Scenarios Generation using Bootstrap in the Multichannel Singular Spectrum Analysis Approach and PAR (P) Structures: Application to Affluent Natural Energy

Moisés Lima de Menezes
Statistics Dept.
Fluminense Federal University
Niterói - Brazil

Reinaldo Castro Souza
Industrial Engineering Dept.
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro - Brazil

José Francisco Pessanha
Institute of Mathematical and Statistcs
State University of Rio de Janeiro
Rio de Janeiro - Brazil

ABSTRACT

The periodic autoregressive model (PAR (p)) becomes a powerful tool when to need generate scenarios. The NEWAVE and GEVAZP models in PAR (p) structures use the lognormal distribution to obtain scenarios using synthetic time series. Singular Spectrum Analysis (SSA) is a powerful statistical tool. SSA can decompose a time series into three components: trend, harmonics and noise and smoothing the series, removing the noisy component. Multichannel Singular Spectrum Analysis (MSSA) is a multivariate version of SSA for more than one time series simultaneously. This paper proposes the use of the bootstrap in noisy time series detected by MSSA for the generation of scenarios in the PAR (p) model for many time series smoothed by SSA and MSSA. Scenarios are generated with the original time series as well as the smoothed time series. Affluent Natural Energy (ANE) times series are used to illustrate the propose.

General Terms

Time series filtering, Multichannel Singular Spectrum Analysis

Keywords

Scenarios generation, MSSA, bootstrap, PAR (p) model, time series

1. INTRODUCTION

The use of probabilistic criteria in the various planning activities generates the need to use adequate modelling for forecasting. A simulation of the operation generates several results, among them the risk indexes. However, what exists is only one scenario: the record observed in the past (historical series). The historical record is insufficient to provide adequate performance indexes. A historical series follows a stochastic process. The stochastic model is estimated on this series and, from this model, to try obtain new achievements of the stochastic process that generated it. These realizations are like synthetic series that are statistically indistinguishable from the historical record. In this planning phase, a first-order probability model is performed. In several sectors, this modelling is done through the periodic autoregressive models PAR (p) with the objective of obtaining characteristics of the historical series to produce synthetic series. In Brazilian Interconnected System (BIS), these series are generated by Newave program-based [1] and by Gevazp (Generation Model of Synthetic Series of Energy and Flow) [2].

Singular Spectrum Analysis (SSA) is a powerful method in statistics with analysis elements of classical time series, multivariate statistics, multivariate geometry, dynamic systems, and signal processing [3]. SSA can be applied in fields like mathematics and physics, economics and financial mathematics, meteorology and oceanography, or social sciences [4]. The first works on SSA were published by [5], but the methodology became well known after publication by [3] and by [6] wherein the authors described the SSA procedure in details. Based on singular value decomposition (SVD), the SSA method decomposes a time series into a sum of components and each of these concentrates a part of the energy contained in the time series. In addition, a small group of components contributes to most of the autocorrelation structure contained in the time series, while the remaining may be considered noise components. Therefore, there are two mutually exclusive groups obtained: signal and noise. The first group comprises the components with information about the series structure whereas the latter covers the noise components. Thus, a smoothed version of the time series can be obtained by adding the signal components. Traditionally, a time series can be expressed as the sum of trend, harmonic and noise Components. Noise component removal may contribute to a better identification of the underlying stochastic process of the time series and, consequently, add accuracy to predictions. [7] showed that the removal of noise in hydrological series improves forecast accuracy. [8] have used SSA associated to PAR(p) to model wind speed series, and [9] and [10] presented SSA filtering results in a series of electricity consumption, the second associated with wavelets. A good review on this matter is presented by [11].

Multichannel singular spectrum analysis (MSSA) is a natural extension of SSA for multivariate time series. SSA and MSSA procedures provide a method of data pre-processing. Therefore, both methods can be used to improve the set of prediction models and help improve the accuracy of predictions [12] [13].

This paper proposes a generation of time series scenarios using PAR (p) structures after the SSA and MSSA filter. For this procedure, the bootstrap process was used to re-input the noise removed in the SSA and MSSA approach and decreasing the standard deviation in the scenarios generated. For analyse the accuracy of PAR (p) model, the statistics Mean Absolute Percentage Error (MAPE) was used. For testing the independence of the noise component removed in SSA and MSSA procedure, the BDS test [14] was used. For testing the equality of average between the original time series and the generated scenarios, the t-test was used and for testing the equality of standard deviation between the original time series and the generated scenarios, the Levene's test [15] was used.

For modelling PAR(p), scenarios generations, tests on scenarios and bootstrap procedure, the software MATLAB was used. For analysis in sample and adherence statistics MAPE, the software Forecast Pro for Windows (PFW) was used. For BDS independence test, the software Eviews was used. For SSA and MSSA filtering, the software CaterpillarSSA was used, and for graphical generations and descriptive statistics, the software Microsoft Excel was used.

This paper is subdivided as follow. In the section 2, the SSA procedure is presented. In the section 3, the MSSA approach is presented. The section 4 presents the PAR(p) structures. The section 5 presents the scenarios generation. The application in ANE is presented in section 6, and the conclusions are presented in section 7.

2. SINGULAR SPECTRUM ANALYSIS

The basic SSA method consists of two complementary stages: decomposition and reconstruction. In the first stage, an original time series is decomposed and in the second one the original series is reconstructed as a less noisy series and its used for modelling and forecasting. One of the main concepts in the study of the goodness of SSA is the concept of "separability", which characterizes how well different components can be separated by others. This concept is verified by the weighted correlation matrix. It is worth noting that, despite some statistical and probabilistic concepts employed in SSA-based methods, it is not necessary to keep in touch with statistics such as stationary of the series or normality of the residues.

2.1 Step 1: Decomposition

The decomposition step involves two stages: embedding and singular value decomposition (SVD).

Embedding is a procedure in which a time series $y_T = (y_1, \dots, y_T)$ is mapped into a sequence of lagged vectors in a matrix choosing a window length L such that $2 \leq L \leq T$. The question of the optimal value of L remains open. [16] illustrates a long discussion about the ideal value of window length if this value can be fixed or variable. In several cases, the general recommendation is to choose the window length at slightly less than half the size of the series: $T/3 \leq L \leq T/2$. To choose the window length, the BDS test is applied to the noise series after decomposition. First, $L = (T+1)/2$ is used if T is odd, or $L = T/2$ if T is pair. If the null hypothesis of independence of noise series in BDS test is rejected at 5% significance, then $L_1 = L - 1$ is used and redo the process in loop until the null hypothesis is not rejected. So, the embedding procedure takes y_T into a $L \times K$ trajectory matrix $\boldsymbol{X}$, such that

$$\boldsymbol{X} = \left(x_{ij}\right)_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & \cdots & y_K \\ y_2 & y_3 & \cdots & y_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_T \end{pmatrix}. \quad (1)$$

The $\boldsymbol{X}$ matrix in (1) is characterized by Hankel Matrix where all elements in all diagonals such that $i + j = constants$ are equals.

