$, 0 otherwise; $DT_t = t - t_b$ if $t > t_b$, 0 otherwise.[86]

(3) Randomicity measurement

Randomicity characteristic is related to how much noise is blended into the observed time series. Actually, noise measurement is a crucial task both in data analysis and in digital signal processing, as the premise of other further researches such as data quality estimation and noise reduction. Noise pattern is popularly measured by signal-to-noise ratio (SNR), which declines monotonously with increasing noise level. SNR can be defined as the ratio between signal of interest and the noise pattern in terms of energy (or variance), amplitude, density (or root mean values) and others, via different ratio forms, such as simple ratio, logarithmic ratio and lower better form.[87,88] In addition, other measurement forms, e.g., noise variance, noise standard deviation and noise ratio in total data, increase with a high level of noise.[89,90] These approaches of noise estimation shed light on pattern importance measuring, which is further discussed in Sec. 2.2.3.

In modeling noise, it is generally assumed that time series comprise of patterns of interest s_t and the noise pattern n_t, i.e., $y_t = s_t + n_t$. Thus, noise pattern can be captured by effectively abstracting various patterns of interest s_t from data. Accordingly different regression models, smoothing methods, spectral analysis and filters can be utilized to model the patterns of interest s_t, and the prediction error approximates the noise pattern.[91] Meantime, direct estimation of the noise pattern in terms of variance, standard deviation, ratio or other statistics is another important way to model noise, e.g., the studies by Gao[92] and Cellucci *et al.*[93] where noise is estimated based on computing logarithmic displacement or time-dependent exponent curves in the reconstructed phase space.

2.2.3. *Main procedure*

According to methodological framework and existing relative tests mentioned in the previous subsections, a main procedure of the proposed integrated data characteristic testing scheme together with recommended testing techniques can be constructed, as shown in Fig. 3.

As can be seen from Fig. 3, the main procedure of the proposed integrated data characteristic testing scheme consisted of two main parts: nature determination and pattern measurement, as elaborated in the following description.

(1) Nature determination
In nature determination, unit root test, nonlinear test and complex test are respectively performed to determine nature characteristics of the dynamics system generating time series data.

(a) Unit root test
In the time series data, stationarity is the most essential data characteristic. Thus unit root tests should first be implemented, aiming to test whether the time series data are stationary or nonstationary in the data generation process of the observed time series. Based on different data generation process and the related testing methods, diverse unit root tests can be formulated with their distinct assumptions. Amongst existing unit root tests, DF, ADF and PP tests with a general data generation assumption of AR(1) or AR(p) process are the most popularly used stationary tests, and most of recently proposed ones are extended from them by incorporating other data characteristics into the original form, such as nonlinearity, seasonality and mutability. Therefore, the above traditional tests are strongly recommended as useful tools to explore stationarity or nonstationarity hidden in time series data.

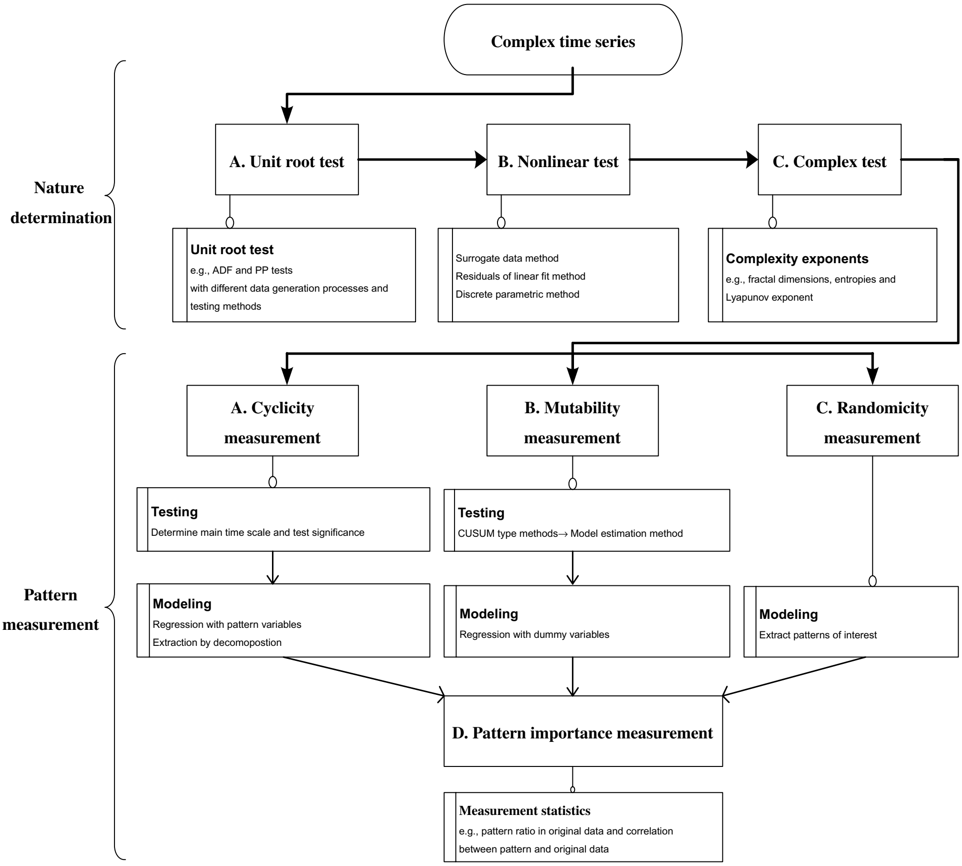

Fig. 3. Main procedure of the proposed integrated data characteristic testing scheme for complex time series data exploration.

