2010 International Symposium on Nonlinear Theory and its Applications
NOLTA2010, Krakow, Poland, September 5-8, 2010

Equatorial Climate Data Analysis and Forecasting by Singular Spectrum Analysis

Naoki Itoh† and Jürgen Kurths

†Center for Dynamics of Complex Systems
University of Potsdam, D-14476 Potsdam, Germany
Email: naoki.itoh@gmail.com

Abstract—This paper describes the techniques of Singular Spectrum Analysis (SSA) and forecasting by the Linear Recurrent Formulae (LRF), which are applied to monthly precipitation and lake-sediments in Kenya. By the SSA algorithm, information, such as a trend, seasonal periodicities, anomaly cycles, and noise, can be extracted from these data series. And then from the results obtained by the SSA, it is possible to forecast the data which are assumed to be governed by the LRF. The goal in this paper is to investigate the properties of the equatorial climate changes, where the global warming is remarkable.

1. Introduction

Climate problems through the term global warming have received a lot of attention in recent years. Interest in the topics spans not only meteorologist and geologist but also physicist, economist and so on as interdisciplinary subjects.

Data observed in nature generally consist of complicated components such as exogenous and endogenous factors. These data series have successfully been analyzed by a number of statistical tools. If such noisy data are, for instance, analyzed for a forecast, the noise has to be reduced from it. In other words, the underlying deterministic dynamics in the data will be extracted by the noise reduction. In some of the previous research the modelling and forecasting have often been performed by a linear model. However, since most of the actual data series are currently well-known as a nonlinear nature, it is necessary to consider both the linear and nonlinear models for the modelling and forecasting. As one of the ideal methods, Singular Spectrum Analysis (SSA) is powerful and useful, and it is especially applicable for the analysis of time series with complex seasonal components and non-stationarity, i.e., it is not necessary to assume satationarity of the series or normality of the residuals. The technique is defined as a nonparametric technique of the time series analysis including the statistical tools such as the classical analysis, dynamical system, signal processing, and so on. Another advantage of this method is that it can well be applied to small sample sizes.

An early study of the SSA is described in the papers by Broomhead and King [1]. Then, the idea of the SSA is independently developed in several groups in Russia, UK, and USA. Especially, the theoretical and practical foundations of this technique are described in the book by Golyandina, Nekrutkin, and Zhigljavsky in the Russian group (2001) [5].

The purpose of SSA is to decompose the original data series into some components with useful and interpretable information (e.g. a slow trend, oscillatory components, and a structureless noise). Their decomposed components are of substantial importance for time series forecasting by the Linear Recurrent Formulae (LRF).

In this paper the Caterpillar-SSA proposed by Golyandina et al. (2001), is applied to data about precipitation and lake-sediments in Kenya in order to compare present climate changes with a paleoclimate and to forecast the new data point by using the LRF [2–7, 9–16].

2. Singular Spectrum Analysis

SSA aims to decompose the observed data series into some meaningful subseries, which are in general identified as a slowly varying trend, harmonic (periodic and quasi-periodic) components, and noise. These kinds of components show essential properties of observed data.

The algorithm of the method generally falls into two stages, the first stage: decomposition and the second stage: reconstruction, and then each of them has two following separate steps: embedding (step 1) and Singular Value Decomposition (SVD) (step 2) in the first stage, and grouping (step 1) and diagonal averaging (step 2) in the second stage. The reconstructed data series are then used for forecasting.

2.1. First Stage: Decomposition

Embedding of the step 1 in the first stage is to transfer a one-dimensional series $Y = (y_1, \cdots, y_N)$ into the multidimensional series $[X_1 : \cdots : X_K]$ with vectors $X_j = (y_j, \cdots, y_{j+L-1})^T \in \mathbb{R}^L$ $(j = 1, \cdots, K)$, where K and L, which are defined as the integer parameters for the SSA, can be described by $K = N - L + 1$, thus the SSA has substantially the single parameter L of the embedding, called window length, restricted by $2 \leq L \leq N/2$. The matrix which consists of the vectors X_j is defined as the trajectory matrix, $\mathrm{X} = [X_1 : \cdots : X_K] = (x_{ij})_{i,j=1}^{L,K}$. Since the trajectory matrix X is a Hankel matrix, all the elements along the diagonal $i + j = const$ are equal [8].