In SVD stage, the trajectory matrix in (1) is expanded by decomposition as in (2):

$$\boldsymbol{X} = \boldsymbol{E}_1 + \boldsymbol{E}_2 + \cdots + \boldsymbol{E}_L, \quad (2)$$

where $\boldsymbol{E}_l = \lambda_l^{1/2} U_l V_l', l = 1, \dots, L,$ are elementary matrices, the sequence $\lambda_1, \lambda_2, \dots, \lambda_L$ are the eigenvalues of the positive semidefinite matrix $\boldsymbol{S} = \boldsymbol{XX'}$ in order of significance such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0$, U_l denotes the respective eigenvectors, and $V_l = X'U_l/\sqrt{\lambda_l}$ are the principal components. The matrices $\boldsymbol{E}_l$ are called singular values, $(\lambda_l)_{l=1}^{L}$ is called singular spectrum, and the vectors U_l and V_l are called left and right singular vectors of $\boldsymbol{X}$. Let d be the rank of the trajectory matrix $\boldsymbol{X}$ (i.e., the number of nonzero eigenvalues), then the identity described in (2) can be rewritten as in (3):

$$\boldsymbol{X} = \boldsymbol{E}_1 + \boldsymbol{E}_2 + \cdots + \boldsymbol{E}_d, \quad (3)$$

where $d \leq L$. The collection (λ_l, U_l, V_l) is called eigentriple of SVD of the trajectory matrix $\boldsymbol{X}$. The contribution of each component in (2) can be measured by the ratio of singular values, given by $\sqrt{\lambda_l}/\sum_{l=1}^{L}\left(\sqrt{\lambda_l}\right)$ for each.

2.2 Step 2: Reconstruction

The reconstruction step also has two stages: grouping and diagonal averaging.

Grouping is a procedure that groups the elementary matrices into $n \leq d$ disjoints groups and adding the matrices within each group. Let $I_i = \left(I_{i_1} I_{i_2}, \dots, I_{i_p}\right)$ be the set of indices of the p_i elementary matrices classified in a same group i. Then the matrix corresponding to the group i is defined as $\boldsymbol{E}_{I_i} = \sum_{j=1}^{p_i} \boldsymbol{E}_{I_{ij}}$, such that $\{\boldsymbol{E}_l\}_{l=1}^{L} = U_{i=1}^{n}\left\{\boldsymbol{E}_{I_{ij}}\right\}_{j=1}^{p_i}$, so the identity (3) can be rewritten as (4):

$$\boldsymbol{X} = \boldsymbol{E}_{I_1} + \boldsymbol{E}_{I_2} + \cdots + \boldsymbol{E}_{I_n}, \quad (4)$$

where $n \leq d$.

Consider the trajectory matrix $\boldsymbol{X}$ and assume that $L^* = \min(L, K)$ and $K^* = \max(L, K)$. Consider that $e_{l,k}^{(i)}$ is an element in the row l and the column k of matrix $\boldsymbol{E}_{I_i}$. The element $y_t^{(i)}$ of SSA component $\left[y_t^{(i)}\right]_{1\times T}$ is computed by the Diagonal Averaging procedure applied to the matrix $\boldsymbol{E}_{I_i}$ as in (5)

$$y_t^{(i)} = \begin{cases} \dfrac{\sum_{l=1}^{t} e_{l,t-l+1}^{(i)}}{t}, for\ 1 \leq t < L^* \\ \dfrac{\sum_{l=1}^{L^*} e_{l,t-l+1}^{(i)}}{L^*}, for\ L^* \leq t < K^* \\ \dfrac{\sum_{l=t-K^*+1}^{T-K^*+1} e_{l,t-l+1}^{(i)}}{T-K^*+1}, for\ K^* \leq t \leq T \end{cases}. \quad (5)$$

Each SSA component $\left[y_t^{(i)}\right]_{1\times T}$ concentrates part of the energy of the original series $[y_t]_{1\times T}$ which can be measured by the ratio of singular values $\sqrt{\lambda_l}/\sum_{l=1}^{L}\left(\sqrt{\lambda_l}\right)$. According to [6], the SSA component $\left[y_t^{(i)}\right]_{1\times T}$ can be classified into three categories: trend, harmonic components (cycle and seasonality) and noise.

3. MULTICHANNEL SINGULAR SPECTRUM ANALYSIS

MSSA is an extension of SSA for analysis and forecasting of multidimensional time series. MSSA follows the same structure of SSA, with the difference of using a set of time series versus a single series. Consider system of Z time series of length T in (6)

$$y^{(z)} = \left(y_t^{(z)}\right)_{t=1}^{T}, \tag{6}$$

where $z = 1, \dots, Z$. The case of MSSA for $z = 1$ is equivalent to SSA procedure [17]. By choosing a single lag window length L for all series, the MSSA embedding stage takes $K = T - L + 1$ lagged vectors $X_k^{(z)} = \left(y_k^{(z)}, \dots, y_{k+L-1}^{(z)}\right), k = 1, \dots, K$ for each series $y^{(z)}, z = 1, \dots, Z$. Therefore, for each series $y^{(z)}$a trajectory matrix can be found as in (7)

$$\boldsymbol{X}^{(z)} = \begin{pmatrix} y_1^{(z)} & y_2^{(z)} & \cdots & y_K^{(z)} \\ y_2^{(z)} & y_3^{(z)} & \cdots & y_{K+1}^{(z)} \\ \vdots & \vdots & \ddots & \vdots \\ y_L^{(z)} & y_{L+1}^{(z)} & \cdots & y_T^{(z)} \end{pmatrix}. \tag{7}$$

The trajectory matrix of multidimensional series $\left(y^{(1)}, y^{(2)}, \dots, y^{(Z)}\right)$ is a $LZ \times K$ dimensional matrix with the equation (8)

$$\boldsymbol{X} = \left[X_1^{(1)} : \dots : X_K^{(1)} : \dots : X_1^{(Z)} : \dots : X_K^{(Z)}\right]' = \left[\boldsymbol{X}^{(1)} : \dots : \boldsymbol{X}^{(Z)}\right]'.$$

Trajectory space is defined by a linear space comprising the lagged vectors (columns of trajectory matrix $\boldsymbol{X}$). After embedding stage in MSSA process, all the steps follow as the steps in SSA procedure and after the diagonal averaging, the reconstructed series are obtained.

