Phys. Lett. A 248 (1998) 191-202.

Detecting modes with nontrivial dynamics embedded in colored noise: Enhanced Monte Carlo SSA and the case of climate oscillations

Milan Paluš
Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic;[*]
and
Dagmar Novotná
Institute of Atmospheric Physics, Academy of Sciences of the Czech Republic
Boční II/1401, 141 31 Prague 4, Czech Republic

August 14, 1998

Abstract

Singular spectrum analysis (SSA) is a useful tool for identification and extraction of oscillatory or other signals from noisy background. Its basic form, however, is reliable when a signal is embedded in white noise, while presence of "colored" noises could lead to spurious results. Recently, Monte Carlo SSA, based on so-called surrogate data technique, has been introduced in order to increase reliability of detecting signals embedded in colored noises, which are usually present in geophysical data. We propose to enhance the Monte Carlo SSA by evaluating and testing regularity of dynamics (quantified by so-called coarse-grained entropy rates) of the SSA modes against the colored noise null hypothesis, in addition to the test based on variance (eigenvalues). We demonstrate that such an approach can improve the test reliability in detection of relatively more regular dynamical modes than those obtained by decomposition of colored noises, in particular, in identification of irregular oscillations embedded in red noise. The method is illustrated in detection of oscillations with a period of eight years in historical temperature records obtained from several European locations, as well as in detection of approximately five-year cycles in the global temperature series.

1 Introduction

Singular spectrum (or singular system) analysis (SSA) in its original form (also known as principal component analysis, or Karhunen-Loève decomposition) is a method for identification and distinction from noise of important information in multivariate data. It is based on an orthogonal decomposition of a covariance matrix of multivariate data under study. The SSA provides an orthogonal basis onto which the data can be transformed making thus individual data components ("modes") linearly independent. Each of the orthogonal modes (projections of the original data onto new orthogonal basis vectors) is characterized by its variance, which is given by the related eigenvalue of the covariance matrix. In dimensionality and/or noise reduction tasks it is supposed that the "information" in the data is confined into a few directions along the basis vectors with the largest eigenvalues (variance) while the rest of the modes with smaller eigenvalues consist just of noise present in the data.

In this paper we will deal with a univariate version of SSA, in which the analyzed data is a univariate time series and the decomposed matrix is a time-lag covariance matrix, i.e., instead of several components of a multivariate data, a univariate[1] time series and its time-lagged versions are considered. This type of the SSA

[*] Corresponding author, e-mail: mp@uivt.cas.cz, mp@santafe.edu.

[1] However, the presented method can be easily generalized for applications to multivariate time series.

application, which recently became frequently used especially in the field of meteorology and climatology [1, 2, 3, 4, 5], can provide a decomposition of the studied time series into orthogonal components (modes) with different dynamical properties and thus "interesting" phenomena such as slow modes (trends) and regular or irregular oscillations (if present in the data) can be identified and retrieved from the background of noise and/or other "uninteresting" non-specified processes.

In the traditional SSA, the distinction of "interesting" components (signal) from noise is based on finding a threshold (jump-down) to a "noise floor" in a sequence of eigenvalues given in a descending order. This approach might be problematic if the signal-to-noise ratio is not sufficiently large, or the noise present in the data is not white but "colored." For such cases statistical approaches utilizing the Monte Carlo simulation techniques have been proposed [3, 6] for reliable signal/noise separation. The particular case of the Monte Carlo SSA (MCSSA) which considers the "red" noise, usually present in climatic data such as historical temperature records, has recently been introduced by Allen & Smith [7].

The latter method itself is a sophisticated statistical test, however, the only discriminant is the variance (eigenvalues) of particular modes. In this paper we demonstrate that extending MCSSA by testing dynamical properties of the modes can increase sensitivity and reliability of the test, as well as it can support interpretation of a positive result as an identification of a process with "more interesting" or nontrivial dynamics in comparison to "trivial" dynamics of the modes obtained by the decomposition of colored noise. In the presented implementation we quantify and test the "regularity" of dynamics of the orthogonal modes by using the ideas of coarse-grained entropy rates [8]. A signal (a process with nontrivial dynamics) is identified if the regularity of some modes is significantly higher than the regularity of related modes obtained from the colored noise surrogate (simulated) data.

The singular system analysis and the Monte Carlo SSA are briefly reviewed in Sec. 2. Basic ideas and tools for quantifying regularity of complex signals by using coarse-grained entropy rates are summarized in Sec. 3. The enhanced MCSSA, proposed as the extension of the MCSSA by testing regularity of the modes, is presented in Sec. 4. The considered methods are demonstrated by using simulated data through the Sections 2 and 4, and applied to real data – historical temperature records from several European locations as well as to a so-called global temperature record in Sec. 5. Other possible applications of the presented method are discussed and conclusion is given in Sec. 6. Appendix is devoted to an algorithm for estimating the marginal redundancy, an information-theoretic functional used for characterizing the regularity of the modes.

2 Monte Carlo singular system analysis

Let a univariate time series $\{y(i)\}$, $i = 1, \dots, N_0$, be a realization of a stochastic process $\{Y(i)\}$ which is stationary and ergodic. A map into a space of n-dimensional vectors $\mathbf{x}(i)$ with components $x^k(i)$, where $k = 1, \dots, n$, is given as

$$x^k(i) = y(i + k - 1). \tag{1}$$

The sequence of the vectors $\mathbf{x}(i)$, $i = 1, \dots, N = N_0 - (n-1)$, is usually referred to as the $n \times N$ trajectory matrix $\mathbf{X} = \{x_i^k\}$, the number n of the constructed components is called the embedding dimension, or the length of the (embedding) window. These terms are related to the application of SSA in the problem of reconstructing (chaotic) attractors of dynamical systems from univariate time series, as proposed by Broomhead & King [9] and discussed, among others, by Paluš & Dvořák [10]. We use here the same implementation of SSA as described in [9] and [10], however, we have to emphasize that the method presented here is devoted to detecting signals of unspecified origin in a specified type of background noise, i.e., *no* specific hypothesis about a process underlying the signal, such as a chaotic attractor, is considered.