In the step 2, the Singular Value Decomposition (SVD) is applied to the trajectory matrix, which can then be written as $\mathrm{X} = \mathrm{X}_1 + \cdots + \mathrm{X}_d$, where $\mathrm{X}_l = U_l \sqrt{\lambda_l} V_l^T$ $(l = 1, \cdots, d)$ defined as a rank-one orthogonal elementary matrix. The collection $(\sqrt{\lambda_l}, U_l, V_l)$, which are a singular value, empirical orthogonal functions (EOFs), and principal components (PCs), respectively, is called the l-th eigentriple of the matrix X. d is the number of non-zero singular values (i.e., $\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_d} > 0$). The relationship among the terms in the collection can be described by $V_l = \mathrm{X}^T U_l / \sqrt{\lambda_l}$. The fact that $\sum_{l=1}^{d} (\sqrt{\lambda_l})^2$ is equal to the squared Frobenius-Perron norm of the trajectory matrix X, and also $(\sqrt{\lambda_l})^2$ is the squared Frobenius-Perron norm of the elementary matrix X_l [2, 5–7], means that the ratio $\sum_{l=1}^{r} \lambda_l / \sum_{l=1}^{d} \lambda_l$ measures the degree of approximation of the trajectory matrix, that is, it shows a contribution of the elementary matrices to the trajectory matrix.

2.2. Second Stage: Reconstruction

So called *eigentriple grouping* will be performed so that the index set $\{1, \cdots, d\}$ of the elementary matrices is reformulated into m disjoint subsets $I_1, \cdots, I_m$:

$$\mathrm{X} = \mathrm{X}_1 + \cdots + \mathrm{X}_d = \mathrm{X}_{I_1} + \cdots + \mathrm{X}_{I_h} + \cdots + \mathrm{X}_{I_m}, \qquad (1)$$

with $\mathrm{X}_{I_h} = \mathrm{X}_{h_1} + \cdots + \mathrm{X}_{h_s}$, $h \in \{1, \cdots, d\}$. These represented matrices are defined as m resultant matrices. In order to find a proper parameter r for the grouping, singular spectrum, $S = \{\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_d}\}$ and weighted-correlation (w-correlation), ρ^w are introduced:

$$\rho_{ij}^w = \frac{(Y^{(i)}, Y^{(j)})_w}{\|Y^{(i)}\|_w \|Y^{(j)}\|_w}, \qquad (2)$$

where $(Y^{(i)}, Y^{(j)})_w = \sum_{k=1}^{N} w_k y_k^{(i)} y_k^{(j)}$, $\|Y^{(i)}\|_w = \sqrt{(Y^{(i)}, Y^{(i)})_w}$ $(i, j = 1, \cdots, d)$, and w is defined by $w_k = \min(k, L, N - k + 1)$. The series $Y^{(i)} = \{y_1^{(i)}, \cdots, y_N^{(i)}\}$ is available from the elementary matrix by using the diagonal averaging which is defined as follows:

$$y_n^{(i)} = \begin{cases} \frac{1}{n} \sum_{m=1}^{n} x_{m,(n-m+1)}^{(i)} & (1 \leq n < L) \\ \frac{1}{L} \sum_{m=1}^{L} x_{m,(n-m+1)}^{(i)} & (L \leq n < K) \\ \frac{1}{N-n+1} \sum_{m=n-K+2}^{N-K+1} x_{m,(n-m+1)}^{(i)} & (K \leq n \leq N) \end{cases} . \qquad (3)$$

If the reconstructed series $Y^{(i)}$ and $Y^{(j)}$ are highly correlated, then they can be grouped into a same component. In contrast, if the w-correlation between these two series is quite low or zero, it means that they are well separable into different groups. The series grouped by the result of the w-correlation can be represented as a decomposed form of the initial series, $Y = Y^{(I_1)} + \cdots + Y^{(I_m)}$, where the labels of Is are equivalent to the m disjoint subsets in the r.h.s. of *eq.* (1).