4. PERIODIC AUTRREGRESSIVE MODELS

According to [18] some time series have an autocorrelation structure that does not exist in the time interval between observations, but also of the observed period. These series can be analysed by autoregressive formulations whose periodic results are analysed. These formulations are a periodic autoregressive model PAR (p), where p is a vector represented by $p = (p_1, p_2, \dots, p_s)$ where s is the period considered as 12 if data are monthly. The PAR (p) model can be represented through standardizing observations of model AR (p) model as in (9)

$$\left(\frac{Y_t - \mu_m}{\sigma_m}\right) = \varphi_1^m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) + \cdots + \varphi_{p_m}^m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right), \tag{9}$$

where Y_t is the s-periodic seasonal time series, $s = 12$, $t = 1, \dots, T$, $m = 1, \dots, s$, μ_m is the seasonal average of period s, p_m is the order of autoregressive operator of period m, and a_t^m is the residual series independents and identically distributed with average zero and variance $\sigma_a^{2(m)}$. For each month, an order p_m autoregressive model is adjusted.

To identify the orders p_m of PAR (p) models, the autocorrelations function (ACF) are computed. The autocorrelation between Y_t and Y_{t-k} for each period m is obtained by ρ_k^m as in (10)

$$\rho_k^m = E\left[\left(\frac{Y_t - \mu_m}{\sigma_m}\right)\left(\frac{Y_{t-k} - \mu_{m-k}}{\sigma_{m-k}}\right)\right], \quad k = 1, 2, \dots \tag{10}$$

By multiplying (9) by $\left(\frac{Y_{t-k} - \mu_{m-k}}{\sigma_{m-k}}\right)$, and calculating the expectation, the expression in (11) is obtained.

$$\rho_k^m = E\left[\left(\frac{Y_t - \mu_m}{\sigma_m}\right)\left(\frac{Y_{t-k} - \mu_{m-k}}{\sigma_{m-k}}\right)\right] = \varphi_1^m E\left[\left(\frac{Y_t - \mu_{m-1}}{\sigma_{m-1}}\right)\left(\frac{Y_{t-k} - \mu_{m-k}}{\sigma_{m-k}}\right)\right] + \cdots + E\left[a_t^m \left(\frac{Y_{t-k} - \mu_{m-k}}{\sigma_{m-k}}\right)\right].$$

For each period m, from (11), the Yule-Walker periodic equations in (12) is obtained.

$$\begin{bmatrix} 1 & \rho_1^{m-1} & \cdots & \rho_{p_m-1}^{m-1} \\ \rho_1^{m-1} & 1 & & \rho_{p_m-2}^{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p_m-1}^{m-1} & \rho_{p_m-2}^{m-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \varphi_1^m \\ \varphi_2^m \\ \vdots \\ \varphi_{p_m}^m \end{bmatrix} = \begin{bmatrix} \rho_1^m \\ \rho_2^m \\ \vdots \\ \rho_{p_m}^m \end{bmatrix}.$$

Let φ_{kk}^m be a partial autocorrelation function (PACF) of period m. Each partial autocorrelation coefficient of order k coincides with the last parameter of an autoregressive model of the same order. So, in an autoregressive process of order p_m, the PACF is nonzero for $k \leq p_m$, and zero for $k > p_m$. The classical identify of PAR (p) model is based on obtain the appropriate orders p_m for the autoregressive operators according to estimate of $\hat{\varphi}_{kk}^m, k = 1, \dots, T/4$ according to Quenouille approximation [19] when $k > p_m$. The estimators of the parameters of PAR (p) model are obtained by moments method. These parameters are as efficient as the maximum likelihood estimators [18].

(8)

5. SCENARIOS GENERATION

The concepts, mathematical formulation, and statistics used to scenarios generation presented in this section are based in the implementation of Newave model [1]. The historic data in the instant t can be interpreted as a sample value associated to random variable of stochastic process in t. The goal of scenarios generation is fitter the synthetic series to represent the possible time series in the samples of process. Once not exist all occurrences of the stochastic process, the objective of adjusting the PAR (p) model so that it is a generator of the historical series and, with output, generate synthetic series that represent as possible temporal series of samples for the process. The adjusted PAR model should then allow for as many draws as are necessary for the problem in question. Thus, each draw is associated with a synthetic series.

Mathematically, from equation (9), when manipulated to isolate Y_t become as in (13). But, Y_t needs be positive, so (13) expresses an inequality.

$$Y_t = \mu_m + \varphi_1^m \sigma_m \left(\frac{Y_t - \mu_{m-1}}{\sigma_{m-1}}\right) + \cdots + \varphi_{p_m}^m \sigma_m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + \sigma_m a_t^m > 0. \tag{13}$$

In other words, (13) becomes (14).

$$a_t^m > \Delta, \tag{14}$$

where Δ is a function of both the first two moments of the period and the autoregressive coefficients and it is given by (15)

$$\Delta = -\frac{\mu_m}{\sigma_m} - \varphi_1^m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) - \cdots - \varphi_{p_m}^m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right).$$

Many researchers assume that normal distribution systems, and a possible non-normality, can be corrected by Box-Cox Transformation [20]. The synthetic series generation model should be applied directly to the original time series without transformation to make it stationary and developing the capacity to handle wastes that know a strong coefficient of asymmetry. For this, it was adopted the adjustment of a log-normal distribution with three permissions for monthly residues a_t^m [21]. So, the random variable ξ_t follow a Normal distribution with mean μ_{ξ_t} and variance $\sigma_{\xi_t}^{2(m)}$, $a_t = e^{\xi_t} + \Delta$, and $a_t \sim LNormal(\mu_{\xi_t}, \sigma_{\xi_t}^{2(m)}, \Delta$. So $\xi_t = \ln\,(a_t - \Delta)$ [22].

After the scenarios generate, the model performance evaluation is applied to these and some statistics tests are applied to this set of synthetics series. The synthetic scenarios obtained based on the proposed method must be able to reproduce the statistical properties of the original series. According to [1], the usefulness of a model can be measured by its ability to reproduce probability distributions of random variables relevant to the process. Are the tests:

a. t-Test for mean difference. Preservation of historical averages in the scenarios is one of fundamental assumptions to evaluate the adequacy of the model. According to [23], if two independent populations follow a Normal distribution, so the t-Test is used to test the equality of its averages.
b. Levene's test for equality between variances. The procedure to test the equality between variances of the historical series and the scenarios drawn through the Levene test [15].
c. Adherence tests. The adherence testes are the non-parametric tests to verify type of distribution. The principal idea of these tests consists in exams the level of concordance between the sample distribution and the populational one. The adherence tests used for the scenarios were the Kolmogorov-Smirnov test [24] [25].
d. Negative sequence analysis. According to [1], the main characteristics of the observed series must be preserved by the synthetic series generation model. Thus, one can measure the usefulness of a model by its ability to produce probability distributions of random variables relevant to the process. The random variables introduced here are related to the representation of critical periods. Thus, the negative sequence concept is used. For time-series analyses, a negative sequence can be defined as a period in which the inflows are continuously below the predetermined values preceded and succeeded by values above these limits. In general, monthly averages are used.