Suppose that the studied time series $\{y(i)\}$ results from a linear combination of m different dynamical modes, $m < n$. Then, in an ideal case, the rank of the trajectory matrix $\mathbf{X}$ is $\text{rank}(\mathbf{X}) = m$, and, $\mathbf{X}$ can be transformed into a matrix with only m nontrivial linearly independent components. Instead of the $n \times N$ matrix $\mathbf{X}$ it is more convenient to decompose the symmetric $n \times n$ matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$, since $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{C})$. The elements of the covariance matrix $\mathbf{C}$ are

$$c_{kl} = (1/N)\sum_{i=1}^{N} x^k(i)x^l(i), \tag{2}$$

where $1/N$ is the proper normalization and the components $x^k(i)$, $i = 1, \ldots, N$, are supposed to have zero mean. The symmetric matrix $\mathbf{C}$ can be decomposed as

$$\mathbf{C} = \mathbf{V}\Sigma\mathbf{V}^T, \tag{3}$$

where $\mathbf{V}= \{v_{ij}\}$ is an $n \times n$ orthonormal matrix, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, σ_i are non-negative eigenvalues of $\mathbf{C}$ by convention given in descending order $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$. If rank($\mathbf{C}$)= $m < n$, then

$$\sigma_1 \geq \ldots \geq \sigma_m > \sigma_{m+1} = \ldots = \sigma_n = 0. \tag{4}$$

In the presence of noise, however, all eigenvalues are positive and the relation (4) takes the following form [9]:

$$\sigma_1 \geq \ldots \geq \sigma_m >> \sigma_{m+1} \geq \ldots \geq \sigma_n > 0. \tag{5}$$

Then, the modes ξ_i^k

$$\xi_i^k = \sum_{l=1}^{n} v_{lk} x_i^l, \tag{6}$$

for $k = 1, \ldots, m$ are considered as the "signal" part, and the modes ξ_i^k, $k = m+1, \ldots, n$, are considered as the noise part of the original time series. The "signal" modes can be used to reconstruct the denoised signal $\tilde{x}_l^k$ as

$$\tilde{x}_i^k = \sum_{l=1}^{m} v_{kl} \xi_i^l. \tag{7}$$

Of course, the original time series x_i^k can be reconstructed back from the modes as

$$x_i^k = \sum_{l=1}^{n} v_{kl} \xi_i^l. \tag{8}$$

In the latter relation – decomposition (8), the modes ξ_i^k can also be interpreted as time-dependent coefficients and the orthogonal vectors $\mathbf{v}_k = \{v_{kl}\}$ as basis functions, usually called the empirical orthogonal functions (EOF's).

The clear eigenvalue-based signal/noise distinction (5) can be obtained only in particularly idealized situation when the signal/noise ratio is large enough and the background consists of white noise. A kind of colored noise, the "red" noise, which is particularly important for its presence in many geophysical processes [7], can be modeled by using an AR(1) model (autoregressive model of the first order):

$$u(i) - \hat{u} = \alpha(u(i-1) - \hat{u}) + \gamma z(i), \tag{9}$$

where $\hat{u}$ is the process mean, α and γ are process parameters, and $z(i)$ is a Gaussian white noise with a zero mean and a unit variance.

The red noises possess power spectra of the $1/f$ type,[2] and their SSA eigenspectra have the same character, i.e., an eigenspectrum of a red noise is equivalent to a coarsely discretized power spectrum, where the number of frequency bins is given by the embedding dimension n. The eigenvalues related to the slow modes are much larger than the eigenvalues of the modes related to higher frequencies. Thus, in the classical SSA approach applied to a red noise, the eigenvalues of the slow modes might incorrectly be interpreted as a (nontrivial) signal, or, on the other hand, a nontrivial signal embedded in a red noise might be neglected, if its variance is smaller than the slow-mode eigenvalues of the background red noise. Therefore the mutual comparison of eigenvalues inside an eigenspectrum cannot lead to a reliable detection of a nontrivial signal, if a red noise is present in studied data. In order to correctly detect a signal in a red noise, the following approach has been proposed [7]:

[2] The colored noises are usually defined as stochastic processes with spectra of the $1/f^\alpha$ type, a more general definition considers linearly filtered white noise. Here we consider a specific kind of the colored (red) noise represented as the AR(1) process (9), applications of the proposed method to problems with more general types of colored noises are mentioned in Conclusion.

First, the eigenvalues are plotted not according to their values, but according to a frequency associated with a particular mode (EOF), i.e., the eigenspectrum in this form becomes a sort of a (coarsely) discretized power spectrum in general, not only in the cases of red noises (when the eigenspectra have naturally this form, as mentioned above).
Second, an eigenspectrum obtained from a studied data is compared, in a frequency-by-frequency way, with eigenspectra obtained from a set of realizations of an appropriate noise model (such as the AR(1) model (9)), i.e., an eigenvalue related to a particular frequency bin obtained from the data is compared with a range of eigenvalues related to the same frequency bin, obtained from the set of so-called surrogate data, i.e., the data artificially generated according to the chosen noise model (null hypothesis) [7, 11, 12, 13].

The detection of a nontrivial signal in an experimental time series becomes a statistical test in which the null hypothesis that the experimental data were generated by a chosen noise model is tested. When (an) eigenvalue(s) associated with some frequency bin(s) differ(s) with a statistical significance from the range(s) of related noise model eigenvalues, then one can infer that the studied data cannot be fully explained by the considered null hypothesis (noise model) and could contain an additional (nontrivial) signal.

This is a rough sketch of the approach, for which we will use the term Monte Carlo SSA (MCSSA), as coined by Allen & Smith [7] (altough the same term was earlier used for other SSA methods, which considered white noise background [3, 6]). Allen & Smith [7] give a detailed account of the MCSSA approach analyzing various levels of null hypotheses and related MCSSA techniques. Here we will consider a slightly simplified MCSSA version, in which we will demonstrate a special kind of extension by testing regularity of the modes (Sec. 4). We will realize the basic version of MCSSA as follows.

1. The studied time series undergoes SSA as described above, i.e., using an embedding window of length n, the $n \times n$ lag-correlation matrix $\mathbf{C}$ is decomposed using the SVDCMP routine [25]. In the eigenspectrum, the position of each eigenvalue on the abscissa is given by the dominant frequency associated with the related EOF, i.e., detected in the related mode. That is, the studied time series is projected onto the particular EOF, the power spectrum of the projection (mode) is estimated, and the frequency bin with the highest power is identified. This spectral coordinate is mapped onto one of the n frequency bins, which equidistantly divide the abscissa of the eigenspectrum.

2. An AR(1) model is fitted on the series under study, the residuals are computed.

3. The surrogate data are generated using the above AR(1) model, "scrambled" (randomly permutated in temporal order) residuals are used as innovations.

4. Each realization of the surrogates undergoes SSA as described in item 1. Then, the eigenvalues for the whole surrogates set, in each frequency bin, are sorted and the values for the 2.5th and 97.5th percentiles are found. In eigenspectra, the 95% range of the surrogates eigenvalue distribution is illustrated by a horizontal bar between the above percentile values.

5. For each frequency bin, the eigenvalue obtained from the studied data is compared with the range of the surrogate eigenvalues. If an eigenvalue lies outside the range given by the above percentiles, the null hypothesis of the AR(1) process is rejected, i.e., there is a probability $p < 0.05$ that the data can be explained by the null noise model.

Performing MCSSA using the embedding window of the length n, there are n eigenvalues in the eigenspectrum, and n statistical tests are done. Therefore the problem of the simultaneous statistical inference should be considered in applications, however, we will not discuss this topic here. (See [13] and references therein.)

