Singular spectrum analysis for time series with missing data

David H. Schoellhamer

U.S. Geological Survey, Sacramento, California

Abstract. Geophysical time series often contain missing data, which prevents analysis with many signal processing and multivariate tools. A modification of singular spectrum analysis for time series with missing data is developed and successfully tested with synthetic and actual incomplete time series of suspended-sediment concentration from San Francisco Bay. This method also can be used to low pass filter incomplete time series.

Introduction

Singular spectrum analysis (SSA) is essentially a principal components analysis in the time domain that extracts information from short and noisy time series without prior knowledge of the dynamics affecting the time series [*Vautard and Ghil*, 1989; *Dettinger et al.*, 1995]. Geophysicists have used SSA to analyze a wide variety of time series such as solar oscillations [*Varadi et al.*, 1999], precipitation [*Prieto et al.*, 1999; *Wang et al.*, 1996], streamflow and sea-surface temperature [*Robertson and Mechoso*, 1998], chemical constituents of ice cores [*Yiou et al.*, 1997], global temperature [*Vautard et al.*, 1992; *Allen and Smith*, 1977], magnetosphere dynamics [*Sharma et al.*, 1993], and suspended-sediment concentration in an estuary [*Schoellhamer*, 1996]. One of the requirements of SSA is a continuous time series, but geophysical time series often contain missing data. The purpose of this paper is to present a modification to SSA that permits analysis of time series with missing data.

My motivation for this work is to develop a tool to analyze time series of suspended-sediment concentration (SSC) collected by automated optical instruments in San Francisco Bay every 15 minutes from 1991 to the present (2001) [*Buchanan and Schoellhamer*, 1999; *Buchanan and Ruhl*, 2000]. Typically, one-half of the data are invalidated by biological fouling, instrument malfunction, and vandalism. SSA has been applied to continuous blocks of data 14-70 days long and several tidal signals were identified, and the fortnightly spring/neap tidal cycle accounted for about one-half of the variance of SSC [*Schoellhamer*, 1996]. Modifying SSA to permit missing data eliminates the need to screen, fill, and subdivide time series prior to applying SSA and allows analysis of longer, but incomplete, time series. The SSC time series typically have large tidal and seasonal signals, so the problem of how to extract a signal from noisy data is not as important as the problem of extracting a signal from incomplete data.

Copyright 2001 by the American Geophysical Union.

Paper number 1999GL000000.
0094-8276/01/1999GL000000$05.00

Singular Spectrum Analysis

Vautard, Yiou, and Ghil [1992] describe singular spectrum analysis (SSA). For a standardized time series x_i, where sample index i varies from 1 to N, and a maximum lag (or window size) M, a Toeplitz lagged correlation matrix (each diagonal has a uniform value), is formed by

$$c_j = \frac{1}{N-j}\sum_{i=1}^{N-j} x_i x_{i+j} \qquad 0 \le j \le M-1 \tag{1}$$

The eigenvalues, λ_k, and eigenvectors (or empirical-orthogonal functions), E_j^k, of this matrix are determined and sorted in descending order of λ_k, where indices j and k vary from 1 to M. The kth principal component is

$$a_i^k = \sum_{j=1}^{M} x_{i+j} E_j^k \qquad 0 \le i \le N-M \tag{2}$$

Each component of the original time series identified by SSA can be reconstructed, with the kth reconstructed component (RC) series given by

$$x_i^k = \frac{1}{M}\sum_{j=1}^{M} a_{i-j}^k E_j^k \qquad M \le i \le N-M+1 \tag{3}$$

Expressions for x_i^k for $i < M$ and $i > N-M+1$ are given by *Vautard, Yiou, and Ghil* [1992]. The fraction of the total variance of the original time series (equal to one for standardized time series) contained in the kth RC is λ_k, so that, with the sorting used, the RCs are ordered by decreasing information about the original time series. Most of the variance is contained in the first several RCs and most or all of the remaining RCs contain noise. SSA typically decomposes a time series into RCs that are nearly periodic with periods less than M and one or two RCs contain variations in the time series with periods greater than M. A pair of RCs with similar λ_k typically represents each period less than M with significant energy in the original time series [*Vautard et al.*, 1992].