From this concept, three variables of interest can be cited: length, sum and sequence intensity. These variables can be defined according to the Table 1.

Table 1. Variables for negative sequence analysis

Variable	Description	Calculation
Sequence length	Interval of negative sequence $[t_1; t_2]$	$L_s = t_2 - t_1$
Sequence sum	Corresponds to the area below the boundary during the sequence	$S_s = \sum_{i=t_1}^{t_2} (Y_i - \mu_i)$
Sequence intensity	average value below limits given by sum over length	$I_s = \frac{S_s}{L_s}$

According to [26], the performance of the model can be evaluated by the proportion of indexes generated larger or smaller than the historical index. If this proportion is very small, there is an indication that historical observation is atypical for the model considered. For this analysis, the following variables can be considered: maximum sequence length, maximum sequence sum and maximum sequence strength.

6. METHODOLOGY

The procedure starts with the choice of the initial L parameter in SSA procedure. According to Literature, the best value of L is between $T/3$ and $T/2$. For the initial choice, the criterion that if T is even is used, then initial L is equal to $T/2$ and if T is odd, the initial L is $(T+1)/2$. From this definition, one proceeds with the SSA / MSSA filtering of the series. Once MSSA works with SSA filtering of several series simultaneously maintaining its dependency structure, then it was done to reduce series noise.

From the SSA / MSSA procedure, a signal component (S) composed of the trend and harmonic components and a noise component (R) is obtained for each series involved. The R component is tested via BDS and, if the test does not reject the null hypothesis of independence, then the R component is classified as noise and is removed from the series so that it can be modelled via PAR (p). If the DBS test rejects the null hypothesis, then a new value of L is estimated by subtracting the previous value from one unit: $L_i = L_{i-1} - 1$, where $i = 2, \ldots$ it indicates the number of attempts to estimate this parameter. This procedure is repeated as many times as necessary until the BDS test does not reject the null hypothesis.

In this way, the series are filtered through MSSA and via SSA and the approximate less noisy filtered series are obtained. With the filtered series, the PAR (p) methodology is applied in order to obtain the transformed model in (16):

$$\left(\frac{\hat{Y}_t - \hat{\mu}_m}{\hat{\sigma}_m}\right) = \varphi_1^m \left(\frac{\hat{Y}_{t-1} - \hat{\mu}_{m-1}}{\hat{\sigma}_{m-1}}\right) + \cdots + \varphi_{p_m}^m \left(\frac{\hat{Y}_{t-p_m} - \hat{\mu}_{m-p_m}}{\hat{\sigma}_{m-p_m}}\right) + a_t^m, \qquad (16)$$

where $\hat{Y}_t$ is the seasonal series filtered via SSA / MSSA period $s = 12,\ t = 1, \ldots, T,\ m = 1, \ldots, s,\ \hat{\mu}_m$ is the filtered SSA / MSSA average, p_m is the order of the autoregressive operator of period m - in this case, the order changes according to the period, and a_t^m is the series of independent residues and identically distributed with zero mean and variance $\sigma_a^{2(m)}$. Thus, the PAR (p) - MSSA / SSA model is determined.

The goal of scenarios generation is obtaining 5,000 scenarios in a horizon of 5 years (60 month). When the series are filtered through the SSA / MSSA approach, they become softer than original series, because they are without the noisy components. This makes the standard deviation of the generated scenarios well below the standard deviations of the historical series. To avoid this difference in behaviour, the noise removed in the SSA / MSSA process is reincorporated into the scenarios using the Bootstrap technique as follows: The series of noises of length T is redistributed month by month to obtain series of monthly noises of length T / 12. Then, the random draw of the 5 years of noise of these T / 12 data and the same replicated 5,000 times, thus obtaining 5,000 series of noises of length 60. For each month, the commands by MATLAB are:

```
data=csvread('month.csv');
          matrix=nan(5000;5);
                    for i=1:5000

indicesDrawn=randi(years, 5,1);
          dataDrawn=data(indiceDrawn);
                    matrix(i,:)=dataDrawn;
end.
```

Scenarios are considered for the original series modeled by PAR (p), for the series filtered through SSA and for the series filtered through MSSA. The generated scenarios are composed of two parts, one generated from the lognormal distribution of the PAR (p) - MSSA / SSA residuals, and another from the inclusion of the noise removed in the SSA / MSSA process through the Bootstrap process.

7. APPLICATION ON AFFLUENT NATURAL ENERGY

The application of scenario generation occurs in the affluent natural energy (ANE) series in Brazil, where more than 80% of the electric energy generated is of hydraulic origin [27]. The Electric System National Operator (ONS) works with the national interconnected system (SIN) that maintains information on power generation subdivided into four subsystems: North (N), Northeast (NE), Southeast / Center-West (SE) and South (S). The data presented are monthly averages of ANE from January 1931 to December 2012. This application is a part of results of [28]. Figure 1 shows the behaviour of the four-original series.