Rejecting the null hypothesis of the AR(1) (or other appropriate) noise model, one can infer that there is something more in the data than a realization of the null hypothesis model, however, it is not an evidence that any interesting nontrivial signal, such as oscillations, is present. Identification of oscillations in the SSA and MCSSA approach by searching pairs of symmetric and anti-symmetric EOFs is discussed in [7]. Below (Sec. 4) we will demonstrate a way to identification of "dynamically interesting" modes in general (by means of testing their regularity), however, oscillations will be considered as an example.

The presented approach is demonstrated here by using numerically generated data, illustrated in Fig. 1. A periodic signal with randomly variable amplitude (Fig. 1a) was mixed with a realization of an AR(1) process with a strong slow component[3] (Fig. 1b), obtaining the signal to noise ratio 1:2 (Fig. 1c), and 1:4 (Fig. 1d). (The given signal/noise ratios are the ratios of the standard deviations.) The latter two series are analyzed by the presented method.

The eigenspectrum of the time series consisting of the signal (Fig. 1a) and the AR(1) noise (Fig. 1b) in the ratio 1:2 (Fig. 1c) is presented in Fig. 2a, where logarithms of the eigenvalues are plotted as the diamonds ("LOG POWER"). The series is considered as unknown experimental data, so that an AR(1) model is fitted on the data and the surrogates are generated as described above. The vertical bars in the eigenspectrum represent the surrogate eigenvalue ranges from 2.5th to 97.5th percentiles, which were obtained from 1500 surrogate realizations (here, as well as in the following examples). The eigenvalues of the AR(1) surrogates uniformly fill all the n frequency bins (here, as well as in the following examples $n = 100$), while in the case of the test data, some bins are empty, others contain one, two or more eigenvalues. We plot the surrogates bars only in those positions, in which (an) eigenvalue(s) of the analyzed data exist(s). Note the $1/f$ character of the surrogate eigenspectrum, i.e., the eigenvalues plotted according to increasing dominant frequency associated with the related modes are monotonously decreasing in a $1/f^{\alpha}$ way. The low-frequency part of the eigenspectrum from Fig. 2a is enlarged in Fig. 2b. The two data eigenvalues related to the frequency 0.02 (cycles per time unit) are clearly above the surrogate bar, i.e., they are significant on the 95% level and the null hypothesis is rejected. Further study of the significant modes shows that they are related to the embedded in noise signal from Fig. 1a, in particular, one of the modes contains the signal together with some noise of similar frequencies, and the other include an oscillatory mode shifted by $\pi/2$ relatively to the former. Note that the simple SSA based on the mutual comparison of the data eigenvalues could be misleading, since the AR(1) noise itself "produces" two or three eigenvalues which are larger than the two eigenvalues related to the signal embedded in the noise.

The same analysis applied to the series possessing the signal/noise ratio 1:4 (Figs. 2c,d), however, fails to detect the embedded signal — all eigenvalues obtained from the test data are well confined between the 2.5th and 97.5th percentiles of the surrogate eigenvalues distributions.

3 Coarse-grained entropy rates and measuring regularity of complex time series

Consider n discrete random variables $X_1, \ldots, X_n$ with sets of values $\Xi_1, \ldots, \Xi_n$, respectively. The probability distribution for an individual X_i is $p(x_i) = \Pr\{X_i = x_i\}$, $x_i \in \Xi_i$. We denote the probability distribution function by $p(x_i)$, rather than $p_{X_i}(x_i)$, for convenience. Analogously, the joint distribution for the n variables $X_1, \ldots, X_n$ is $p(x_1, \ldots, x_n) = \Pr\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\}$, $(x_1, \ldots, x_n) \in \Xi_1 \times \ldots \times \Xi_n$.

The marginal redundancy [15, 22, 8, 13] $\varrho(X_1, \ldots, X_{n-1}; X_n)$, in the case of two variables also known as mutual information $I(X_1; X_2)$, quantifies the average amount of information about the variable X_n, contained in the $n-1$ variables $X_1, \ldots, X_{n-1}$, and is defined as

$$\varrho(X_1, \ldots, X_{n-1}; X_n) \;=\; \sum_{x_1 \in \Xi_1} \cdots \sum_{x_n \in \Xi_n} p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_{n-1}) p(x_n)}. \tag{10}$$

Now, let $\{X_i\}$ be a stochastic process, i.e., an indexed sequence of random variables, characterized by the joint probability distribution function $p(x_1, \ldots, x_n)$. The entropy rate of $\{X_i\}$ is defined as

$$h = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n), \tag{11}$$

where $H(X_1, \ldots, X_n)$ is the joint entropy of the n variables $X_1, \ldots, X_n$ with the joint distribution $p(x_1, \ldots, x_n)$:

$$H(X_1, \ldots, X_n) = - \sum_{x_1 \in \Xi_1} \cdots \sum_{x_n \in \Xi_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n). \tag{12}$$

A studied time series $\{y(t)\}$ can be considered as a realization of a stochastic process, i.e., a sequence of stochastic variables. Uncertainty in a stochastic variable is measured by its *entropy*. The rate in which the

[3] The used model is defined as $x_i = 0.933 x_{i-1} + \xi_i$, where ξ_i are Gaussian deviates with a zero mean and a unit variance.

stochastic process "produces" uncertainty is measured by its *entropy rate,* defined above (11). The concept of entropy rates is common to the theory of stochastic processes as well as to the information theory where the entropy rates are used to characterize information production by information sources [15].

Alternatively, the time series $\{y(t)\}$ can be considered as a projection of a trajectory of a dynamical system, evolving in some measurable state space. Then, it can be characterized by the Kolmogorov-Sinai entropy (KSE) [16, 17, 18, 19, 20] which is a topological invariant, suitable for classification of dynamical systems or their states, and is related to the sum of the system's positive Lyapunov exponents (LE) according to the theorem of Pesin [21].

A way from the entropy rate of a stochastic process to the Kolmogorov-Sinai entropy (KSE) of a dynamical system can be straightforward due to the fact that any stationary stochastic process corresponds to a measure-preserving dynamical system, and vice versa [19]. Then for a definition[4] of the KSE we can consider the equation (11), however, the variables X_i should be understood as m-dimensional variables, according to a dimensionality of the dynamical system. If the dynamical system is evolving in continuous (probability) measure space, then any entropy depends on a partition ξ chosen to discretize the space and the KSE is defined as a supremum over all finite partitions [18, 19, 20].

Above we have introduced the entropy rates for both stochastic processes and dynamical systems, i.e., the entropy rates are meaningful quantities for characterization of stationary processes irrespectively of their origin. However, possibilities to compute the entropy rates from experimental data are limited to a few exceptional cases: for stochastic processes it is possible, e.g., for finite-state Markov chains [15]. In a case of a low-dimensional dynamical system on continuous measure space, Fraser [22] proposed to estimate its KSE from an asymptotic behavior of the marginal redundancy, computed from a time series generated by the dynamical system. Then the variables X_i are substituted as

$$X_i = y(t + (i-1)\tau), \tag{13}$$

and, considering stationarity of the series (underlying system), the marginal redundancy

$$\varrho^n(\tau) \equiv \varrho(y(t), y(t+\tau), \ldots, y(t+(n-2)\tau); y(t+(n-1)\tau)) \tag{14}$$

is a function of the embedding dimension n and the time delay τ, independent of the time t.

