Gap Filling of Precipitation Data by SSA - Singular Spectrum Analysis

A S F Filho, G A R Lima
Departamento de Engenharia Sanitária e Ambiental, Universidade Federal de Mato Grosso Av. Fernando Corrêa da Costa, nº 2367, 78060-900, Cuiabá, MT, Brazil
E-mail: garlima@ufmt.br

Abstract. From the macroscopic standpoint, the precipitation time series is obtained from observation of natural systems rather than in the laboratory. These time series are often full of gaps (missing values) due to the conditions under which the measurements are made. Missing values give rise to various problems in spectral estimation, inhibit statistical analysis and in specifying boundary conditions for numerical models. Hence, gap filling is necessary in environmental science. The aim of this study is to highlight the application of the SSA forecasting algorithm to fill in missing values to real-life time series. It was applied to several monthly precipitation time series recorded over a large savannah area in Brazil. The results are promising and the accuracy and reliability depend on the pattern and relative length of the gaps with respect to the total length of the time series and presence of noise.

1. Introduction
Although, precipitation long varying trend analyses on different spatial and temporal scales has been of great concern because of the attention given to global climate change [1], the worldwide number of active monitoring points is still quite low [2]. Furthermore, time series are affected by non-climatic factors caused by malfunctioning or maintenance of measuring equipment, changes in location of monitoring points, systematic errors and data transcription. As a consequence, time series generally contain many missing values. Missing values impose gradual bias and are the major source of problems in the analysis and modelling of spatial and temporal variability. In other words, missing values impose a limit on the quality of information that can be extracted from the observed values (records). Incomplete time series reveal only part of the reality. Examining the full extent of the precipitation time series, with the missing data filled in, allows for better significance testing in the spectral analysis because all the information contained in the time series can be extracted and the information quality maximized. In this sense, SSA is a widely used method of time series analysis [3-5] and filling gaps [6-10]. State-of-the-art gap filling algorithms based on the SSA exploit the information contained in periodic temporal patterns to fill in the observations. Today there are three SSA based approaches for filling in missing values. The first approach was designed to deal with stationary time series only and preserved only pairs of valid vector components in the vectors’ inner products [6]. In the second approach the missing values are estimated iteratively by successive application of SSA and the complete time series are composed by imputation of estimated missing values. But the other values are taken from the initial time series [7]. The third approach is an extension of SSA forecasting algorithms known as Caterpillar-SSA where the main methodological
concept is the components’ separability. In this approach, neither a parametric model nor stationarity conditions have to be assumed for the time series. [8]. Following the last approach, in this paper we highlight the application of the algorithm to fill gaps in precipitation time series recorded at wide savannah area in Brazil.

2. Methodology - SSA Algorithm
The SSA algorithm decomposes the initial time series into a sum of a small number of interpretable components such as a slowly varying trend, seasonal, periodic or oscillating components, and also structureless components of noises. It is based on the singular value decomposition (SVD) of a trajectory matrix constructed upon the initial time series. The algorithm consists of two complementary stages: decomposition and reconstruction. Each stage, comprises two steps: Embedding and SVD are the steps of the first stage, grouping and diagonal averaging are the steps of the second stage. The extension of the SSA algorithm for the analysis of time series with missing values has the same structure, but the steps are somewhat different [11]. A brief description of the algorithm is presented below. A detailed discussion can be found in [11,12] and references therewith.

Let \( X = (x_1, ..., x_N) \) be a time series (input) consisting of \( N \) data points, some part of which is unknown (missing values).

2.1 First stage - Decomposition
Step one – Embedding. The embedding procedure transforms the initial time series into the sequence of \( L \)-dimensional lagged vectors \( \{X_i\}_{i=1}^K \) where \( K = N - L + 1 \). Some of the lagged vectors contain missing components. The window length \( L \) is the only parameter in the method. Let us fix the \( L, 1 < L < N \). Selection of the proper \( L \) depends on the problem in hand and on preliminarily information about the time series to be analyzed. Let \( C \) be the set of indices such that the lagged vectors \( X_i \) with \( i \in C \) are complete. Let us collect all complete lagged vectors \( X_i, i \in C, \) into the matrix \( \tilde{X} \). If there are no missing values, then the matrix \( \tilde{X} \) coincides with the trajectory matrix of the series \( X \).

Step two – SVD. Let \( \tilde{S} = \tilde{X}^T \tilde{X} \). Denote by \( \lambda_1 \geq \ldots \geq \lambda_d \geq 0 \) the ordered eigenvalues of the matrix \( \tilde{S} \) and by \( U_1, \ldots, U_d \) the orthonormal system of the eigenvectors of the matrix \( \tilde{S} \) corresponding to these eigenvalues, \( d = \max \{i : \lambda_i > 0\} \).