(b) Nonlinear test

Nonlinear tests are mainly performed to determine whether the observed time series are generated by nonlinear or linear dynamics. The most popularly used tests can be generally divided into three main methods in terms of different principles, i.e., surrogate data-based method, residuals of linear fit-based method and parametric method. Usually, the surrogate data-based method explored nonlinearity in an indirect way by using surrogate data, while the other two methods test nonlinearity by directly modeling linear or nonlinear parts in time series data. Therefore, different kinds of nonlinear tests are advised to perform simultaneously in order to analyze nonlinearity of time series data from a more comprehensive perspective.

(c) Complexity test

The complexity of time series data reflects the complex state between regular and irregular relationships. In this study, complexity is defined in terms of characteristics in various theories of complex, chaotic or fractal systems, such as chaoticity, fractality, memorability, regularity (or irregularity) and others. Therefore, the related exponents corresponding to the above characteristics, such as fractal dimensions, entropies and Lyapunov exponents, can be employed to measure the complexity of time series data in terms of governing power of regular rules. Besides, some other tools are also useful for visualizing the complexity of time series, e.g., diagrams of recurrence plots, power spectral density and others.

(2) Pattern measurement

After determining nature characteristics, pattern characteristics further display the hidden patterns of time series data, including cyclical, seasonal, mutable and noise patterns. In pattern measurement, not only the verification of the presence of related patterns is necessary, but also pattern importance should be measured in order to determine the most important pattern hidden in the observed time series data. Thus, two sub-steps are involved in each pattern measurement, i.e., pattern testing and modeling and pattern importance determination.

(a) Cyclicity testing and modeling

Exploring the main time scale is a crucial task in cyclicity (or seasonality) test. In the cyclicity testing, the most useful tools can use power spectral analysis, autocorrelation analysis, recurrence plots chart, R/S analysis and others for testing.

The standard testing for cyclicity (or seasonality) comprises of two main types, i.e., regression with pattern variables and extraction by decomposition. The former type introduces pattern variables to capture the possible cyclicity or seasonality, the latter attempts to abstract cyclical or seasonal patterns from original time series by decomposing data into a set of components.

(b) Mutability testing and modeling

Mutability testing aims to test whether any structurally mutable break exists in the observed time series data. Generally speaking, two popular methods, the model estimation methods and the CUSUM method, can be used. Usually, the former approaches often focus on given possible breaks, while the latter tends to test each point one by one. In the CUSUM method, it is commended to perform first to select suspected points and these breaks are then tested by model estimation methods.

Similarly, introducing structural change dummies into regression equation is a widely used way to model mutable patterns. Furthermore, structural changes in terms of level and tendency should be both considered.

(c) Randomicity testing and modeling

Time series data are always polluted by noise pattern. Noise testing can be implemented by extracting the patterns of interest from the original data or directly estimating the impact of noise pattern on data. For consistence with other patterns, extracting the patterns of interest is strongly recommended in the proposed testing scheme. Accordingly the residuals of regression on main tendency and patterns of interest will approximate the noise pattern.

(d) Pattern importance determination
After testing, patterns importance should be measured to estimate the impact of each pattern on the original data and finally to determine the most important pattern. For such purpose, measurement statistics of pattern importance should be given. According to existing pattern measurements, especially for seasonality and noise, pattern ratio in total data can be proposed as a useful description to the governing power of the pattern over the time series, in terms of various features of time series data, such as power (i.e., variance), amplitude (i.e., the difference between the maximum and minimum) and density (i.e., the root mean values). Besides, the correlation between the pattern series and original time series data can be also utilized as a measurement statistic.

3. Empirical Results

For illustration, four main Chinese economic time series data are selected as testing targets and their data characteristics will be explored by using the proposed integrated data characteristic testing scheme. In this section, a brief description about the sample data will be first given in Sec. 3.1, and then Sec. 3.2 reports the detailed testing steps and the corresponding experimental results.

3.1. *Data descriptions*

For illustration, the proposed integrated data characteristic testing scheme is implemented to test the data characteristics of main Chinese economic time series data. Specially, time series data of gross domestic product (GDP), total social consumption (TSC), consumer price index (CPI) and energy production (ENE) are selected as testing samples, which are obtained from Wind database (http://www.wind.com.cn/), as shown in Table 1 and Fig. 4.

3.2. *Data characteristic testing*

Based on the main procedure as described in Sec. 2.2.3, the data characteristics of the four main Chinese economic time series data are tested by the proposed integrated data characteristic testing scheme.

Table 1. Descriptions of testing samples.

Time series	Frequency	Starting date	Ending date	Size	Unit
GDP	Quarterly	Mar-92	Jun-11	78	Hundred million Yuan (Rmb)
TSC	Monthly	Jan-95	Jul-11	199	Hundred million Yuan (Rmb)
UER	Monthly	Jan-90	Jul-11	259	Dollar/Yuan (Rmb)
ENE	Monthly	Jan-95	Feb-11	182	Ten thousand ton (ce)