It was shown [22, 23, 24] that if the underlying dynamical system is m-dimensional and the marginal redundancy $\varrho^n(\tau)$ is estimated using a partition fine enough (to attain so-called generating partition [18, 20, 22]), then the asymptotic behavior

$$\varrho^n(\tau) \approx H_1 - |\tau| h \tag{15}$$

is attained for $n = m+1, m+2, \ldots$, for some range of τ. The constant H_1 is related to $\varrho^n(0)$. For observation of the behaviour (15), however, even in a case of a low-dimensional chaotic system, a large amount of practically noise-free data is necessary [24, 8], which is usually unavailable in experimental practice. Therefore Paluš [8] has proposed to give up the effort for estimating the exact entropy rates, and defined "coarse-grained entropy rates" (CER's) instead. The CER's are not meant as estimates of the exact entropy rates, but as quantities which can depend on a particular experimental and numerical set-up, however, quantities which have the same meaning as the exact entropy rates, i.e., which are measures of systems' production of uncertainty, or, in other words, which can be used as measures of regularity and predictability of analyzed time series, however, in the relative sense: Two or several datasets can be compared according to their regularity and predictability, providing they were measured in the same experimental conditions and their CER's were estimated using the same numerical parameters.

Paluš [8] defines one type of CER as follows: In a particular application, we compute the marginal redundancies $\varrho^n(\tau)$ for all analyzed datasets and find such τ_{max} that for $\tau' \geq \tau_{max}$: $\varrho^n(\tau') \approx 0$ for all the datasets. Then we define a norm of the marginal redundancy

$$||\varrho^n|| = \frac{\sum_{\tau=\tau_0}^{\tau_{max}} \varrho^n(\tau)}{\tau_{max} - \tau_0}. \tag{16}$$

[4] A more detailed and rigorous KSE definition can be found in monographs [18, 19, 20] or in the paper of Paluš [24] and references therein.

The CER $h^{(1)}$, defined as

$$h^{(1)} = \frac{\varrho^n(\tau_0) - ||\varrho^n||}{||\varrho^n||}, \tag{17}$$

has been found able to discern systems with different exact entropy rates, in particular, to discern different states of chaotic systems, as well as stochastic systems with different correlation lengths, in cases of both numerically simulated and real experimental data [8, 27].

4 Extending MCSSA by measuring and testing regularity of the modes

Estimating the marginal redundancy (10) using the marginal equiquantization (see Appendix) the data are effectively transformed into a uniform marginal distribution. Then, choosing in (17) $\tau_0 = 0$, the marginal redundancy norm (16) is the only discriminating member in (17). Therefore we will use the marginal redundancy norm (16) itself as an informal measure of regularity (and predictability) for studied time series, and thereafter referred to it as a *regularity index.* Indeed, the marginal redundancy norm (16) reflects an average level of dependence between $y(t)$ and $y(t+\tau)$, for the chosen range of τ, that means, it is proportional to regularity and predictability of analyzed time series.

In the following the regularity index (16) is used for characterizing dynamics of the SSA modes. After the decomposition of the $n \times n$ lag-correlation matrix $\mathbf{C}$, the regularity index is estimated for each of the n modes (the projections of the studied time series $\{y(t)\}$ onto the orthogonal basis of EOFs — eigenvectors of $\mathbf{C}$). The same procedure is then applied to the surrogate data and the regularity indices are processed by the same way as the mode variances (the eigenvalues of $\mathbf{C}$), i.e., they are plotted according to the dominant frequencies associated with the related modes. For each of the n frequency bins, the regularity index obtained for the particular projection of the studied time series $\{y(t)\}$ is compared with the chosen range (e.g., from the 2.5th to the 97.5th percentile of the distribution) of the surrogate regularity indices obtained for the same frequency bin. Again, finding a regularity index (indices) significantly different from the range of the related surrogate indices, one infers that the chosen null hypothesis (noise model) does not explain the data. In particular, if a regularity index (indices) is (are) significantly larger than the related surrogate indices, one can infer that the mode contains a signal which is more regular (and potentially better predictable) than the related surrogate mode, i.e., a linearly filtered colored noise.

The above described artificial series (Fig. 1) which underwent the standard MCSSA (Fig. 2), were tested by using the regularity index (16) and the MCSSA approach, as described above. In the case of the signal to noise ratio 1:2 (Figs. 3a,b) one data regularity index has been found significantly higher than the related surrogate indices. It was obtained from the mode related to the frequency bin 0.02, as in the case of the significant eigenvalues in Figs. 2a, b. This is the mode which contains the embedded signal (Fig. 1a) together with some noise of similar frequencies. The orthogonal mode, related to the same frequency bin, which has the variance comparable to the former (Figs. 2a,b), has its regularity index close to the 97.5th percentile of the surrogate regularity indices distribution. In other words, if a (nearly) periodic signal is embedded in a (colored) noise, the SSA approach, in principle, is able to extract this signal together with some noise of close frequencies, and produces an orthogonal “ghost” mode which has comparable variance, however, its dynamical properties are closer to those of the modes obtained from the pure noise (null model), as measured by the regularity index (16). Nevertheless, the regularity index used as a test statistic in the MCSSA manner is able to detect the embedded signal with a high statistical significance in this case (signal:noise = 1:2), as well as in the case of the signal to noise ratio 1:4 (Figs. 3c,d), when the standard (variance-based MCSSA) failed (Figs. 2c, d). In the latter case the orthogonal “ghost” mode did not appear, and the regularity index of the signal mode is lower than in the previous case, since the mode contains larger portion of the isospectral noise, however, the signal mode regularity index is still safely above the surrogate bar, i.e., significant with $p < 0.05$ (Fig. 3d).

5 Application of the enhanced MCSSA to temperature records

Monthly average surface air temperature series from ten European stations (Stockholm, De Bilt, Paris – Le Bourget, Geneve – Cointrin, Berlin – Tempelhof, Munich – Riem, Vienna – Hohe Warte, Budaors, Wroclaw II, obtained from the Carbon Dioxide Information Analysis Center (CDIAC) Internet server;[5] and a series from Prague[6]) from the period 1781 – 1988 were analyzed by using the above introduced MCSSA enhanced by the evaluation of the regularity indices. The long-term monthly averages were subtracted from the data, so that the annual cycle was effectively filtered-out. Also the Jones global North hemisphere surface air temperature series[7] [26], from the period 1851 – 1984, has been analysed.