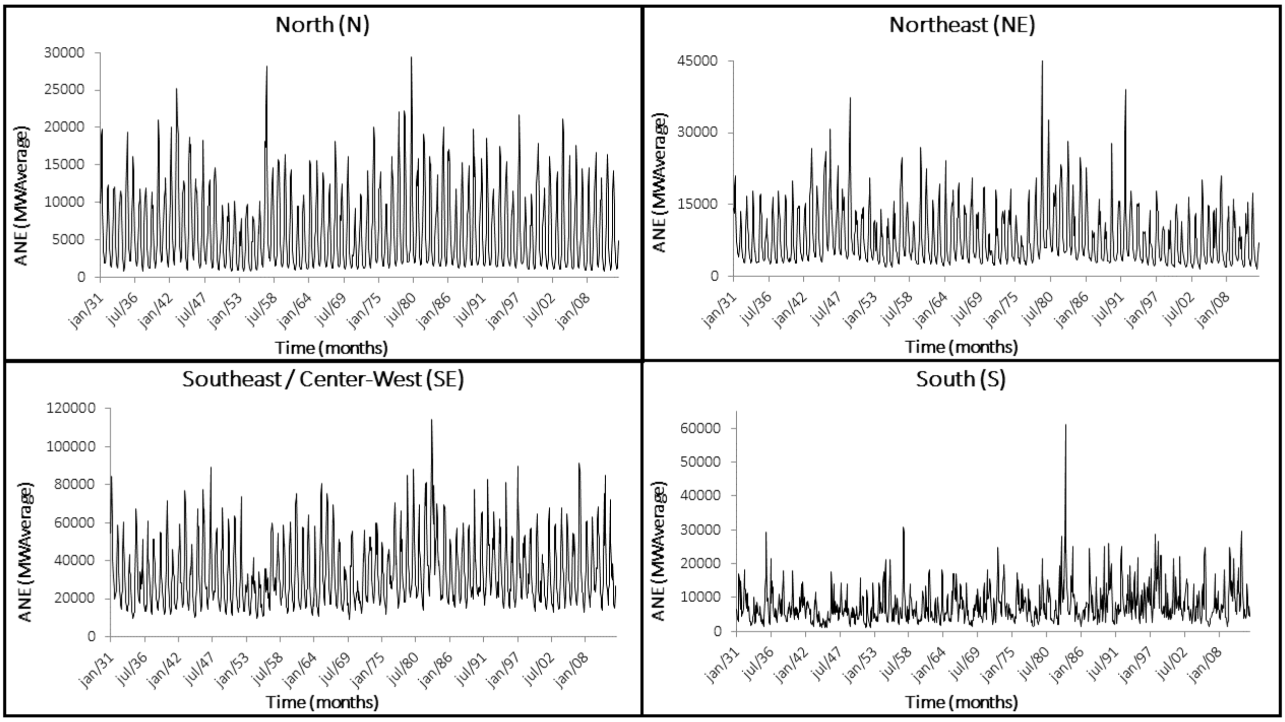

Fig 1: Originals series of monthly average of ANE from the four subsystems

Aiming to know statistically the series of the four ANE subsystems studied, the Table 2 shows the principal descriptive statistics from the four subsystems.

Table 2. Descriptive statistics from (MWAverage) Jan/1931 – Dec/2012

Descriptive Statistics	Subsystems			
	N	NE	SE	S
Length	984	984	984	984
Mean	6,276.42	8,175.30	32,993.65	7,753.11
Median	4,158.10	5,849.71	27,353.94	6,339.31
Minimum	736.53	1,443.00	9,115.39	992.44
Maximum	29,424.81	46,262.92	114,307.50	61,043.77
Std Dev	5,250.08	5,989.95	17,246.56	5,444.24
Skewness	1.08	1.57	1.08	2.26
kurtosis	0.56	3.45	0.82	10.99

Before PAR (p) modelling for forecasting and scenario generation, the four-subsystem series undergo MSSA-based filtering. For the MSSA case, the multidimensional trajectory matrix obtained in the embedding phase has dimension $(4K \times L)$. Once each series has length $T = 984$, the window length chosen is $L = 492$ according to the pre-established criteria. Thus, the size of the trajectory matrix is $(1{,}968 \times 493)$. In this way, the L eigenvalues and their eigenvectors U_l are obtained, as well as the principal components V_l. Consequently, there are 492 eigenvectors with dimension $4L \times 1 = 1{,}968 \times 1$, and 493 principal components with dimension $K \times 1 = 493 \times 1$. The first nine singular vectors and their contributions to the series are given in Figure 2, and Figure 3 are known some singular vector pairs (scatterplots) that can contribute with a graphical analysis.

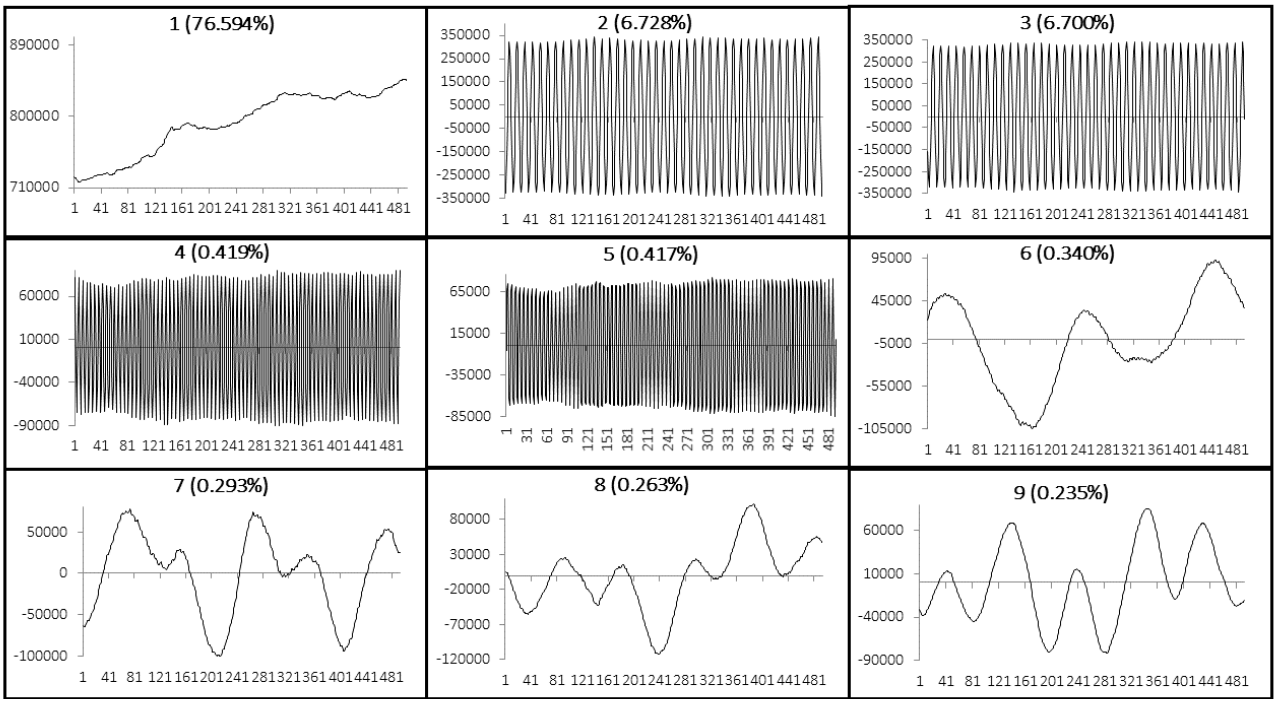

Fig 2: First nine singular vectors (principal components) in MSSA procedure

Because the soft behaviour, the singular vectors 1, 6, 7, 8, and 9 make up the trend component, and the singular vectors 2, 3, 4, and 5, make up the harmonic component because the sinusoidal behaviour. Another way of noting that a singular vector belongs to the harmonic component is the scatterplots between two vectors in sequence as in figure 3. If the scatterplots show a regular polygon or circumference, the two vectors belong to the harmonic component.