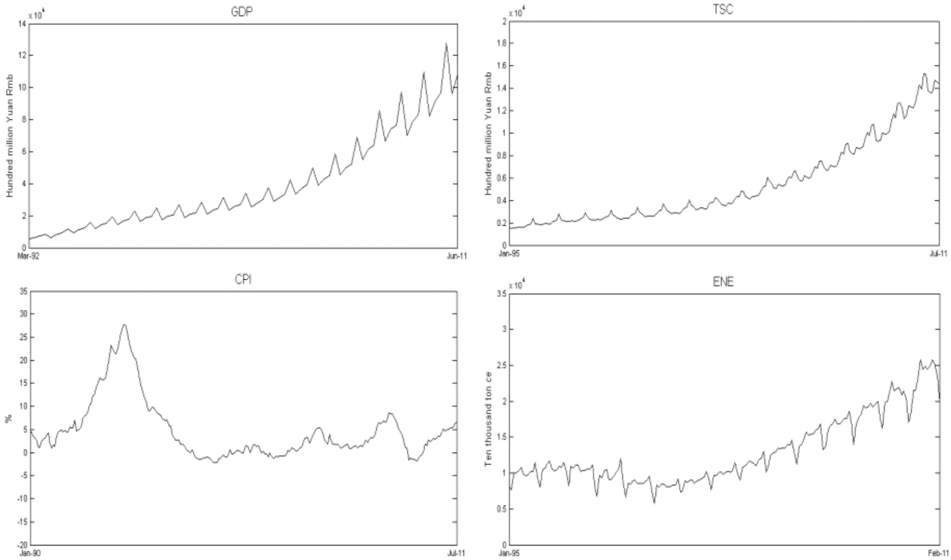

Fig. 4. Four main Chinese economic time series data.

3.2.1. *Nature determination*

(1) Determination of Stationarity and Nonstationarity

For stationarity and nonstationarity, the most popular unit root test, ADF test,[34] is utilized in this study. Accordingly the computational results are reported in Table 2. In this testing, the data generation functions whether with or without constants or trends are selected based on Schwarz information criterion. From the reported results, it is easy to see that at the confidence level of 95%, the existence of unit roots cannot be refused in the four time series data, and thus all sampling data are proved to be nonstationary time series.

(2) Determination of Nonlinearity and Linearity

As for nonlinear analysis, surrogate data-based method and residuals of linear fit method are employed in this study. In the surrogate data-based method, Fourier transform (FT) are employed. Using the FT approach, the surrogating data can first

Table 2. Testing results for stationarity and nonstationarity characteristics in terms of ADF test.

	ADF test				
Time series	Equation (Constant, Trend, Lages)	t-Stat.	Critical values	Prob.	Stationarity
GDP	(1,1,5)	1.9637	−3.1640	1.0000	×
STC	(1,1,12)	4.3574	−3.1410	1.0000	×
CPI	(0,0,12)	−1.2012	−1.6158	0.2101	×
ENE	(1,1,12)	0.1028	−3.1424	0.9971	×

Table 3. Testing results for nonlinear characteristic in terms of surrogate data-based method.

Time series	h	KS-Stat.	Prob.	Linearity
GDP	0.9840	0.3439	0.0091	×
STC	0.9980	0.3352	0.0007	×
CPI	1.0000	0.4460	0.0000	×
ENE	0.9820	0.3005	0.0055	×

be generated by multiplying the FT of the data in random phases and then transforming back to the time domain.[50] Subsequently, the distributions of the surrogate data and the original data are compared via Kolmogorov−Smirnoff (KS) tests[54] to determine the nonlinearity hidden in the time series data. According to above steps, a total of 1000 surrogate data with different random phases are generated. Table 3 reports the corresponding results, where h, KS-Stat. and Prob. represent the ratio of times of rejecting the null hypothesis in all simulations at the confidence levels of 95%, the mean calculated KS statistics and the mean p-value, respectively. From the table, it can be seen that h are all above 95%, and the mean p-values are far below 5%, indicating that all the observed time series data are statistically proved to have the characteristic of nonlinearity with the confidence level of 95%.

In the residuals of fit method, the observed data are first model via linear AR(p) models, in which the order p is determined via AIC criterion. Then, BDS tests are implemented to test the independence of the residuals, where the dimension m is set to 2−5, and the parameter r is set to 0.7 times of the variance of data in phase-space reconstruction.[59] Accordingly, Table 4 presents the related testing results. Similar conclusion can be drawn that all p-values are far smaller than 5%, indicating that the residuals of linear fit cannot be proved to be independent as the null hypothesis assumes, and thus the observed time series data is demonstrated to possess nonlinear characteristic at the confidence level of 95%.

(3) Determination of Complexity

In complex analysis, complexity exponent is built based on the most popular fractal dimension, i.e., correlation dimension, which is estimated by using G-P algorithm in this study.[22] In phase-space reconstruction, the lag τ is determined via

Table 4. Tests results for nonlinear characteristic in terms of residuals of fit method.

	m-dimensional space								
	2		3		4		5		
Time series	Stat.	Prob.	Stat.	Prob.	Stat.	Prob.	Stat.	Prob.	Linearity
GDP	0.1132	0.0000	0.1934	0.0000	0.2398	0.0000	0.2716	0.0000	×
STC	0.0276	0.0003	0.0971	0.0000	0.1145	0.0000	0.1183	0.0000	×
CPI	0.0187	0.0006	0.0315	0.0003	0.0424	0.0000	0.0477	0.0000	×
ENE	0.0147	0.0000	0.0291	0.0000	0.0432	0.0000	0.0568	0.0000	×

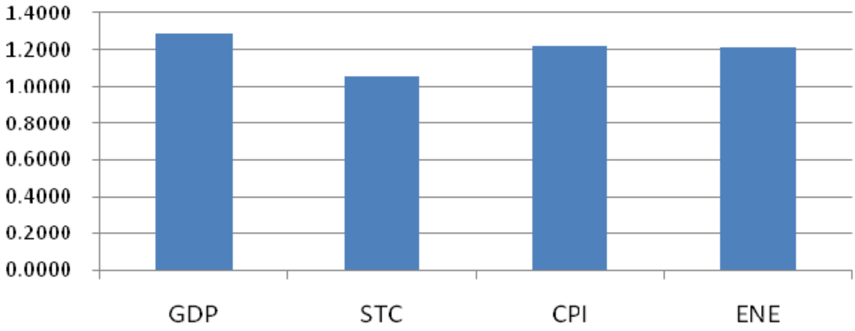

Fig. 5. Testing results for complexity characteristic in terms of correlation dimension.

autocorrelation analysis, and embedding dimension m via the Kennel *et al.*' approach.[94] Figure 5 presents the corresponding testing results. It is not hard to find that all the sample data have the obvious complexity, since all complexity exponents in terms of correlation dimensions are greater than one in the dimension of time series data.