Singular Spectrum Analysis For Time Series with Missing Data

Singular spectrum analysis for time series with missing data (SSAM) alters the SSA calculation of lagged autocorrelation and principal components. An overbar in the following notation indicates that the series either contains missing data or is calculated from a series that contains missing data. First, the lagged autocorrelation is computed by ignoring any pair of data points with a missing value

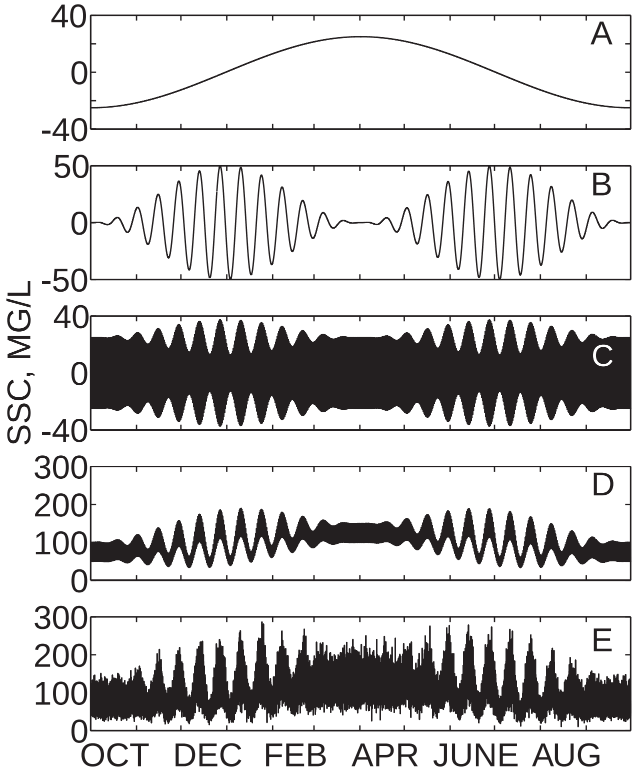

Figure 1. Synthetic time series of suspended-sediment concentration. A) Seasonal component, B) Fortnightly spring/neap component, C) Semidiurnal component, D) Synthetic signal $c_s(t)$, and E) Synthetic time series $c(t)$.

$$\bar{c}_j = \frac{1}{N_l}\sum_{l\le N-j}\bar{x}_l\bar{x}_{l+j}$$

$0 \le j \le M-1,$ for N_l pairs with no missing data. (4)

Eigenvalues and eigenvectors are computed for the lagged autocorrelation matrix as with SSA. The eigenvectors contain no missing data. Computation of the kth principal component ignores missing data points

$$\bar{a}_i^k = \frac{M}{N_l}\sum_{l\le M}\bar{x}_{i+l}\bar{E}_l^k$$

$0 \le i \le N-M,$ for $N_l\, x_{i+1}$ with no missing data. (5)

If $N_l < fM$, where $0 \le f \le 1$ is a specified fraction of allowable missing data points within window size M, then $\bar{a}_i^k$ is assigned a missing value. Reconstructed components are calculated as with SSA. If any principal component value in the sum is missing, then the RC value will be missing.

Table 1. Percent Variance of Synthetic Time Series in Subtidal, Semidiurnal, and Noise Components

	Actual	SSA	SSAM-first half missing	SSAM-missing values from Bay example	SSAM-one-half randomly missing
Subtidal	50.3	51.3	51.6	51.8	51.6
Semidiurnal	21.0	21.0	20.8	20.9	21.2
Noise	28.7	27.7	27.6	27.3	27.2

Table 2. Median Percent Error of Reconstructed Synthetic Time Series Relative to $c_s(t)$

SSA	SSAM-first half missing	SSAM-missing values from Bay example	SSAM one-half randomly missing
2.1	2.5	5.5	4.2

Application of SSAM to a Synthetic Time Series

To test the ability of SSAM to properly process missing data, a synthetic SSC time series was analyzed with SSA and variants containing missing data were analyzed with SSAM. The synthetic SSC time series has the following characteristics intended to be similar to SSC time series in San Francisco Bay: one water year in length, starts at the beginning of a water year (0000 hours October 1), 15-minute time step, periodic components oscillating about a mean value of 100 mg/L, and a random noise component. A seasonal component simulates the effect of seasonal winds on sediment resuspension in the Bay, which is least during autumn and greatest during spring [*Schoellhamer*, 1996]. The fortnightly spring/neap cycle affects the magnitude of tidal currents, sediment resuspension, and advection, and varies seasonally with the greatest amplitude in winter and summer. Advection of suspended sediment, which typically is between relatively deep channels and more turbid shallow water, produces a semidiurnal component of SSC that is modified by the spring/neap and seasonal cycles. Thus, the 'true' SSC signal was generated with the mean and periodic terms