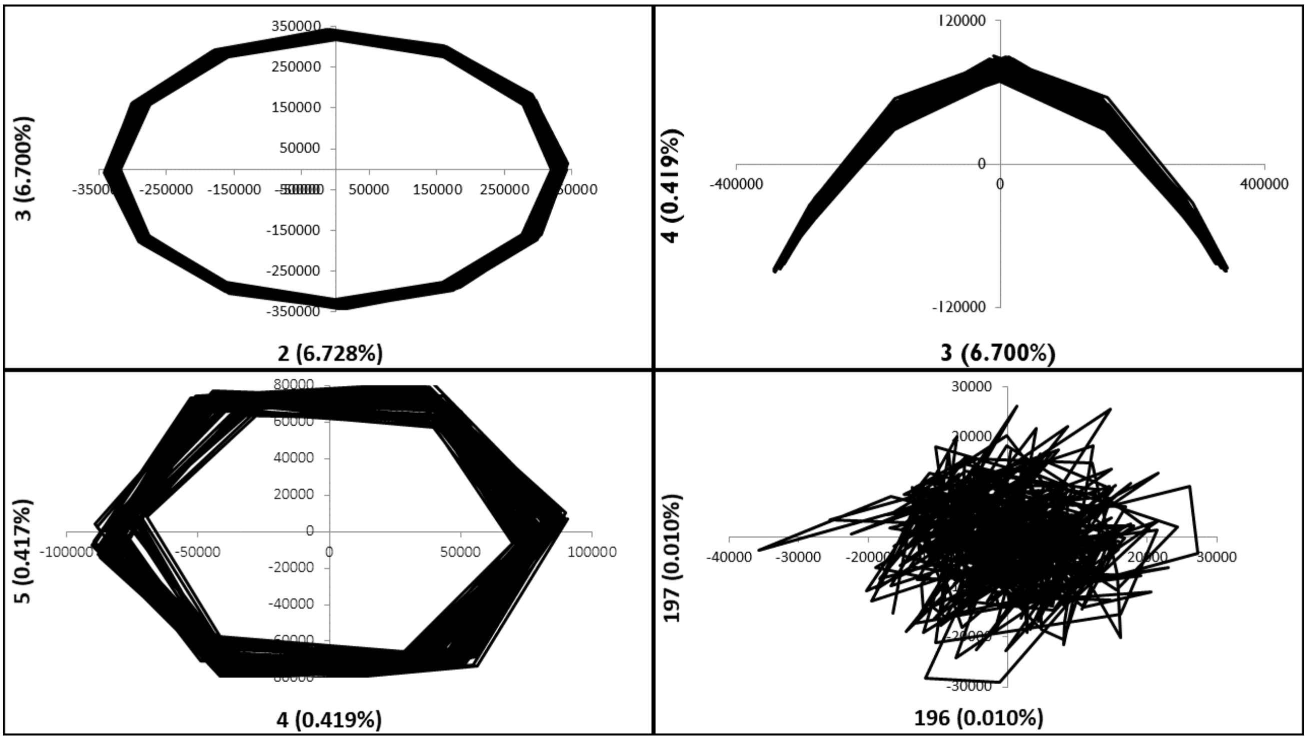

Fig 3: Some scatterplots in MSSA procedure

In figure 3 the pairs (2,3) and (4,5) present the shape of a 12-sided polygon, in the first case and with 6-sided, in the second case. Both belong to the harmonic component. The pair (3,4) presents the behaviour of the scatterplots of two vectors belonging to the harmonic component, but with different periods. Finally, the pair (196,197) presents the chaotic behaviour of two vectors belonging to the noisy component.

After graphical analysis of singular vectors, the components were grouped as in Table 3.

Table 3. Principal components decomposed in MSSA

Component	Singular Vectors
Trend	1, 6 - 11
Harmonic	2 – 5, 12 - 19
Noise	20 - 492

Figures 4 – 6 present the three components for each ANE time series decomposed by MSSA.

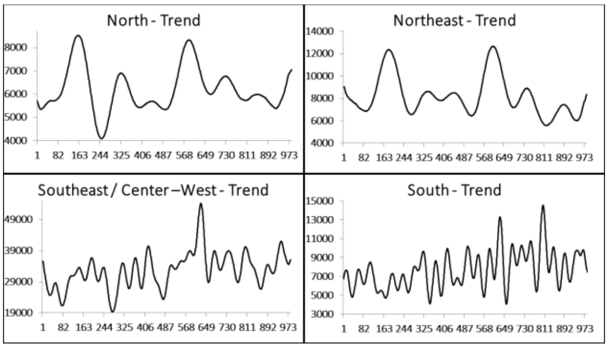

Fig 4: Trend component for the four subsystems in MSSA procedure

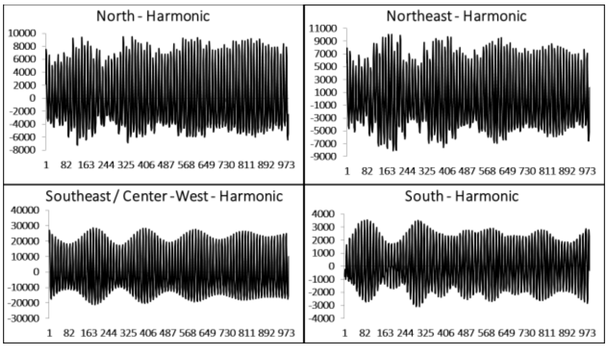

Fig 5: Harmonic component for the four subsystems in MSSA procedure

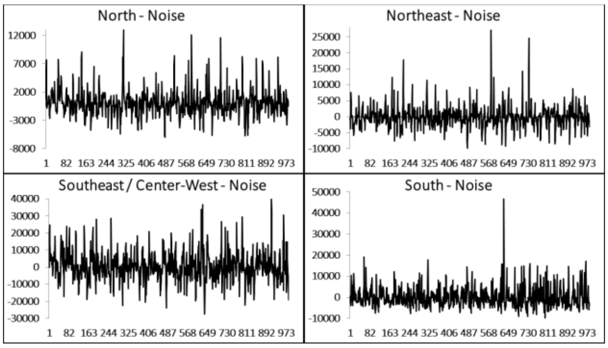

Fig 6: Noise component for the four subsystems in MSSA procedure

The weighted correlation shown in Table 4 confirm that the components are well separated.

Table 4. Weighted correlation between the three components in MSSA procedure

Components	Trend	Harmonic	Noise
Trend	1	0.001	0.011
Harmonic	0.001	1	0.042
Noise	0.011	0.042	1

For a verification purpose and to prove that a noisy component represents a noise, the BDS test was applied on components in each subsystem. Table 5 presents the results of the BDS test applied to the noise series of the four subsystems. The results in this table show that the independence hypothesis is not rejected at the 5% level of significance in all measured dimensions, proving to be noisy as components evaluated in the test.