3.2.2. *Pattern measurement*

(1) Testing and Modeling for Cyclicity
Seasonality, as a special case of cyclicity, is first considered due to its importance in macroeconomic time series. Autocorrelation analysis with one-year lag period is implemented to determine whether seasonal pattern exists in the observed time series data.[71] The testing results indicate that all the observed time series data demonstrate to have seasonality because all autocorrelations are larger than 0.5. In detail, the relative numerical results of GDP, TSC, CPI and ENE are 0.796, 0.755, 0.691 and 0.757, respectively.

For cyclical patterns, R/S analysis[10,73] is employed to determine the main time scale in terms of long memory in the time series data. Figure 6 shows the log–log plot of R/S against t of the four observed data series. From the computational results, it can be seen that the main time scales of TSC and ENE are identical about one year, which may be considered as an indicator with strong seasonality. For other two time series data, a cycle of three year (12 quarters or 36 moths) might be present in the data of GDP and CPI.

When modeling seasonality, the Census-X12-ARIMA, a most popular seasonal model based on moving average, is utilized to abstract the seasonal patterns hidden in time series data, where the additive forms are selected. For cyclicity, an exponentially weighted moving average method, the Holt–Winter's method,[95] is employed to model the cyclical patterns of three year in GDP and CPI. The Holt–Winter's method with additive seasonal forms assumes that the time series y_t is determined by local level l_{t-1}, local growth b_{t-1} and local cyclical component s_{t-m}, where m is the time scale of cyclical pattern and set to three years. Thus, extracted series s_t presents the cyclical pattern hidden in original time series data.

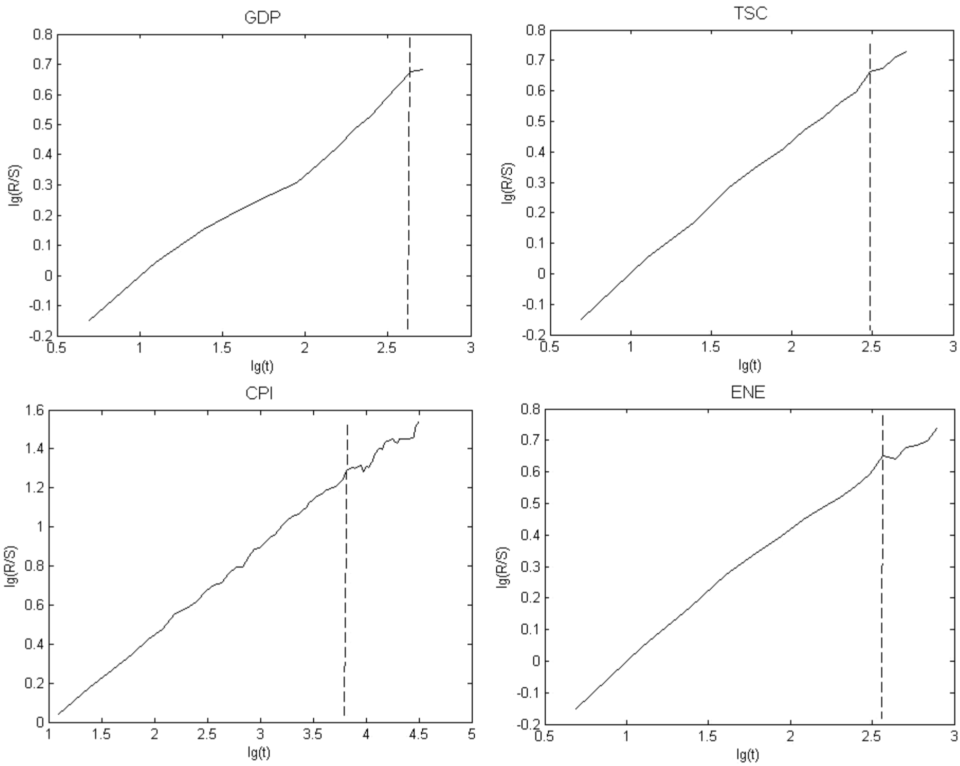

Fig. 6. Testing results of log−log plot of R/S against t.

(2) Testing and Modeling for Mutability

To test mutability, ICSS test,[26] an excellent CUSUM type approach, is first employed to test the residuals of linear ARIMA model based on AIC. The point with the largest ICSS statistic is selected as the possible structural break. Then, the original data is segmented into two subsets before and after the possible break, and the two subsets are modeled respectively. Chow statistics, log likelihood ratio and Wald test are computed to test the similarity between the two subsets.[84] If one point is tested to be break point at confidence level of 95%, the two sub-series are respectively tested for structural change in a similar way, and these steps repeat until no break point can be found with 95% confidence level.