$$\begin{aligned} c_s(t) = 100 - 25\cos\omega_s t + 25(1-\cos 2\omega_s t)\sin\omega_{sn} t \\ +25(1+0.25(1-\cos 2\omega_s t)\sin\omega_{sn} t)\sin\omega_a t \end{aligned} \qquad (6)$$

in which the seasonal angular frequency $\omega_s = 2\pi/365\ day^{-1}$, the spring/neap angular frequency $\omega_{sn} = 2\pi/14\ day^{-1}$, and the advection angular frequency $\omega_a = 2\pi/\ (12.5/24)\ day^{-1}$. Noise tends to increase with SSC so the synthetic time series

$$c(t) = 0.2Rc_s(t) + c_s(t) \qquad (7)$$

in which R is a time series of normally distributed random numbers with zero mean and standard deviation of one. The three periodic components, synthetic signal $c_s(t)$, and synthetic time series $c(t)$ are shown in Figure 1.

SSA was applied with a 30-hour (M=120) window intended to separate the semidiurnal component from the seasonal and fortnightly components. This window size was used by *Schoellhamer* [1996] to analyze SSC data in San Francisco Bay. The first four modes from SSA contained the periodic components and 72.3% of the total variance, 51.6% in two modes containing the subtidal (seasonal and fortnightly) components and 21.0% in two modes containing the semidiurnal component (Table 1). The remaining modes contained noise. The reconstructed time series from the first four modes is virtually identical to that shown in Figure 1d and has a median error of 2.1% (Table 2). SSAM produced the same result with this complete time series.

SSAM was successfully applied to several versions of the time series with differing distributions of missing data. The fraction of allowable missing data points for computation of a principal component (f) was set equal to 0.5. Analysis

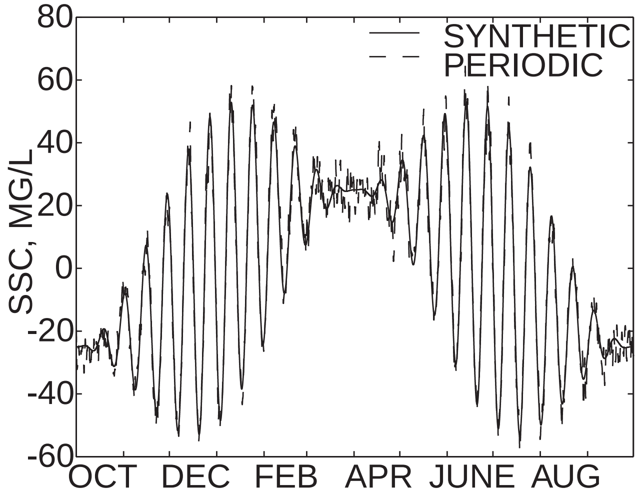

Figure 2. Comparison of subtidal components from the synthetic time series and from SSAM analysis of the synthetic SSC time series with missing values distributed the same as for the real SSC time series in Figure 3.

of the time series with a single missing value imbedded in the middle of the time series produced virtually the same result as with the complete time series. A 6-month gap was placed at the beginning of the synthetic time series $c(t)$ and the error of the time series reconstructed by SSAM was identical to SSA (Table 2). Similar results were obtained when a 6-month data gap was placed in the center and at the end of the synthetic time series. The partitioning of variance between the components for these tests was nearly identical to that for the actual time series and SSA (Table 1).

Another test assigned 61% of the data missing values which were distributed the same as for the real SSC time series presented in the next section. The median error of the reconstructed time series (5.5%, Table 2) was greater than that found for SSA or the previous examples, but the standard deviation of $c(t)$ was 39.5% of the mean.

This example demonstrates that SSAM can be used as a low-pass filter for time series with missing data. SSAM subtidal components account for 51.8% of the variance and primarily contain signals with periods greater than the 30-hour window size, in this case the seasonal and spring/neap components. The sum of the actual seasonal and spring/neap components contains 50.3% of the variance and is nearly identical to the SSAM subtidal components when data are available (Fig. 2).