The enhanced MCSSA analysis of the Prague temperature series is presented in Fig. 4. In the classical MCSSA test based on eigenvalues (Figs. 4a, b) the only significance has been found for the zero frequency mode, i.e., there is present a significant long-term trend inconsistent with the hypothesis of the AR(1) noise, however, no oscillations or other dynamical phenomena have been detected. The situation is different using the test based on the regularity index (Figs. 4c, d), when, in addition to the significant long-term trend, also another mode, related to oscillatory dynamics with a period of approximately eight years (0.01 cycles per month, Fig. 4d), has been found significantly different from the AR(1) null hypothesis.

Similar result has been found in the analysis of the Berlin series (Fig. 5a, c) and in the series from Wroclaw and De Bilt. In the data from the other six stations only the long-term trend has been found significant, but no oscillations. This result could lead to the question of simultaneous statistical inference, namely to the probability of randomly occurring significances in a part of the data set. Considering geographical locations of the stations, however, we can see a nonrandom pattern in the occurrence of the significant results, since the eight-year cycle has been found in the stations located slightly over 50 degrees of northern latitude.

The standard (variance-based) MCSSA applied to the Jones global temperature series (Fig. 5b) discovered three modes with the variance significantly different from the AR(1) null hypothesis – the long term trend, just like in the above European series, and a mode related to the annual cycle (0.086 cycles per month, Fig. 5b), together with its orthogonal pair. Applying the test based on the regularity index, in addition to the above three significances, also another mode has been found significant, which is related to oscillations with a period about five years (0.015 cycles per month, Fig. 5d).

6 Conclusion

An extension of the Monte Carlo SSA method has been proposed, based on evaluating and testing regularity of dynamics (quantified by a regularity index inspired by so-called coarse-grained entropy rates) of the SSA modes against the colored noise null hypothesis in addition to the test based on variance (eigenvalues). It has been demonstrated that such an approach could enhance the test sensitivity and reliability in detection of relatively more regular dynamical modes than those obtained by decomposition of colored noises, in particular, in detection of irregular oscillations embedded in red noise. The method has been illustrated in detection of oscillations with a period of eight years in historical temperature records obtained from several European locations, as well as in detection of approximately five-year cycles in the global temperature series, related probably to the El Niño Southern Oscillations cycle. The detailed account of the climate related issues will be published elsewhere, here we presented the method in general, since it can be adapted also for different problems, namely the AR(1) noise null hypothesis can be replaced by different ones, depending on the problem solved. For instance, a study of a possibility to combine the MCSSA enhanced by the regularity index testing with the isospectral surrogates [12, 13] in detection of deterministic oscillations or unstable periodic orbits embedded in band-pass filtered noise of close frequencies is currently in progress.

[5] The Internet address is `ftp://cdiac.esd.ornl.gov/pub/ndp041`.

[6] Used by the courtesy of the Czech Hydrometeorological Institute, Prague.

[7] Also available at the CDIAC Internet server at `ftp://cdiac.esd.ornl.gov/pub/ndp003r1`.

Acknowledgements

The authors would like to thank I. Charvátová for valuable discussions and cooperation in the climatology project. This research is supported by the Grant Agency of the Czech Republic (grant No. 205/97/0921).

Appendix - Estimation of the marginal redundancy

When the discrete variables $X_1, \ldots, X_n$ are obtained from continuous variables in a continuous probability space, then the redundancies $\varrho(X_1, \ldots, X_{n-1}; X_n)$ depend on a partition ξ chosen to discretize the space. Various strategies have been proposed to define an optimal partition for estimating redundancies of continuous variables (see Refs. [23, 24, 13] and references therein). We have found that satisfactory results can be obtained by using simple box-counting method and by observing the following two rules:
a) The partition is defined by the marginal equiquantization method, i.e., the marginal histogram bins are defined not equidistantly but equiprobably, i.e., so that there is approximately the same number of samples in each marginal bin.
b) The relation between the number Q of quantization levels (marginal bins) and the effective[8] series length N in the computation of n-dimensional redundancy should be

$$N \geq Q^{n+1},$$

otherwise the results may be heavily biased [24, 8].

This algorithm does not provide unbiased estimates of absolute values, however, the absolute values of the redundancies are not important here. Applying this simple recipe and using the same parameters (N, n, Q and the τ-range), the redundancy estimator should bring consistent estimates in the relative sense that the marginal redundancies $\varrho^n(\tau)$ and, consequently, the CER's and/or the regularity indices (16), obtained from different datasets, are mutually comparable, and the CER's provide the classification of the datasets equivalent to the classification given by the exact entropy rates [8].

In all the examples presented in this paper, the marginal redundancy $\varrho^n(\tau)$ was estimated using $n = 2$ and $Q = 4$ equiquantal marginal bins, and $\tau_0 = 0$ and $\tau_{max} = 240$ (samples or months) were used for computing the regularity index (16). The total series lengths N_0 were 2560 samples in the case of the simulated data, and 2496 and 1608 monthly samples in the cases of the European and the Jones global temperature series, respectively.

References

[1] R. Vautard, M. Ghil, *Physica D* **35** (1989) 395.

[2] M.R. Allen, L.A. Smith, *Geophys. Res. Lett.* **21** (1994) 883.

[3] M. Ghil, R. Vautard, *Nature* **350(6316)** (1991) 324.

[4] C.L. Keppenne, M. Ghil, *J. Geophys. Res.* **97** (1992) 20449.

[5] P. Yiou, M. Ghil, J. Jouyel, D. Paillard, R. Vautard, *Clim. Dyn.* **9** (1994) 371.

[6] R. Vautard, P. Yiou, M. Ghil, *Physica D* **58** (1992) 95.

[7] M.R. Allen, L.A. Smith, *J. Climate* **9(12)** (1996) 3373.

[8] M. Paluš, *Physica D* **93** (1996) 64.

[8] If a univariate series is used to construct a time-delay n-dimensional embedding (Eq. 13), the effective series length N is $N = N_0 - (n-1)\tau$, where N_0 is the total series length, n is the embedding dimension (dimensionality of the marginal redundancy), and τ is the time delay. Note that the embedding dimension used in the SSA procedure is different from the dimensionality of the marginal redundancy used to characterize the mode dynamics.

[9] D.S. Broomhead, G.P. King, *Physica D* **20** (1986) 217.

[10] M. Paluš, I. Dvořák, *Physica D* **55** (1992) 221.

[11] L.A. Smith, *Physica D* **58** (1992) 50.

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

[13] M. Paluš, *Physica D* **80** (1995) 186.

[14] M. Paluš, *Phys. Lett. A* **213** (1996) 138.

[15] T.M. Cover and J.A. Thomas, *Elements of Information Theory* (J. Wiley & Sons, New York, 1991).

[16] A.N. Kolmogorov, *Dokl.Akad.Nauk SSSR* **124** (1959) 754.

[17] Ya.G. Sinai, *Dokl.Akad.Nauk SSSR* **124** (1959) 768.

[18] I.P. Cornfeld, S.V. Fomin, Ya.G. Sinai, *Ergodic Theory* (Springer, New York, 1982).