Table 5 reports the testing results of mutable point, where p-values are listed. The reported results show that the data of STC, CPI and ENE are statistically proved to have mutability because at least one test refuses the null hypothesis, supporting that there exist certain structural breaks in the time series at 95% confidence level. For GDP, the computational results indicate that there might not be obvious breaks in the data.

Table 5. Testing results for mutable breaks in terms of ICSS, Chow, log likelihood ratio and Wald statistics.

Time series	Break point	Prob.				Structural change
		ICSS	Chow	Log likelihood ratio	Wald	
GDP	—	—	—	—	—	×
STC	2004.08	0.0359	0.0000	0.0000	0.0000	√
	2007.11	0.0000	0.0000	0.0000	0.0000	√
CPI	1994.09	0.0562	0.0001	0.0000	0.0000	√
	1998.03	0.0728	0.0436	0.0289	0.0136	√
ENE	2001.01	0.1906	0.0000	0.0000	0.0177	√
	2003.11	0.0728	0.0026	0.0006	0.0021	√

(3) Modeling for Randomicity
For consistence purpose, the residuals from modeling above patterns of interest are considered as noise pattern. Particularly, the noise pattern can be calculated in terms of irregularity component abstracted by the Sensus-X12-ARIMA model, residues after extracting cyclical, level and growth components by Holt−Winter's smoothing, and prediction errors from ARIMA models with mutable dummy variables, and their mean values will be used to approximate noise in this study.

(4) Pattern Importance Determination
In this study, variance ratios of pattern in total data are computed as patterns importance to measure the impacts of the related pattern on the original time series. Accordingly the computational results are shown in Table 6.

As can be seen from Table 6, it is easy to find that STC and ENE have strong seasonality in terms of pattern importance measurement and thus seasonal pattern is their main patterns. Meantime, there is an obvious cyclical pattern with three-year duration in the data of CPI and GDP. However, noise pattern seriously polluted the data of CPI. In addition, seasonal pattern and cyclical pattern of three years can be both seen as the main hidden patterns of GDP, which accounting for about 0.44062% and 0.44060%.

3.2.3. *Further discussions*

Through testing via the proposed integrated testing scheme, the main data characteristics of the four main Chinese economic time series data can be obtained, as shown in Table 7.

Table 6. Pattern importance measurement in terms of variance ratios.

	Seasonal pattern	Cyclical pattern	Mutable pattern	Noise pattern
GDP	0.44062%	0.44060%	—	0.01867%
STC	1.07837%	—	0.06590%	0.04713%
CPI	0.04452%	0.18276%	0.16134%	0.39763%
ENE	4.08753%	—	0.81445%	0.66349%

Table 7. Main data characteristic of Chinese economic time series data.

	Nature determination			Pattern measurement			
Time series	Stationarity	Linearity	Complexity	Seasonality	Cyclicity	Mutability	Randomicity
GDP	×	×	1.2958	√	√		
TSC	×	×	1.0571	√			
CPI	×	×	1.2187		√		√
ENE	×	×	1.2124	√			

As can be seen from Table 7, it is easy to find that there are high complex nature characteristics for all four testing time series data because their values of complexity dimension are greater than 1 of the dimension of time series data. The main reason of high complexity may lie in that they are all nonstationary and nonlinear time series. Accordingly, conventional linear techniques might be insufficient for fitting the dynamics of these sample data due to the characteristics of nonstationarity and nonlinearity. In addition, according to the complexity characteristic, AI tools or hybrid approaches are strongly recommended as the promising modeling techniques.

Furthermore, the main patterns hidden in the testing time series data are tested, which shed light on modeling function design. From Table 7, it can be seen that GDP, TSC and ENE demonstrate strong seasonality and a three-year cycle is involved in the data of GDP and CPI in terms of pattern testing and pattern importance measurement. Therefore, when formulating some models for GDP, TSC and ENE, seasonal variables should be introduced. Similarly, cyclical factors should be considered in the case of GDP and CPI. Model adjustment based on main pattern characteristics can effectively improve analysis performance, which will be verified and demonstrated in the future research.

4. Concluding Remarks

The main contribution of this paper is to formulate an integrated data characteristic testing scheme for complex time series data exploration. In the proposed testing scheme, data characteristics can be divided into two main categories, i.e., the nature characteristic and pattern characteristic. Accordingly two related steps, i.e., nature determination and pattern measurement, are involved. In nature determination, unit root tests, nonlinear tests and complex tests are respectively performed to determine nature of time series data. In pattern measurement, characteristics of cyclicity (and seasonality), mutability and randomicity are analyzed not only to make sure whether cyclical (and seasonal), mutable and noise patterns exist in time series data, but also to measure their respective impacts on original data in terms of pattern importance, in order to capture the main pattern hidden in the observed time series data.

For illustration, data characteristics of four main Chinese economic time series data are used to the testing targets via the proposed integrated data characteristic testing scheme. The empirical results show that in nature characteristics, all the time series data are complex since they are demonstrated to be nonstationary and nonlinear, and all their complexity exponents are at high levels. Furthermore, the main patterns hidden in these data are captured and some interesting results can be found. For example, GDP, TSC and ENE demonstrate to have strong seasonality. For GDP and CPI, there exists a three-year cycle within the sampling data. These results also indicate that the proposed integrated testing scheme can be used as an effective tool to explore data characteristics for complex time series data from a thorough and comprehensive perspective.

In addition, the testing results drawn from the proposed data characteristic testing scheme are pretty helpful for selecting and formulating the most appropriate analysis models for the observed time series data. For example, the nature characteristic can shed light on compatible techniques, e.g., traditional economic methods, artificial intelligence tools or hybrid approaches, and accordingly the modeling forms should consider the main pattern hidden in the observed data. We will look into these issues in the near future.