Table 5. The BDS test for the four noisy series in MSSA procedure

Dim	p-value			
	North	Northeast	Southeast	South
2	0.9697	0.8950	0.2889	0.9793
3	0.9864	0.9201	0.9775	0.9585
4	0.9840	0.8668	0.9992	0.9504
5	0.9821	0.8441	0.9999	0.9452
6	0.9234	0.8415	0.9999	0.9375

After this, the noisy component was removed and the filtered time series via MSSA were obtained. The PAR(p) model was fitted on originals time series and on the filtered time series. The adherence statistics MAPE shown in Table 6, show that MSSA improve the accuracy of model, once the values of MAPE to the filtered series fitted are less than the original series fitted.

Table 6. Adherence statistics MAPE (%) of PAR(p) models before and after filter MSSA

Month	North		Northeast		Southeast		South	
	PAR(p)	PAR(p) MSSA	PAR(p)	PAR(p) MSSA	PAR(p)	PAR(p) MSSA	PAR(p)	PAR(p) MSSA
Jan	21.31	**4.05**	20.09	**4.07**	19.10	**3.59**	43.86	**6.15**
Feb	22.49	**3.25**	29.17	**4.50**	20.43	**3.71**	42.42	**5.49**
Mar	14.01	**2.09**	24.62	**4.42**	17.95	**3.22**	31.30	**8.83**
Apr	14.15	**2.33**	29.20	**5.01**	12.33	**3.29**	45.13	**14.77**
May	12.96	**2.88**	14.68	**6.90**	10.38	**4.66**	79.80	**8.08**
Jun	10.51	**5.56**	7.27	10.93	10.44	**5.17**	55.22	**5.11**
Jul	6.83	**4.96**	5.19	12.67	8.01	**5.50**	46.30	**5.20**
Aug	5.74	10.35	4.52	12.35	9.24	**7.15**	67.32	**4.39**
Sep	8.30	21.18	6.77	11.94	14.19	**8.09**	51.13	**4.22**
Oct	13.29	**9.51**	13.62	**5.99**	18.25	**5.98**	48.87	**3.23**
Nov	19.93	**5.79**	24.59	**7.52**	13.68	**5.53**	12.13	**4.48**
Dec	22.55	**5.52**	25.19	**4.29**	15.42	**4.61**	46.57	**7.26**

For scenarios generation, 5,000 synthetic series were generated for a 5-year horizon (60 months). Scenarios were considered for the original series modelled by PAR (p) and for the series modelled via PAR (p) - MSSA. In order to evaluate if the synthetic series have statistical behaviour identical to the historical series, the mean and standard deviation of the scenarios generated from the PAR (p) and PAR (p) - MSSA models were calculated and to obtain the five years of the historical series, the mean and standard deviation of each month in this series were calculated and, thus, generated a series year of historical averages and the same for the standard deviation. To obtain the 60 months required for comparison with the synthetic series, the series of historical means and standard deviations was replicated 5 times.

With the MSSA filtering, the standard deviations of the scenarios generated from the PAR (p) - MSSA model are very below the standard deviations of the original historical series. Therefore, to compare the standard deviations obtained through this modelling, it is necessary to add the noise series drawn in the MSSA filtering to the scenarios generated in this modelling. Thus, the standard deviations, which in principle were much lower, approximate the equivalents obtained in the historical series. To include these noises in the scenarios, the 984-length noise series was redistributed month by month in order to obtain 82 series of monthly noises. Then, a random draw of the 5 years of noise carried out through the Bootstrap process. This procedure carried out 5,000 times. This yields 5,000 series of noises of length 60 following the same probability distribution as the original series of noise extracted in the MSSA filtering. These noises are added to the scenarios and their standard deviations are compared with the historical standard deviations and the scenarios generated from the PAR (p) modelling. According to [28], the southern subsystem has its own characteristic and deserves to be filtered separately, as well as its scenario generation. Thus, with the MSSA filter in the other subsystems, the southern subsystem was filtered separately from the others via SSA. With these MSSA / SSA filtering results, the PAR (p) model was adjusted and the scenarios were generated based on the Bootstrap process to redistribution of the noisy series. In this new situation, the historical monthly averages and the monthly standard deviations can be compared with the averages of the scenarios generated. The analysis of model’s performance with the southern subsystem filtered together with the other subsystems can be seen in [28]. Tables 7-10 show the results of the Levene, t, and negative sequences tests for the scenarios generated with the original series and the MSSA-filtered series considering filtering of southern subsystem separately.

Table 7. Percentages of non-rejection of the null hypothesis of equality between averages in the t-Student test compared to the historical series

Subsystem	Does not reject null hypothesis (%)	
	PAR(p)	PAR(p) - MSSA
North	100	85
Northeast	99	100
Southeast	100	94
South	97	99

Table 8. Percentages of non-rejection of the null hypothesis of equality between variances in the Levene test compared to the historical series

Subsystem	Does not reject null hypothesis (%)	
	PAR(p)	PAR(p) - MSSA
North	100	97
Northeast	99	97
Southeast	100	97
South	97	97

Table 9. Negative sequence test for PAR (p) modelling

Subsystem	Length Critical value: 3.84	Sum p-value min: 0.05	Intensity p-value min: 0.0.5
North	2.33	0.99812	0.25267
Northeast	0.86	0.80479	0.65092
Southeast	0.18	0.83240	0.77302
South	0.02	0.96319	0.03753

Table 10. Negative sequence test for PAR (p) - MSSA modelling

Subsystem	Length Critical value: 3.84	Sum p-value min: 0.05	Intensity p-value min: 0.0.5
North	1.06	0.76024	0.70471
Northeast	0.16	0.01220	0.07203
Southeast	0.03	0.37662	0.23918
South	0.80	0.10420	0.41750

The results presented in tables 7 – 10 shows that the capacity of the scenarios generated reproduce the historical series behaviour is verified by using the MSSA approach before modelling PAR(p) and Bootstrap in noise component, such that from 85% to 100% of scenarios tested by t-Student test not reject the null hypothesis of equality between the averages of historical series and fitted scenario. The same information can be observed when the Levene test is used to testing the equality between the variances of historical series and the scenarios fitted. In this case, 97% of scenarios not reject the null hypothesis of equality. From the negative sequence test, the results in table 10 show that the scenarios generated can reproduce critical periods of drought. The results presented in tables 7 – 10 shows that the capacity of the scenarios generated reproduce the historical series behaviour is verified by using the MSSA approach before modelling PAR(p), such that from 85% to 100% of scenarios tested by t-Student test don't reject the null hypothesis of equality between the averages of historical series and of the fitted scenarios. The same information can be observed when the Levene test is used to testing the equality between the variances of the historical series and of the scenarios generated. In this case, 97% of scenarios not reject the null hypothesis of equality. From the negative sequence test, the results in tables 9, and 10 shows that the scenarios generated can reproduce critical periods of drought with or without the filter MSSA. This result show that the scenarios generate are adequate.