2.3. Forecasting: Linear Recurrent Formulae

The SSA forecasting is in general started with the assumption that the data series is approximately governed by the Linear Recurrent Formulae (LRF) [4–7, 9–13, 15, 16]:

$$y_{i+d} = \sum_{k=1}^{d} a_k y_{i+d-k}, \quad 1 \leq i \leq N - d, \qquad (4)$$

where d is the dimension for forecasting and $a_1, \cdots, a_d$ are defined as coefficients for the LRF. Note that if the original data series Y satisfies an LRF, there exist at most d non-zero singular values $(\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_d} > 0)$ within a window length L. Therefore, in this forecasting technique it is necessary to prepare at most d elementary matrices X_i in order to reconstruct the series.

The SSA recurrent forecasting algorithm can be explained as follows: Recall the eigenvector $U \in \mathbb{R}^L$ computed in the SVD step. Let us denote that the vector of the first $L - 1$ components of the U as $U^\nabla \in \mathbb{R}^{L-1}$ and set $\nu^2 = \pi_1^2 + \cdots + \pi_r^2 < 1$ as a sum of the square of the last components $(\pi_i := u_{Li}, i = 1, \cdots, r)$ of U. It can be proved that $y_L = a_1 y_{L-1} + \cdots + a_{L-1} y_1$ where the coefficients

$$A = (a_1, \cdots, a_{L-1}) = \frac{1}{(1 - \nu^2)} \sum_{i=1}^{r} \pi_i U_i^\nabla. \qquad (5)$$

3. Application

The data analyzed and forecasted by the SSA in this study are precipitation and lake-sediments profile in Kenya to investigate the structure of climate in the equatorial zone of East Africa.

3.1. Data: Precipitation and Lake-Sediments

Precipitation shown in the figure 1, had been recorded at the stations of Kenyan three towns, Nakuru, Naivasha, and Narok, with different time length.

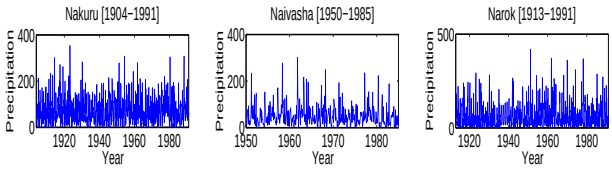

Figure 1: Monthly precipitation in Nakuru, [1904-1991] (left panel), in Naivasha, [1950-1985] (middle panel), and in Narok, [1913-1991] (right panel) from GHCN v2 database.

A lake-sediments profile had been taken from the lake Nakuru, which is shown in the figure 2 (top). The bottom one shows the color intensity of the profile. Depth of this profile is ca. 4.6 cm, which corresponds to the time length for ca. 63 years [17].

3.2. Analysis

The first mode component extracted by the SSA is in general a slowly-varying trend. The figure 3 shows that

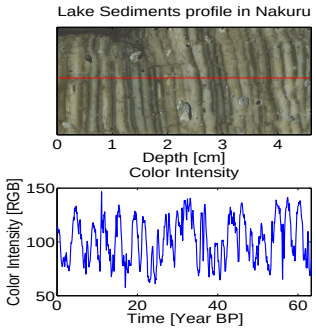

Figure 2: Lake-sediments profile from the lake Nakuru (top) and the color intensity of a layer shown on the red line in this profile (bottom). The depth is ca. 4.6 cm (for 63 years). It becomes deeper toward the right. The color intensity is supposed to be linked with rainfall variability [17].

the larger window length L, the slower the trend variation. The component can be shown as a smoothing for different purpose.

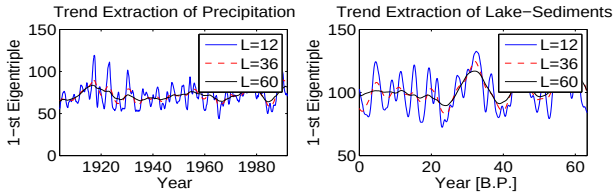

Figure 3: The trends of precipitation (left panel) and lake-sediments (right panel) in Nakuru for L = 12, 36, and 60. The longer L, the more slowly-varying is the trend.