Acknowledgments

Authors would like to express their sincere appreciation to the editors and the three independent reviewers in making valuable comments and suggestions to the paper. Their comments have improved the quality of the paper immensely. This work is partially supported by grants from the National Science Fund for Distinguished Young Scholars (NSFC No. 71025005) and the National Natural Science Foundation of China (NSFC Nos. 90924024, 91224001) and the Fundamental Research Funds for the Central Universities in BUCT (Project No. ZY1320).

References

1. Y. Peng, G. Kou, Y. Shi and Z. X. Chen, A descriptive framework for the field of data mining and knowledge discovery, *International Journal of Information Technology & Decision Making* **7**(4) (2008) 639–682.
2. W. Huang, K. K. Lai, Y. Nakamori and S. Y. Wang, Forecasting foreign exchange rates with artificial neural networks: A review, *International Journal of Information Technology & Decision Making* **3**(1) (2004) 145–165.
3. J. P. Li, X. L. Sun, W. He, L. Tang and W. X. Xu, Modeling dynamic correlations and spillover effects of country risk: Evidence from Russia and Kazakhstan, *International Journal of Information Technology & Decision Making* **8**(4) (2009) 803–818.
4. Y. Shi, Y. Peng, G. Kou and Z. X. Chen, Classifying credit card accounts for business intelligence and decision making: A multiple-criteria quadratic programming approach, *International Journal of Information Technology & Decision Making* **4**(4) (2005) 581–599.
5. Y. Shi, Y. Peng, W. X. Xu and X. W. Tang, Data mining via multiple criteria linear programming: Applications in credit card portfolio management, *International Journal of Information Technology & Decision Making* **1**(1) (2002) 131–151.
6. C. R. Nelson and C. I. Plosser, Trends and random walks in macroeconomics time series: Some evidence and implications, *Journal of Monetary Economics* **10** (1982) 139–162.
7. H. Tong, Birth of the threshold time series model, *Statistica Sinica* **17** (2007) 8–14.

8. L. Tang, L. A. Yu, J. P. Li and S. Y. Wang, A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting, *Applied Energy* **93** (2012) 432–443.
9. S. Wang, L. A. Yu, L. Tang and S. Y. Wang, A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China, *Energy* **36**(11) (2011) 6542–6554.
10. H. E. Hurst, Long-term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers* **116** (1951) 770–779.
11. T. Y. Li and J. Yorke, Period three implies chaos, *American Mathematics Monthly* **82** (1975) 985–992.
12. W. T. Chen, J. Zhuang, W. X. Yu and Z. Z. Wang, Measuring complexity using FuzzyEn, ApEn, and SampEn, *Medical Engineering & Physics* **31** (2009) 61–68.
13. B. B. Mandelbrot, *Fractals: Form, Chance, and Dimension* (W.H. Freeman and Company, San Francisco, CA, 1977).
14. Y. C. Weng, N. B. Chang and T. Y. Lee, Nonlinear time series analysis of ground-level ozone Dynamics in Southern Taiwan, *Journal of Environmental Management* **87** (2008) 405–414.
15. T. Jagrič, M. Marhl, D. Štajer, Š. T. Kocjančič, T. Jagrič, M. Podbregar and M. Perc, Irregularity test for very short electrocardiogram (ECG) signals as a method for predicting a successful defibrillation in patients with ventricular fibrillation, *Translational Research* **149**(3) (2007) 145–151.
16. V. Zarnowitz and A. Ozyildirim, Time series decomposition and measurement of business cycles, trends and growth cycles, *Journal of Monetary Economics* **53** (2006) 1717–1739.
17. J. Jouini and M. Boutahar, Evidence on structural changes in U.S. time series, *Economic Modelling* **22** (2005) 391–422.
18. J. Hu, J. B. Gao and K. D. White, Estimating measurement noise in a time series by exploiting nonstationarity, *Chaos, Solitons and Fractals* **22** (2004) 807–819.
19. D. A. Dickey and W. A. Fuller, Distribution of estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association* **74** (1979) 427–431.
20. E. N. Lorenz, Deterministic non-periodic flow, *Atmospheric Sciences* **20** (1963) 130–141.
21. Q. A. Memon, Synchronized choas for network security, *Computer Communications* **26** (2003) 498–505.
22. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica D* **9** (1983) 189–208.
23. J. P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, *Reviews of Modern Physics* **57** (1985) 617–656.
24. S. Hylleberg, R. F. Engle, C. W. J. Granger and B. S. Yoo, Seasonal integration and cointegration, *Journal of Econometrics* **44** (1990) 215–238.
25. G. C. Chow, Tests of equality between sets of coefficients in two linear regressions, *Econometrica* **28** (1960) 591–605.
26. C. Inclán and G. C. Tiao, Use of cumulative sums of squares for retrospective detection of changes of variance, *Journal of the American Statistical Association* **89**(427) (1994) 913–923.
27. X. Zhang, K. K. Lai and S. Y. Wang, A new approach for crude oil price analysis based on Empirical Mode Decomposition, *Energy Economics* **30**(3) (2008) 905–918.
28. A. Galka and T. Ozaki, Testing for nonlinearity in high-dimensional time, *Physica D* **158** (2001) 32–44.
29. D. Pena and J. Rodriguez, Detecting nonlinearity in time series by model selection criteria, *International Journal of Forecasting* **21** (2005) 731–748.
30. N. H. Packard, J. P. Crutchfield, J. D. Farmer and R. S. Shaw, Geometry from a time series, *Physical Review Letters* **45**(3) (1980) 712–716.