2.2 Second stage - Reconstruction
Step three a - Reconstruction stage. This stage demands the grouping to make subgroups of the decomposed trajectory matrices and diagonal averaging to reconstruct the new time series from the subgroups choosing the subspace and projection of the complete lagged vectors. Let a set of indices \( I_i = \{i_1, \ldots, i_r\} \subset \{1, \ldots, d\} \) be chosen and the subspace \( M_i = \text{span}(U_{i_1}, \ldots, U_{i_r}) \) be formed. The choice of the eigenvectors (i.e., their indices) corresponding to \( X_i \) is the same as in Basic SSA. The complete lagged vectors can be projected onto the subspace \( M_i \) in the usual way:

\[
\hat{X}_i = \sum_{k \in I_i} (X_i, U_k) U_k, \quad i \in C
\]

Step three b - Projection of the incomplete lagged vectors. For each \( Q \)-incomplete lagged vector with missing components in the positions from the set \( Q \), the given step consists of two parts: (a) calculation of \( \hat{X}_i \) \( i \in C, \) and (b) calculation of \( \hat{X}_i \) \( i \in C. \) Since adjacent lagged vectors have common information which enables processing of empty vectors with \( Q = J_i = \{1, \ldots, L\}. \) Note that Step three b may change the vectors \( \hat{X}_i, i \in C. \) The result of Steps 3a and 3b is the matrix
\( \hat{X} = [\hat{X}_1 : \ldots : \hat{X}_k] \) which serves as an approximation to the trajectory matrix of the series \( X_i' \), under the proper choice of the set \( I_i \).

Step four - Diagonal averaging. In the last step of the algorithm, the matrix \( \hat{X} \) is transformed into the new series \( Y_{w}^{(1)} \) (the reconstructed time series) by means of the diagonal averaging with a simultaneous filling in of the missing data.

The algorithm was applied to the longest available precipitation time series in order to fill gaps and identify slowly varying trends and periodic components from 31 monitoring points located at Alto Médio Cuiabá River Basin as is illustrated in figure 1. The source of data is available in the Hidroweb geodatabase [13].

![Figure 1. Spatial and temporal distribution of monitoring points at Alto-Médio-Cuiabá River Basin](image)

2.3 Gap-filling procedures - As noise is produced by error of measurements or uncertainties, in accordance with [9] when the distribution of gaps follows a seasonal pattern, it generates spurious, intermittent peaks in the reconstructed time series. In the case of precipitation records, during the reconstruction stage, the spectral method also gives rise to negative values for imputation of some missing values in the dry season. To avoid this, the idea is thus to fill in missing values of small time series formed by breaking the initial time series into 12 sub-time series by means of a delay (or lag with \( k=12 \)) according to each month of the calendar year. We consider reconstruction using signal plus noise components by the successive application of the SSA algorithm for each month. Then all predicted values are inserted in the initial time series where the values are missing. Thus a complete time series is obtained as seen in [7]. As imputation means to substitute missing values with plausible values for a stated purpose and all the predicted missing values fall in the range of minimum and maximum records for each hydrologic year of each time series, the ANOVA-LR (Analysis of Variance of Linear Regression Analysis) between complete and reconstructed time series was selected as a measure of “goodness of fit” between the models and time series (confidence interval 95%). Finally, the complete time series was again analyzed to illustrate the capability of the algorithm to extract a long varying trend, and periodic and noise components. All of the results were obtained by means of CatMV1.0 and Caterpillar-SSA 3.30 software. The window lengths were assigned by following the theoretical recommendation [14].

3. Results

Although the study was carried on for 31 time series, the results are shown for only four. These four time series have the highest number of missing values and all types of gaps uniformly distributed (figure 2 (a)), seasonal (figures 2 (b) and (c)) and prolonged gaps (figure 2 (d)). These are also found in
the other 27 times series. Figure 2 shows the summarized results of the decomposition stage for the initial time series. Visual inspection shows that independently of the type of gaps, series length and noise quantity, the algorithm is capable of extracting the slowing varying trend and seasonality, which correspond to the start and the end of each hydrological year.

**Figure 2.** Initial precipitation time series and decomposition after SVD

Figure 3 shows the results of the decomposition stage for the complete time series after the imputation of missing values. Visual inspection of figures 3 (b) and (c) shows new extracted periodic components [4-5], which correspond to the start and the end of the rainy and dry seasons respectively.