8. CONCLUSIONS

This paper has presented a combined modelling of the periodic autoregressive models with Multichannel Singular Spectrum Analysis as well as the choice of the best L parameter in MSSA using the BDS test and a proposal of synthetic scenarios generation based on the Bootstrap approach. In order to evaluate the performance of PAR (p) models with and without MSSA filtering, the MAPE adherence statistic was used and to test the quality of the generated scenarios, several tests such as Levene and negative sequence tests were performed to verify the ability of the generated scenarios of reproduce the behaviour of the historical series.

In order to verify this approach, an application to data of affluent natural energy carried out and in it 5000 scenarios were generated for a horizon of 60 months. The MAPE statistic has shown that the use of the MSSA filtering improved the accuracy of the fitted model, so it which it is a good approach to be used. As far as scenario generation is concerned, it was clear that Bootstrap is necessary since filtering the series does not reproduce the standard deviation of the historical series if the MSSA filtering is applied and the Bootstrap Is not performed and that the scenarios generated in this way, both with, and without the MSSA filtering, present adequate behavior according to the applied tests.

9. REFERENCES

[1] CEPEL. Newave Model Reference Manual (in Portuguese). Technical Report. 1st ed. Rio de Janeiro: CEPEL; 2001. 102 p.

[2] Maceira, M. E. P., Penna, D. D. J. Synthetics scenarios generation of energy for operational planning (in Portuguese). In: BSWR, editor. XVI Brazilian Symposium of Water Resources; 20 - 24 November; João Pessoa, Brazil. 2005.

[3] Elsner, J. B., Tsonis, A. Singular spectrum analysis. A new tool in time series analysis. 2nd ed. New York and London: Plenum Press; 2010. 164 p.

[4] Hassani, H. Singular spectrum analysis: methodology and comparison. Journal of Data Science. 2007; 5: 239 - 257.

[5] Broomhead, D. S., King, G. P. Extracting qualitative dynamics from exponential data. Physica D. 1986; 20: 217 - 236.

[6] Golyandina, N., Nekrutkin, V., Zhigljavsky, A. Analysis of time series structure: SSA and related techniques. 1st ed. New York: Chapman & Hall/CRC; 2001. 305 p.

[7] Jayawardena, A. W., Gurung, A. B. Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach. Journal of Hydrology. 2000; 228: 242 - 264.

[8] Menezes, M. L., Souza, R. C., Pessanha, J. F. M. Combining singular spectrum analysis and PAR(p) structures to model wind speed time series. Journal of Systems Science and Complexity. 2014; 27: 29 - 46.

[9] Menezes, M. L., Souza, R. C., Pessanha, J. F. M. Electricity consumption forecasting using singular spectrum analysis. DYNA. 2015;82(190):138-146.

[10] Teixeira Jr., L. A. T., Menezes, M. L. Cassiano, K. M., Pessanha, J. F. M., Souza, R. C. Residential electricity consumption forecasting using a geometric combination approach. International Journal of Energy and Statistics. 2013;1(2):113-125. DOI: 10.1142/S2335680413500087

[11] Elshorbagy. A, Simonovic, S. P., Panu, U. S. Noise reduction in chaotic hydrologic times series:facts and doubts. Journal of Hydrology. 2002; 256: 147 - 265.

[12] Hassani, H., Heravi, S., Zhigljavsky, A. Forecasting European industrial production with singular spectrum analysis. International Journal of Forecasting. 2009; 25: 103 - 118.

[13] Hassani, H., Mahmoudvand, R. Multivariate singular spectrum analysis: a general view and new vector forecasting approach. International Journal of Energy and Statistics. 2013;1(1):55-83.

[14] Brock, W. A., Dechert, W., Scheinkman, J., LeBaron, B. A test for independence based on the correlation dimension. Econometric Reviews. 1996;15(3):197-235.

[15] Levene, H. Robust tests for equality of variances. In: Contributions to probability and statistics: essays in honour of Harold Hotteling. Stanford: Stanford University press.; 1960. p. 278-292.

[16] Golyandina, N. On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. 2010; 3: 259 - 279.

[17] Golyandina, N. Stepanov, D. SSA-based approaches to analysis and forecast of multidimensional time series. In: Fifth Workshop on Simulation; 2005; St. Petersburg. St. Petersburg: Department of Mathematics, St. Petersburg State University; 2005. p. 293-298.

[18] Hipel, K. W., McLeod, A. I. Time Series Modelling of Water Resources and Environmental Systems. 1st ed. Amsterdam: Elsevier; 1994.

[19] Quenouille, M. Approximate tests of correlation in time series. Mathematical Proceedings of the Cambridge Philosophical Society. 1949;45(3):483-484.

[20] Box, G. E. P., Cox, D. R. An analysis of transformations. Journal of the Royal Statistical Society. 1964; 26: 211 - 252.

[21] Maceira, M. E. P. Optimal Reservoir Operation with Affluent Forecasting (in portuguese) [dissertation]. Rio de Janeiro:1989.

[22] Charbeneau, R. J. Comparison of the two- and three-parameter log normal distributions used in stream of synthesis. Water Resources Research. 1978; 14: 149-150.

[23] Casela, G., Berger, R. L. Statistics Inference (in Portuguese). São Paulo: Cengage Learning; 2010.

[24] Kolmogorov, A. Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari. 1933; 4: 83-91.

[25] Smirnov, N. Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics. 1948; 19: 279 - 281.

[26] Souza, R. M. Modeling of Periodic Series via PAR (p) Structure using Wavelet Shrinkage (in Portuguese) [thesis]. Rio de Janeiro: Pontifical Catholic University of Rio de Janeiro; 2013. 109 p.

[27] ONS - Electric System National Operator. Daily Operation Preliminary Information (Updated Daily) [Internet]. 01/01/2012 [Updated: 09/23/2012]. Available from: http://www.ons.org.br/publicacao/ipdo/Ano_2012/IPDO-23-09-2012.pdf [Accessed: 09/23/2012].

[28] Menezes, M. L. PAR (p) and Singular Spectrum Analysis Approach in the Modelling and Scenarios Generation (in Portuguese) [thesis]. Rio de Janeiro: Pontifical Catholic University of Rio de Janeiro; 2014. 126 p. Available from: https://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=23300@2.