31. S. Janjarasjitt, M. S. Scher and K. A. Loparo, Nonlinear dynamical analysis of the neonatal EEG time series: The relationship between sleep state and complexity, *Clinical Neurophysiology* **119** (2008) 1812–1823.
32. A. D. Silva, R. J. Hyndman and R. Snyder, A multivariate innovations state space Beveridge-Nelson decomposition, *Economic Modeling* **26** (2009) 1067–1074.
33. C. S. Hilas, S. K. Goudos and J. N. Sahalos, Seasonal decomposition and forecasting of telecommunication data: A comparative case study, *Technological Forecasting & Social Change* **73** (2006) 495–509.
34. D. A. Dickey and W. A. Fuller, Likelihood ratio statistics for autoregressive time series with a unitroot, *Econometrica* **49** (1981) 1057–1072.
35. P. C. B. Phillips and P. Perron, Testing for a unit root in time series regression, *Biometrika* **75**(2) (1988) 335–346.
36. D. Kwiatkowski, P. C. B. Phillips, P. Schmidt and Y. Shin, Testing the null of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* **54** (1992) 159–178.
37. G. Elliott, T. Rothenberg and J. H. Stock, Efficient tests for an autoregressive unit root, *Econometrica* **64** (1996) 813–839.
38. P. Perron and S. Ng, Useful modifications to some unit root tests with dependent errors and their local asymptotic properties, *Review of Economics Studies* **63** (1996) 435–463.
39. D. W. Shin, S. Sarkar and J. H. Lee, Unit root tests for time series with outliers, *Statistics & Probability Letters* **30**(3) (1996) 189–197.
40. P. Perron, The great crash, the oil price shock, and the unit root hypothesis, *Econometrica* **57** (1989) 1361–1401.
41. D. Kim and P. Perron, Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses, *Journal of Econometrics* **148**(1) (2009) 1–13.
42. P. M. M. Rodrigues and A. M. R. Taylor, Efficient tests of the seasonal unit root hypothesis, *Journal of Econometrics* **141**(2) (2007) 548–573.
43. J. Lee, M. C. Strazicich and B. C. Yu, LM threshold unit root tests, *Economics Letters* **110**(2) (2011) 113–116.
44. A. Rahman and S. Saadi, Random walk and breaking trend in financial series: An econometric critique of unit root tests, *Review of Financial Economics* **17**(3) (2008) 204–212.
45. P. Schmidt and P. C. B. Phillips, LM tests for a unit root in the presence of deterministic trends, *Oxford Bulletin of Economics and Statistics* **54** (1992) 257–287.
46. D. W. Shin and J. H. Lee, Consistency of the maximum likelihood estimators for nonstationary ARMA regressions with time trends, *Journal of Statistical Planning and Inference* **87** (2000) 55–68.
47. W. A. Fuller, *Introduction to Statistical Time Series*, 2nd edn. (Wiley, New York, 1996).
48. P. M. M. Rodrigues, ON LM type tests for seasonal unit roots in quarterly data, *Econometrics Journal* **5** (2002) 176–195.
49. M. J. Hinich, Testing for Gaussianity and linearity of a stationary time series, *Journal of Time Series Analysis* **3** (1982) 169–176.
50. J. Theiler, A. Longlin, B. Galdrikian and D. D. Farmer, Testing for nonlinearity in time series: The method of surrogate data, *Physica D* **58** (1992) 77–94.
51. R. Kunhimangalam, P. K. Joseph and O. K. Sujith, Nonlinear analysis of EEG signals: Surrogate data analysis, *IRBM* **29** (2008) 239–244.
52. J. H. Moore, Detection of linear and nonlinear dependencies in time series using the method of surrogate data in S-PLUS, *Computer Methods and Programs in Biomedicine* **63**(2) (2000) 117–121.