**Figure 3.** Complete precipitation time series and its decomposition after SVD
Complete time series tend to have more information about oscillations and noise reduction. Figure 4 shows the $w$-correlation matrix for the first 14 reconstructed components of the complete time series in a 20-grade scale from white to black corresponding to the absolute values of correlations from 0 to 1. It is possible to see the approximate separability of groups such as the slowly varying trend represented by eigentriple $|1\rangle$, seasonal components by eigentripe $|2-3\rangle$ and eigentripe $|4-5\rangle$ and noise by others eigentriples $|6 ... k\rangle$. This means that the selection of the proper window length allows extraction (separation) of all the components of the series and the filling in of missing values at the reconstruction stage.

![Figure 4. Matrix of w-correlations for the first 14 reconstructed components](image)

A relative measure of goodness fit is shown in Table 1. The correlation coefficient $R$ measures the degree of linear association between complete and reconstructed time series. The determination of the coefficient $R^2$ of the regression line indicates the percentage of those records in complete time series that can be explained by the records predicted during the reconstruction stage. In other words, more than 90% of the precipitation records can be explained by the linear regression and consequent gap-filling strategy.

| $N$  | $R$  | $R^2$   | $SD$ | $F$    | $p$-value |
|------|------|---------|------|--------|-----------|
| 1755003 | 251  | 0.986   | 0.9723 | 43.34 | 17839.45 | 0.000     |
| 1556002 | 1238 | 0.991   | 0.9820 | 59.94 | 19634.34 | 0.000     |
| 1556005 | 520  | 0.925   | 0.8556 | 44.62 | 3061.90  | 0.001     |
| 1556007 | 456  | 0.989   | 0.9781 | 16.32 | 20820.42 | 0.000     |
In Table 1, the column SD presents the standard deviation of the data in relation to the regression line (unexplainable variation) in the records. The $F$ represents the ratio of explained by unexplained variance and determines the overall statistical significance of the regression equation. As the $p$-value is less than the confidence level 5%, the regression is not rejected. According to LRA results, all time series were reconstructed successfully.

4. Conclusions
One of the conclusions is that the SSA caterpillar algorithm can deal with the non-stationary nature inherent in precipitation records by filling in missing values and in the simultaneous extraction of long varying trends and periodic components. The gap-filling strategy adopted while utilizing a conceptually simple and computationally efficient algorithm represents an adaptation of existing techniques to create an operational method that is applicable to precipitation time-series. In all cases, the adopted strategy prevented negative values and spurious, intermittent peaks in the reconstruction stage and was particularly helpful for situations with continuously missing data, which quickly become prohibitive, without breaking the initial time series by means of a delay (or lag with $k=12$). The results are promising but further analysis is required in order to address real-world limitations associated with sensitivity of types of gaps and the impact of noise on the liability of predicted missing values.

While the study was restricted to precipitation, the strategy could be readily adapted to a wide variety of environmental time-series and combined with the SSA caterpillar algorithm to have a very wide range of applicability.

Acknowledgements
A S F Filho thanks Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT) for a graduate fellowship and to John C. Carpenter for feedback on grammar and writing.

5. References
[1] Longobardi A and Villani P 2009 Int. J. Climatol. DOI: 10.1002/joc.2001
[2] Hulme M 1995 Weather 50, 34-42
[3] Vautard M Ghil M 1989 Physica D, 35, 395-424
[4] Allen M and Smith L 1996 J. Clim., 9, 3373-3404
[5] Vautard R You P and Ghil M 1992 Physica D, 58, 95-126
[6] Schoellhamer D 2001 Geophys. Res. Lett., 28, 16, 3187-3190
[7] Kondrashov D and Ghil M 2006 Nonlin. Processes Geophys. 13, 2, 151-159
[8] Golyandina N and Osipov E 2007 J. Stat. Plan. Infer 137, 8, 2642-2653
[9] Musial J P Vertraete M M Gobron N 2011 Atmos. Chem.Phys., 11, 7905-7923
[10] Butlta J V Zscheischler J and Mahecha M D 2014 Nonlin. Processes Geophys. 21, 203-215
[11] Golyandina N and Zhigljavsky A 2013 Singular Spectrum Analysis for Time Series Springer
[12] Golyandina N Nekrutkin V and Zhigljavsky A 2001 Analysis of Time Series Structure: SSA and related techniques, Chapman & Hall-CRC, New York - London
[13] http://http://www.snirh.gov.br/hidroweb/, August 2015
[14] Golyandina N 2010 Statistics and Its Interface, 3 (3), 259–279