[19] K. Petersen, *Ergodic Theory* (Cambridge University Press, Cambridge, 1983).

[20] Ya. G. Sinai, *Introduction to Ergodic Theory* (Princeton University Press, Princeton, 1976).

[21] Ya.B. Pesin, *Russian Math. Surveys* **32** (1977) 55.

[22] A.M. Fraser, *IEEE Transactions on Information Theory* **35** (1989) 245.

[23] M. Paluš, In: A.S. Weigend and N.A. Gershenfeld, (eds.), *Time Series Prediction: Forecasting the Future and Understanding the Past,* Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XV (Addison–Wesley, Reading, Mass., 1993), p. 387.

[24] M. Paluš, *Neural Network World* **3/97** (1997) 269. (`http://www.uivt.cas.cz/~mp/papers/rd1a.ps`)

[25] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge 1986).

[26] P.D. Jones, S.C.B. Raper, R.S. Bradley, H.F. Diaz, P.M. Kelly, T.M.L. Wigley, *J. Climate and Applied Meteorology* **25(2)** (1986) 161.

[27] M. Paluš, C. Schöfl, A. Von Zur Mühlen, K. Brabant, K. Prank, Coarse-grained entropy rates quantify fast Ca^{2+} dynamics modulated by pharmacological stimulation, Proceedings of Pacific Symposium on Biocomputing, Maui, Hawaii, January 1998.

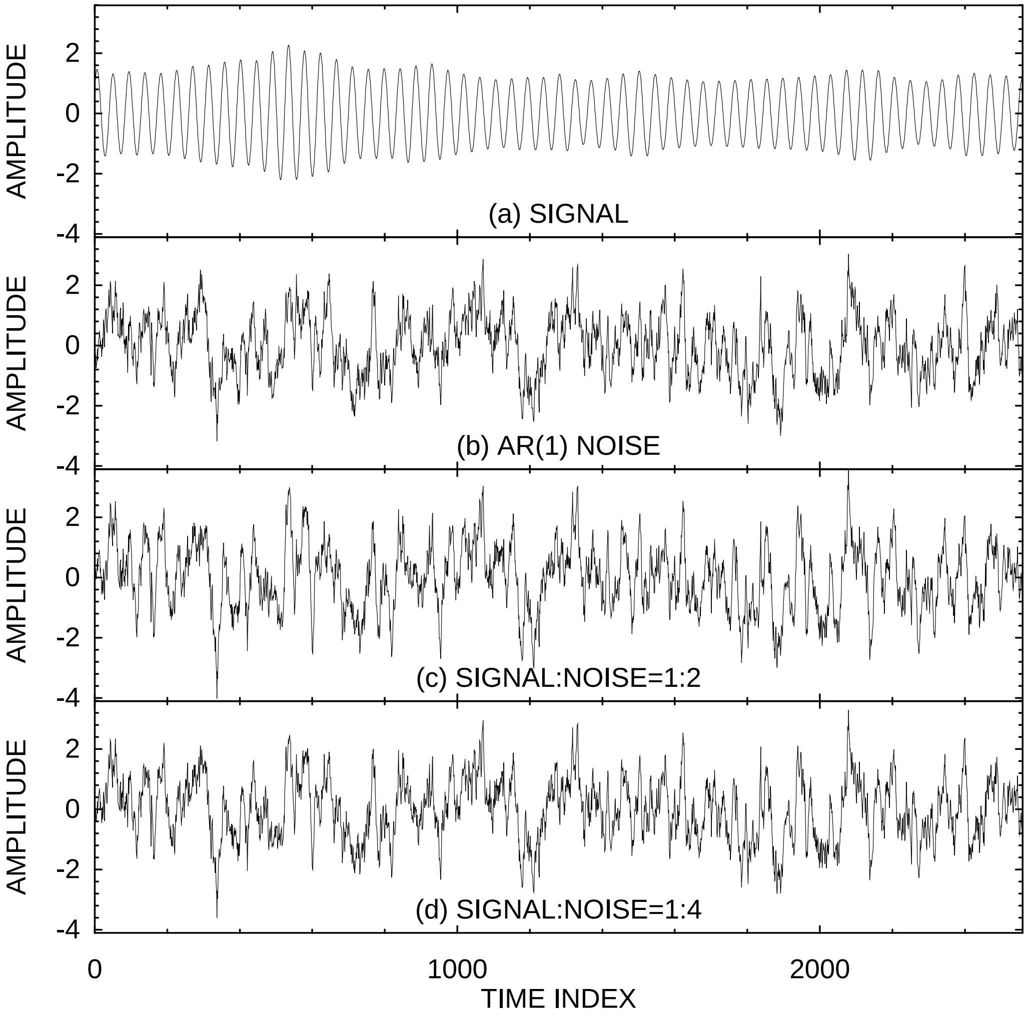

Figure 1: *Numerically generated test data: (a) A periodic signal with randomly variable amplitude was mixed with (b) a realization of an AR(1) process with a strong slow component, obtaining the signal to noise ratio 1:2 (c), and 1:4 (d).*