53. T. Suzuki, T. Ikeguchi and M. Suzuki, Algorithms for generating surrogatedata for sparsely quantized time series, *Physica D* **231**(2) (2007) 108−115.
54. C. P. Unsworth, M. R. Cowper, S. McLaughlin and B. Mulgrew, A new method to detect nonlinearity in a time-series: Synthesizing surrogate data using a Kolmogorov−Smirnoff tested, hidden Markov model, *Physica D* **155** (2001) 51−68.
55. S. Spasić, Surrogate data test for nonlinearity of the rat cerebellar electrocorticogram in the model of brain injury, *Signal Processing* **90**(12) (2010) 3015−3025.
56. C. Gan, Pseudo-periodic surrogate test to sample time series in stochastic softening Duffing oscillator, *Physics Letters A* **357**(3) (2006) 204−208.
57. D. M. Keenan, A Tukey nonadditivity-type test for time series nonlinearity, *Biometrika* **72** (1985) 39−44.
58. R. S. Tsay, Nonlinearity tests for time series, *Biometrika* **73** (1986) 461−466.
59. W. Brock, D. Dechert, J. Scheinkman and B. LeBaron, A test for independence based on the correlation dimension, *Econometric Reviews* **15** (1996) 197−235.
60. A. I. McLeod and W. K. Li, Diagnostic checking ARMA time series models using squared-residual autocorrelations, *Journal of Time Series Analysis* **4** (1983) 269−273.
61. D. Peña and J. Rodríguez, The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series, *Journal of Statistical Planning and Inference* **136**(8) (2006) 2706−2718.
62. G. Rodríguez and M. J. Sloboda, Modeling nonlinearities and asymmetries in quarterly revenues of the US telecommunications industry, *Structural Change and Economic Dynamics* **16** (2005) 137−158.
63. E. E. Peters, *Fractal Market Analysis* (Wiley, New York, 1994).
64. J. Alvarez-Ramirez, J. C. Echeverria and E. Rodriguez, Performance of a high-dimensional method for Hurst exponent estimation, *Physica A: Statistical Mechanics and its Applications* **387**(26) (2008) 6452−6462.
65. J. Fell and J. Roschke, Discrimination of sleep stages: A comparison between spectral and nonlinear EEG measures, *Electroencephalography and Clinical Neurophysiology* **98**(5) (1996) 401−410.
66. J. S. Richman and J. R. Moorman, Physiological time-series analysis using approximate and sample entropy, *American Journal of Physiology: Heart and Circulatory Physiology* **278**(6) (2000) H2039−H2049.
67. P. Walters, *An Introduction to Ergodic Theory* (Springer, New York, 1982).
68. N. Marwan, M. C. Romano, M. Thiel and J. Kurths, Recurrence plots for the analysis of complex systems, *Physics Reports* **438** (2007) 237−329.
69. G. Cottone and M. D. Paola, A new representation of power spectral density and correlation function by means of fractional spectral moments, *Probabilistic Engineering Mechanics* **25**(3) (2010) 348−353.
70. M. Uebele and A. Ritschl, Stock markets and business cycle comovement in Germany before World War I: Evidence from spectral analysis, *Journal of Macroeconomics* **31**(1) (2009) 35−57.
71. T. Hart, T. Coulson and P. N. Trathan, Time series analysis of biologging data: Autocorrelation reveals periodicity of diving behaviour in macaroni penguins, *Animal Behaviour* **79**(4) (2010) 845−855.
72. W. Chen, Use of recurrence plot and recurrence quantification analysis in Taiwan unemployment rate time series, *Physica A: Statistical Mechanics and its Applications* **390**(7) (2011) 1332−1342.
73. K. Shi, C. Q. Liu, N. S. Ai and X. H. Zhang, Using three methods to investigate time-scaling properties in air pollution indexes time series, *Nonlinear Analysis: Real World Applications* **9**(2) (2008) 693−707.

74. A. C. Harvey, *Forecasting Structural Time Series Models and the Kalman Filler* (Cambridge University Press, Cambridge, 1989).
75. F. Canova and B. E. Hansen, Are seasonal patterns constant over time? A test for seasonal stability, *Journal of Business and Economic Statistics* **13** (1995) 237–252.
76. M. Caner, A locally optimal seasonal unit-root test, *Journal of Business and Economic Statistics* **16** (1998) 349–356.
77. W. K. Tam and G. C. Reinsel, Tests for seasonal unit root in ARIMA models, *Journal of the American Statistical Association* **92** (1997) 725–738.
78. D. A. Dickey, D. P. Hasza and W. A. Fuller, Testing for unit root in seasonal time series, *Journal of the American Statistical Association* **79** (1984) 355–367.
79. R. A. Woods, Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks, *Advances in Water Resources* **32**(10) (2009) 1465–1481.
80. G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control* (Holden-Day, San Francisco, 1976).
81. J. W. Taylor, Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles, *International Journal of Forecasting* **26**(4) (2010) 627–646.
82. R. Kaiser and A. Maravall, Combining filter design with model-based filtering (with an application to business-cycle estimation), *International Journal of Forecasting* **21**(4) (2005) 691–710.
83. R. E. Quandt, Tests of the hypothesis that a linear regression obeys two separate regimes, *Journal of the American Statistical Association* **55** (1960) 324–330.
84. M. Hušková, C. Kirch, Z. Prašková and J. Steinebach, On the detection of changes in autoregressive time series II. Resampling procedures, *Journal of Statistical Planning and Inference* **138** (2008) 1697–1721.
85. R. L. Brown, J. Durbin and J. M. Evans, Techniques for testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society: Series B* **37** (1975) 149–163.
86. E. Zivot and D. Andrews, Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis, *Journal of Business and Economic Statistics* **10** (1992) 251–270.
87. S. C. Wong and A. A. Barhorst, Stochastic analysis of Poisson impact series using discrete form, spectrum analysis and time correlation, *Mechanical Systems and Signal Processing* **21** (2007) 151–164.
88. W. M. Lin, H. J. Gowa and M. T. Tsai, Hybridizing particle swarm optimization with signal-to-noise ratio for numerical optimization, *Expert Systems with Applications* **38**(11) (2011) 14086–14093.
89. S. Aja-Fernández, A. Tristán-Vega and C. Alberola-López, Noise estimation in single- and multiple-coil magnetic resonance data based on statistical models, *Magnetic Resonance Imaging* **27**(10) (2009) 1397–1409.
90. S. Rangachari and P. C. Loizou, A noise-estimation algorithm for highly non-stationary environments, *Speech Communication* **48**(2) (2006) 220–231.
91. A. T. Johansson and P. R. White, Instantaneous frequency estimation at low signal-to-noiseratios using time-varying notch filters, *Signal Processing* **88**(5) (2008) 1271–1288.
92. J. B. Gao, Recognizing randomness in a time series, *Physica D* **106** (1997) 49–56.
93. C. J. Cellucci, A. M. Albano, P. E. Rapp, R. A. Pittenger and R. C. Josiassen, Detecting noise in a time series, *Chaos* **7** (1997) 414–422.
94. M. B. Kennel, R. Brown and H. D. I. Abarbanel, Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A* **45**(6) (1992) 3403–3411.
95. P. R. Winters, Forecasting sales by exponentially weighted moving averages, *Management Science* **6** (1960) 324–342.