Group	1	2	3	4	5	6
Nakuru	12	6	4	15	3	10
Naivasha	12	6	13-14	68	10	–
Narok	12	6	4	68	25	15
Sediments	46-47	35	27	23	17-19	13-15

Table 1: The groups with oscillation components for L =60. The values show monthly periodicities.

The table 1 lists the periodicity and quasi-periodicity of harmonics components from the precipitation and the lake-sediments.

From the results of precipitation, indeed seasonal cycles are dominant in all the data because such periodicities are shown in the first 3 groups. On the other hand, in other groups of the higher modes, several irregular cycles in an annual sense are shown. Since these results show partly common cycles (10, 15, and 68 months cycles), they may be considered as individual characteristics in this area. Although the rest of the components is in general assumed as noise, it still remains a matter of debate because they are not well interpreted.

The result obtained from the analysis of sediments is that the dominant periodicities are obviously longer than those of precipitation.

3.3. Forecast

The LRF technique for forecasting the new data points needs another parameter r in the *eq.* (5) which can be produced through the window length L, which will be used to control an accuracy of the forecast. In order to find such optimal parameters, let us measure the Mean Absolute Error (MAE := $\frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|$, where y_i is an original data, $\tilde{y}_i$ a forecasted value, and n the number of forecasted points.) between the forecasted value and the original data by varying the window length L.

In this paper, the window length parameter L will be varied from 12 to 120 and then let us define it as a parameter optimal in the sense of a prediction error when the MAE is minimal.

The results are depicted in the last 6 points of each data in the figure 4. All of them can approximately be forecasted. Their $\hat{L}$s are equal to 49 (Nakuru), 98 (Naivasha), 48 (Narok), and 120 (Lake-sediment), respectively. This means that both data complexities consist of the components of similar climate background. On the other hand, the value of $\hat{L}$ for Naivasha is relatively high. As a prime suspect, it would appear that the amount of data observed in Naivasha is smaller than in the other two towns. $\hat{L}$ of the sediments data results in an even higher value. This means that these sediments data have already been simplified, i.e., there is a possibility that some information vanished. Therefore, the forecast of the sediments data can be achieved with the small errors.

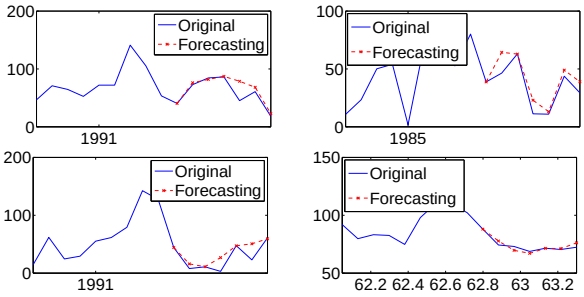

Figure 4: Forecasts; Nakuru for $\hat{L}$ = 49 (left-top panel), Naivasha for $\hat{L}$ = 98 (right-top panel), Narok for $\hat{L}$ = 48 (left-bottom panel), and Lake-sediment for $\hat{L}$ = 120 (right-bottom panel).

4. Conclusions

In this study precipitation and lake-sediments profile in Kenya have been analyzed and forecasted by using the

SSA.

The dominant information of these precipitation are some seasonal periodicities, 12 months, 6 months, and 4 months, which mean that cycles such as rainy and dry seasons are regularly repeated in a year. In particular, since there are two rainy seasons in Kenya (in Spring and Autumn), the interval corresponds to a 4 months cycle. As minor properties besides the above cycles, some irregular cycles can be found in the higher modes. These kinds of properties may be considered as cycles not belonging to the seasonal one. From the analysis of the lake-sediments profile, the results is that the dominant cycle is longer than that of precipitation.

In the SSA forecast a complexity of the observed data can be obtained by the parameter L. The precipitation in Nakuru and Narok can be forecasted by an almost equivalent optimal parameter $\hat{L}$. The lake-sediments data analyzed as a hint about a paleoclimate structure can be forecasted by a relatively large $\hat{L}$, which means that the structure of this data set is simpler than that of the other data sets.