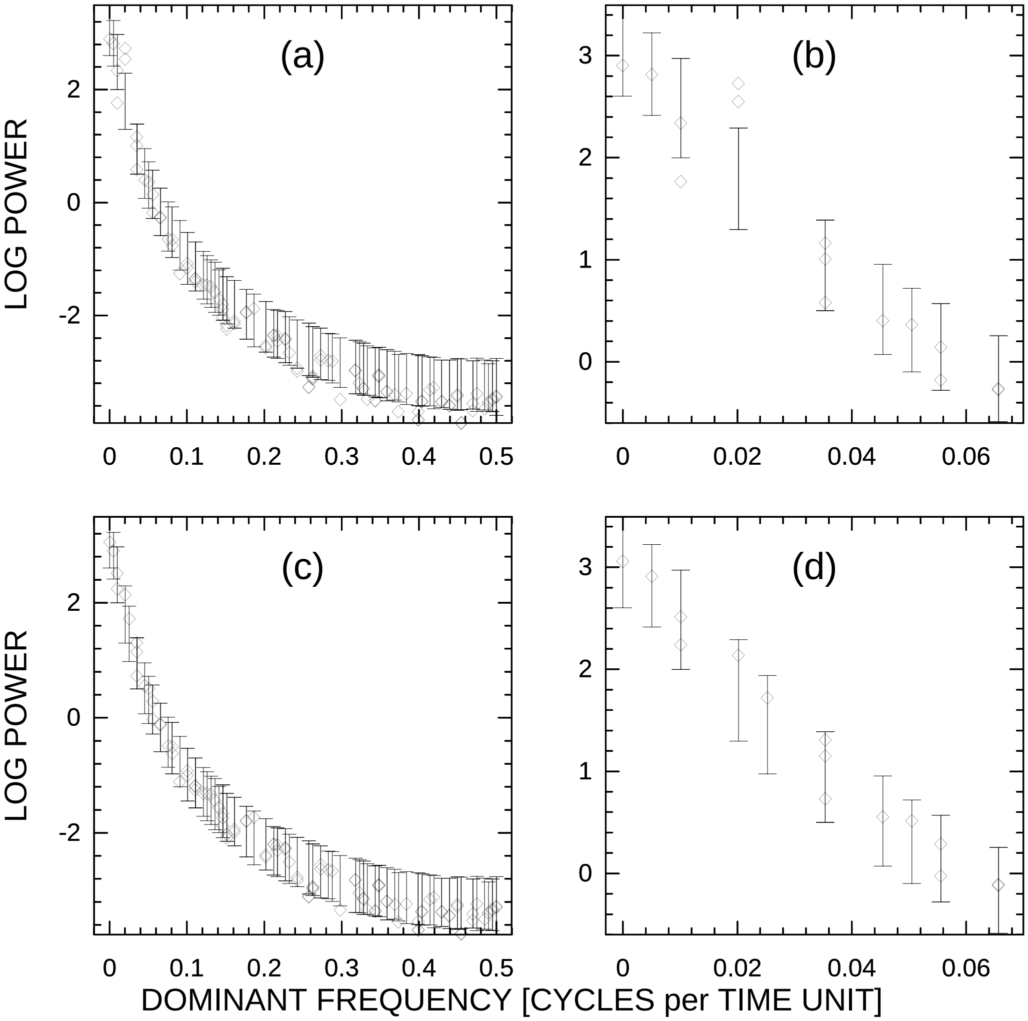

Figure 2: *Standard MCSSA analysis of the numerical data, presented in Fig. 1. Plots (a, b): Signal to noise ratio 1:2. (a) Logarithms of eigenvalues ("LOG POWER") plotted according to the dominant frequency associated with particular modes. Diamonds – eigenvalues for the analysed data; bars – 95% of the surrogate eigenvalues distribution, i.e., the bar is drawn from the 2.5th to the 97.5th percentiles of the surrogate eigenvalues distribution. (b) Low-frequency part of the eigenspectrum from Fig. 2a. Plots (c, d): The same results as presented in Figs. 2a, b, respectively, but for the signal to noise ratio 1:4.*

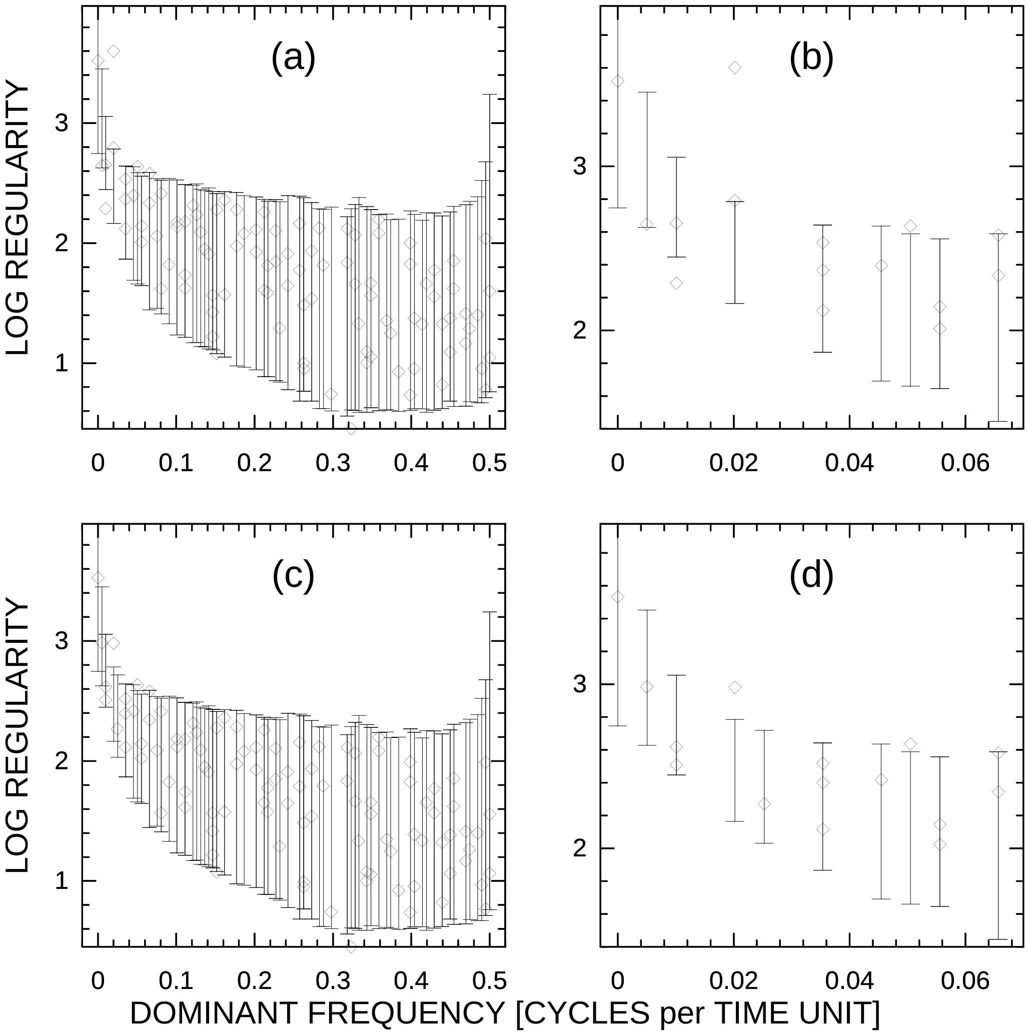

Figure 3: *Enhanced MCSSA analysis of the numerical data, presented in Fig. 1. Plots (a, b): Signal to noise ratio 1:2. (a) Logarithms of the regularity index ("LOG REGULARITY") plotted according to the dominant frequency associated with particular modes. Diamonds – regularity indices for the analysed data; bars – 95% of the surrogate regularity indices distribution, i.e., the bar is drawn from the 2.5th to the 97.5th percentiles of the surrogate regularity indices distribution. (b) Low-frequency part of the regularity indices spectrum from Fig. 3a. Plots (c, d): The same results as presented in Figs. 3a, b, respectively, but for the signal to noise ratio 1:4.*