SSAM also was successfully applied to the synthetic time series with one-half of the values randomly assigned a missing value. The fraction of allowable missing data points for computation of a principal component (f) was set equal to 1. The resulting modes and variances were virtually identical to the actual values and those found by SSA for the complete time series (Table 1) and the median error of the reconstructed time series was 4.2% (Table 2).

Application of SSAM to SSC in San Francisco Bay

SSAM was applied to an SSC data set from San Francisco Bay for which the general outcome has been independently determined and the fraction of missing data is large. *Buchanan and Schoellhamer* [1999] collected SSC time series at mid-depth at the San Mateo Bridge in water year 1997 (Fig. 3). This time series was chosen because it was from a site at which SSA had been previously applied to continuous blocks of SSC data [*Schoellhamer*, 1996], the water year 1997 data were not used in the previous analysis, and the amount of valid data (39%) was the smallest of all of the Bay sites during water year 1997. The fraction of allowable missing data points for computation of a principal component (f) was set equal to 0.5 and a 30-hour window (M=120) was used.

The individual modes resulting from SSAM are in good agreement with previous SSA analysis of other SSC data from San Francisco Bay. The first 10 modes of the SSAM analysis contain 89.1% of the variance. Mode 1 contains 53.7% of the variance and is associated with subtidal frequencies due to the spring/neap and seasonal cycles. Previous SSA found that about one-half of the variance of SSC was subtidal and primarily due to the spring neap/cycle. Modes 2 and 3 account for 14.2% of the variance and represent semidiurnal tides, modes 4 and 5 account for 10.6% of the variance and represent diurnal tides, modes 6 and 7 account for 5.8% of the variance and fluctuate at a quarter diurnal period, mode 8 accounts for 2.0% of the variance and contains a mix of tidal frequencies, and modes 9 and 10 account for 2.8% of the variance and represent terdiurnal (8-hour) tidal fluctuations. Previous application of SSA also found several tidal modes that contained periods of about n/24, where n is an integer. Diurnal and semidiurnal fluctuations were greatest (n=1 and 2) and terdiurnal and quarter diurnal fluctuations (n=3 and 4) were present, but smaller. Thus, application of SSAM to time series with missing data produces similar results to application of SSA to complete time series.

The SSC time series reconstructed by the first 10 RCs is in good agreement with the original time series (Fig. 4). The reconstructed time series contains some negative concentrations when the measured concentration was nearly zero.

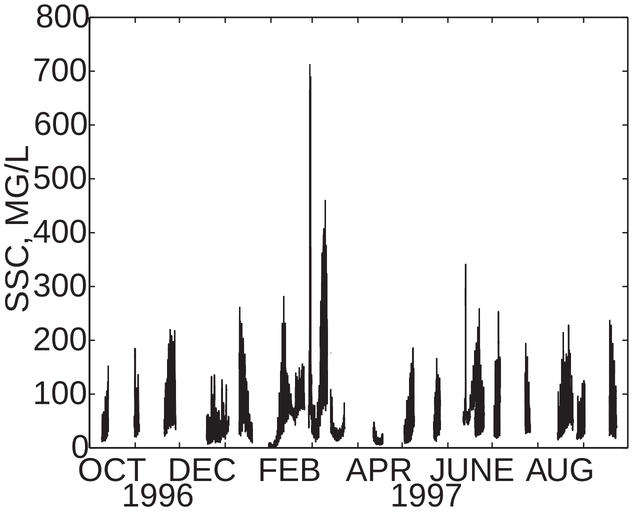

Figure 3. Time series of mid-depth suspended-sediment concentration at San Mateo Bridge during water year 1997.

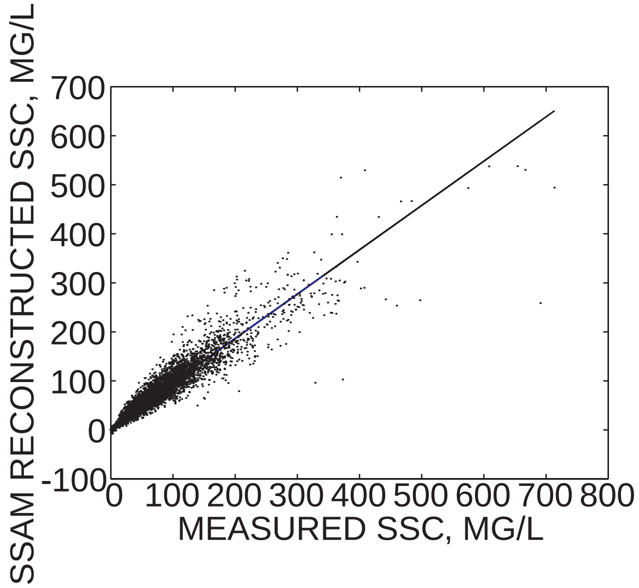

Figure 4. Comparison of measured and SSAM reconstructed mid-depth suspended-sediment concentration at San Mateo Bridge during water year 1997. These 11536 data points have a squared correlation coefficient of 0.905 and the regression is significant at <0.001 level.