Acknowledgments

The precipitation data from GHCN v2 database (`http://www.ncdc.noaa.gov/`) used in this study was provided by Norbert Marwan (PIK Potsdam), he and Udo Schwarz (Center for Dynamics of Complex Systems, University of Potsdam) gave me some quite valuable comments and suggestions. We would like to acknowledge the help of them.

References

[1] D.S. Broomhead and G.P. King, ”EXTRACTING QUALITATIVE DYNAMICS FROM EXPERIMENTAL DATA,” *Physica 20D*, pp. 217–236, 1986.

[2] H. Björnsson, S.A. Venegas, ”A Manual for EOF and SVD analyses of Climate Data,” *Centre for Climate and Global Change Research*, Report No. 97–1, pp.52, 1997.

[3] J.M. Wallace, C. Smith, and C.S. Bretherton, ”Singular Value Decomposition of Wintertime Sea Surface Temperature and 500-mb Height Anomalies,” *Journal of Climate*,vol.5, pp.561–577, Jun. 1992.

[4] D.L. Danilov, ”Principal Components in Time Series Forecast,” *Journal of Computational and Graphical Statistics*,,vol.6, pp.112–121, 1997.

[5] N. Golyandina, V. Nekrutkin and A. Zhigljavsky, ”Analysis of Time Series Structure,” *SSA and related techniques. Chapman & Hall/CRC.*, pp.303, 2001.

[6] N. Golyandina, and D. Stepanov, ”SSA-based approaches to analysis and forecast of multidimensional time series,” *the 5th St. Petersburg Workshop on Simulation*, pp.293–298, 2005.

[7] H. Hassani, ”Singular Spectrum Analysis: Methodology and Comparison,” *Journal of Data Science.*,vol.5, pp.239–257, 2007.

[8] J.L. Phillips, ”The Triangular Decomposition of Hankel Matrices,” *MATHEMATICS OF COMPUTATION*,vol.5, no.115, Jul. 1971.

[9] A. Serita, K. Hattori, C. Yoshino, M. Hayakawa, and N. Isezaki, ”Principal component analysis and singular spectrum analysis of ULF geomagnetic data associated with earthquakes,” *Natural Hazards and Earth System Sciences*,vol.5, pp.685–689, 2005.

[10] H. Hassani, A. Zhigljavsky, ”SINGULAR SPECTRUM ANALYSIS: METHODOLOGY AND APPLICATION TO ECONOMICS DATA,” *Jrl Syst Sci & Complexity, 22*, pp.372–394, 2009.

[11] S. Polukoshko, J. Hofmanis, ”USE OF ”CATERPILLAR”– SSA METHOD FOR ANALYSIS AND FORECASTING OF INDUSTRIAL AND ECONOMIC INDICATORS,” *the 7th International Scientific and Practical Conference*, vol. II, 2009.

[12] H. Hossein, A. Zhigljavsky, ”Singular Spectrum Analysis Based on the Minimum Variance Estimator,” *the World Congress on Engineering 2008*, vol. II, Jul. 2008.

[13] H. Hossein, S. Heravi, and A. Zhigljavsky, ”Forecasting European industrial production with singular spectrum analysis,” *International Journal of Forecasting*, 25, pp.103–118, 2009.

[14] N. Itoh, and J. Kurths, ”Singular Spectrum Analysis of Equatorial Precipitation Data,” *Nonlinear Theory and its Applications NOLTA 2009*, Oct., 2009.

[15] G. Stea, and M. Bilancia, ”Singular Spectrum Analysis: a new decomposition technique applied to environmental systems,” MTISD 2008–Methods, *Models and Information Technologies for Decision Support Systems*, Sep., 2008.

[16] D.H. Schoellhamer, ”Singular spectrum analysis for time series with missing data,” *Geophysical Research Letters*, vol.28, pp. 3187–3190, 2001.

[17] N. Marwan, A. Junginger, M. Trauth, A. Bergner, and Y. Garcin, ”Recurrence in climate variability–a comparison of modern climate data from Nakuru, Kenya, with Early Holocene paleo-climate records,” *General Assembly of EGU*, Poster, Apr., 2007.