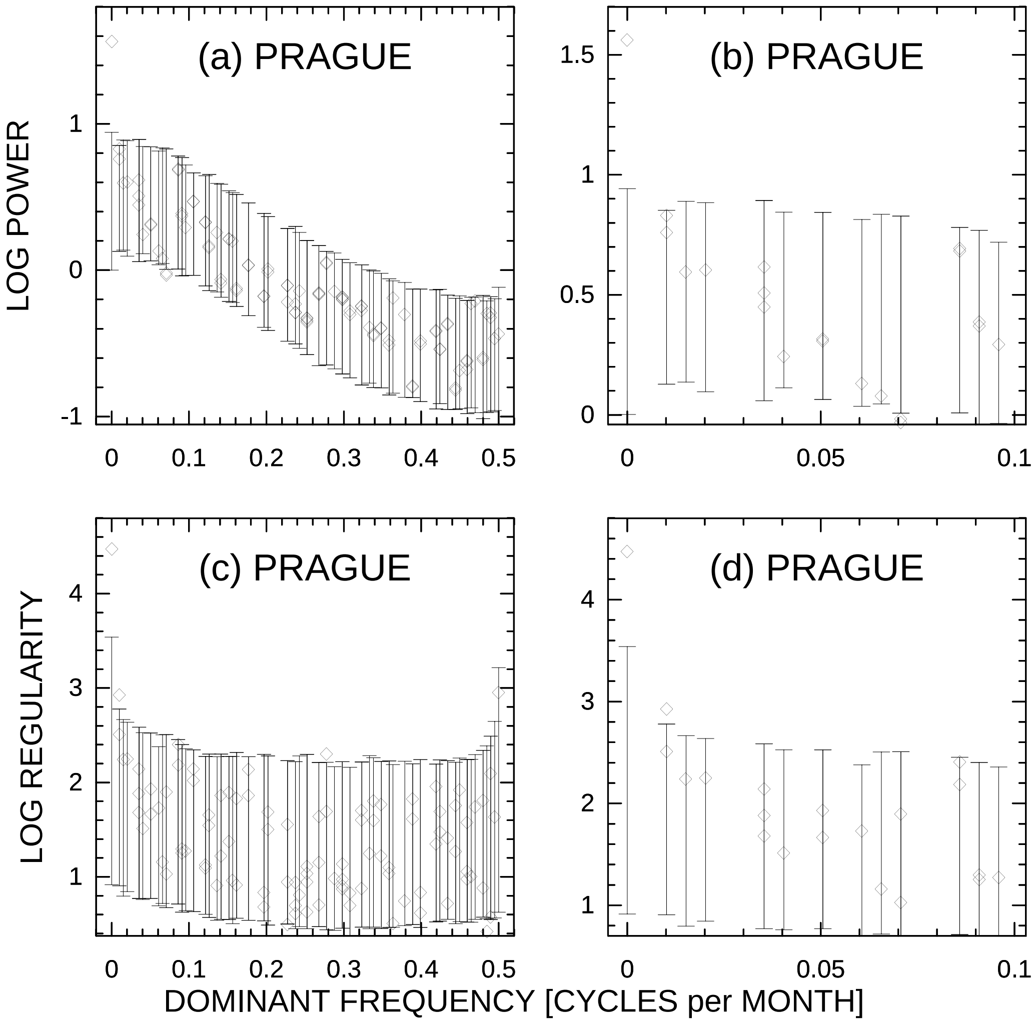

Figure 4: *Enhanced MCSSA analysis of the Prague surface air temperature series. (a) Eigenspectrum, plotted in the same way as the eigenspectra in Fig. 2, (b) low-frequency part of the eigenspectrum from Fig. 4a. (c) Regularity indices spectrum, plotted in the same way as the regularity indices spectra in Fig. 3, (d) low-frequency part of the spectrum from Fig. 3c.*

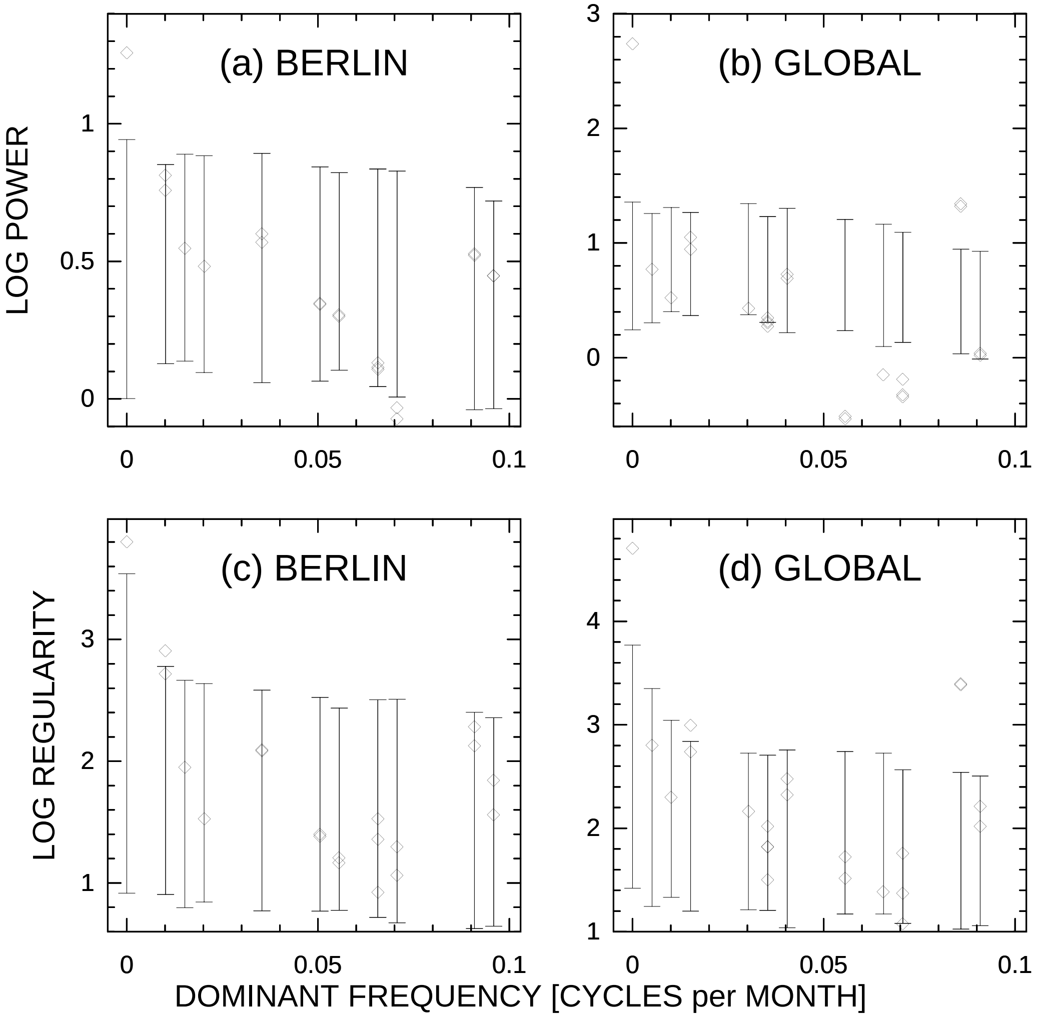

Figure 5: *Enhanced MCSSA analysis of the Berlin surface air temperature series (a,c) and the Jones global surface air temperature series (b,d), low-frequency parts of the spectra. (a, b) Eigenspectra, (c, d) regularity indices spectra.*