Conclusions

Singular spectrum analysis has been modified to allow missing data and SSAM can successfully analyze and reconstruct synthetic and actual incomplete time series. SSAM also can be used to low pass filter incomplete time series. An example of source code for SSAM is available at http://ca.water.usgs.gov/ja/grl/.

Acknowledgments. The support of the U.S. Army Corps of Engineers, San Francisco District, as part of the San Francisco Estuary Regional Monitoring Program for Trace Substances, and the U.S. Geological Survey Place-Based Program is gratefully acknowledged. Helpful comments have been provided by Andy Robertson, Brent Troutman, and two anonymous reviewers. Lisa Lucas and Jan Thompson provided invaluable help with this manuscript.

References

Allen, M. R., and L. A. Smith, Investigating the origins and significance of low-frequency modes of climate variability, *Geophysical Research Letters*, *21,* 883-886, 1994.

Buchanan, P.A., and C.A Ruhl, Summary of suspended-solids concentration data, San Francisco Bay, California, water year 1998, *Open File Report 99-189,* 41 pp., U.S. Geological Survey, 2000.

Buchanan, P.A., and D. H. Schoellhamer, Summary of suspended-solids concentration data, San Francisco Bay, California, water year 1997, *Open File Report 00-88 URL http:// ca.water.usgs.gov/rep/ofr99189/,* 52 pp., U.S. Geological Survey, 1999.

Dettinger, M.D., M. Ghil, C. M. Strong, W. Weibel, and P. Yiou, Software expedites singular-spectrum analysis of noisy time series, *Eos Trans. AGU, 76*(2), 12, 1995.

Prieto, R, L. Gimeno, R. Garcia, R. Herrera, E. Hernandez, and P. Ribera, Interannual variability of hail-days in the Andes region since 1885, *Earth and Planetary Science Letters 171*(3), 503-509, 1999.

Robertson, A.W. and C. R. Mechoso, Interannual and decadal cycles in river flows of southeastern South America, *Journal of Climate, 11*(10), 2570-2581, 1998.

Schoellhamer, D.H., Factors affecting suspended-solids concentrations in South San Francisco Bay, California, *J. Geophys. Res., 101*(C5), 12087-12095, 1996.

Sharma, A.S., D. Vassiliadis, and K. Papadopoulos, Reconstruction of low-dimensional magnetospheric dynamics by singular spectrum analysis, *Geophysical Research Letters, 20*(5), 335, 1993.

Varadi, F., J.M. Pap, and C.J. Henney, Searching for signal in noise by random-lag singular spectrum analysis, *The Astrophysical Journal, 526*(2), 1052-1061, 1999.

Vautard, R., P. Yiou, and M. Ghil, Singular-spectrum analysis: A toolkit for short, noisy, chaotic signals, *Physica D, 58*, 95-126, 1992.

Vautard, R. and M. Ghil, Singular spectrum analysis in nonlinear dynamics with applications to paleoclimatic time series, *Physica D, 35*, 395-424, 1989.

Wang, X.L., J. Corte-Real, and X. Zhang, Intraseasonal oscillations and associated spatial-temporal structures of precipitation over China, *J. Geophys. Res., 101*(D14), 19035-19042, 1996.

Yiou, P., K. Fuhrer, L.D. Meeker, J. Jouzel, S. Johnsen, and P.A. Masked, Paleoclimatic variability inferred from the spectral analysis of Greenland and Antarctic ice-core data, *J. Geophys. Res., 102*(C12), 26441-26454, 1997.

D.H. Schoellhamer, U.S. Geological Survey, 6000 J Street, Placer Hall, Sacramento, California 95819.
(e-mail: dschoell@usgs.gov)

(Received November 30, 2000; revised March 15, 2001; accepted May 17, 2001.)