Soft Comput
DOI 10.1007/s00500-011-0767-3

FOCUS

Singular spectral analysis of ill-known signals and its application to predictive maintenance of windmills with SCADA records

Luciano Sánchez · Inés Couso

Abstract A generalization of the singular spectral analysis (SSA) technique to ill-defined data is introduced in this paper. The proposed algorithm achieves tight estimates of the energy of irregular or aperiodic oscillations from records of interval or fuzzy-valued signals. Fuzzy signals are given a possibilistic interpretation as families of nested confidence intervals. In this context, some types of Supervisory Control And Data Analysis (SCADA) records, where the minimum, mean and maximum values of the signal between two scans are logged, are regarded as fuzzy constrains of the values of the sampled signal. The generalized SSA of these records produces a set of interval-valued or fuzzy coefficients, that bound the spectral transform of the SCADA data. Furthermore, these bounds are compared to the expected energy of AR(1) red noise, and the irrelevant components are discarded. This comparison is accomplished using statistical tests for low quality data, that are in turn consistent with the possibilistic interpretation of a fuzzy signal mentioned before. Generalized SSA has been applied to solve a real world problem, with SCADA data taken from 40 turbines in a Spanish wind farm. It was found that certain oscillations in the pressure at the hydraulic circuit of the tip brakes are correlated to long term damages in the windmill gear, showing that this new technique is useful as a failure indicator in the predictive maintenance of windmills.

L. Sánchez (✉)
Computer Science Department, University of Oviedo,
Campus de Viesques, 33071 Gijón, Asturias, Spain
e-mail: luciano@uniovi.es

I. Couso
Facultad de Ciencias, Statistics Department,
University of Oviedo, 33071 Oviedo, Asturias, Spain
e-mail: couso@uniovi.es

1 Introduction

The causes of the mechanical failures in windmill gears are not fully understood. Many different types of breakage risks have been studied, but the reasons why similar windmills, placed in similar locations, have different behaviors in the long term are not clear yet (Ramesh and Jithesh 2008). Notwithstanding this, it is assumed that frequent stoppage and starting of the windmill may result in broken teeth in gear wheels and pinions, and thus an anomalous frequency content of the tip brake pressure can be related to future breakdowns. Therefore, the frequency analysis of the tip brake pressure might be used for detecting abnormal stresses in the gear and for anticipating costly mechanical failures.

Braking produces a sudden descent in the hydraulic pressure of the corresponding circuit, that quickly recovers its steady value. In most cases, this pressure is routinely monitored and logged by Supervisory Control And Data Acquisition (SCADA) systems. However, braking pulses are too fast for the operating scan intervals in most SCADA systems, that average the signals over a period which is normally much larger than the braking time. Since narrow pulses are filtered out in this process, there is no enough information in the records of the brake pressure for characterizing the dynamics of the braking subsystem. This hinders the use of this type of information for doing predictive maintenance, unless the sampling rate is increased or a secondary sensory system is deployed that complements SCADA data (Martnez-Rego et al. 2011).

An exclusive use of the SCADA data allows for cost-conscious diagnosis. However, increasing the sampling rate is not normally feasible. The scan interval in SCADA systems is determined as a compromise between the amount of data that must be transferred and the sample rate

Published online: 08 October 2011

needed for capturing the dynamic behavior of the signals. In particular, windmill SCADA systems monitor in real time the temperature, pressure, speed, voltage and current at different points of the generators (Zaher et al. 2009). There are some hundreds of signals for each windmill, justifying a scan interval in the range of minutes. This is an adequate period for temperatures, but such a rate is too slow for monitoring electrical magnitudes or hydraulic pressures in the brakes, for instance.

In practice, there is an intermediate solution that allows for transmitting a useful part of the frequency contents of the signal without lowering the sampling period. Some SCADA systems do not only record the average value of the signal between two scans, but also its minimum and maximum values, as shown in the example in Fig. 1. This figure displays a case where there is a fast oscillation between two records. The oscillation has a negligible influence in the average signal, but it is detected when the minimum and the maximum are also transmitted. The frequency contents of this compound signal is still incomplete; we cannot know, for instance, how many cycles the oscillation lasted, neither the frequency of this oscillation; this triple logging allows perceiving low frequencies in detail, while the highest frequencies are imprecisely known. Nevertheless, it can be assumed that this ill-defined signal still carries enough information for certain kinds of diagnosis. Consequently, there is an interest in extending the concept of "spectral transform" to ill-defined signals, and in particular to those of the type minimum–average–maximum. It is also relevant to learn whether this extended spectral transform, when applied to minimum–average–maximum SCADA records of the tip brake pressure, can be successfully used for detecting abnormal oscillations, and correlated with mechanical failures in the gears of windmills.

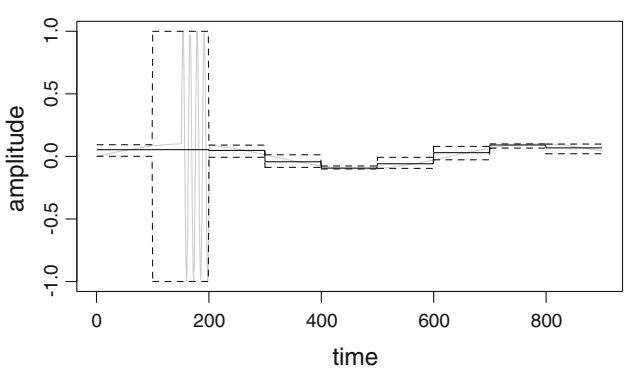

Fig. 1 SCADA records of a fast signal (drawn in *grey*). Values are logged with a sample period of 100 s. The logged signal (*solid black*) comprises the average values of the original series. The fast oscillation about $t = 150$ is lost in this discretization, but some information about it can be kept if the minimum and the maximum (*dotted black*) are logged too

The spectral technique chosen for this purpose is the singular spectrum analysis (SSA) (Ghil 2002). SSA is suitable for this problem because all the physical magnitudes in a windmill are correlated with the wind at the farm. Wind-originated oscillations are likely to be irregular and aperiodic, and SSA, which is mostly used in meteorology-related applications (Allen 1992), is better suited to this task than Fourier analysis, Gabor transform or wavelets, to name some (Allen et al. 1996). SSA does not rely on a previously defined basis of functions, as done for instance in the mentioned Fourier or wavelet analysis, but produces its own basis from the available data. As a consequence of this, SSA can be used for characterizing nearly chaotic signals, as well as it is useful for separating signal and noise when the energy of the noise decays with the frequency, where Fourier analysis is prone to detecting non-existing oscillations at low frequencies (Allen and Robertson 1996). Both uses will be discussed in the next section.

Because of this and the other reasons mentioned before, in this paper SSA will be extended to ill-defined data. In this respect, certain possibilistic interpretations of fuzzy values as nested families of confidence intervals (Couso and Sánchez 2008a) are adequate for this problem, and therefore fuzzy techniques for spectral analysis are suitable too. However, while there are previous papers combining fuzzy logic and SSA (Fang et al. 1993), the use of these algorithms with vague data is scarce, and mostly centered in Fourier (Butkiewicz 2006) and wavelet (Ho et al. 2001) transforms. Up to our knowledge, the problem of extending SSA to fuzzy data has not been addressed before.

The structure of this paper is as follows: in Sect. 2 SSA for crisp data is introduced and its suitableness for finding irregular oscillations and separating signal and noise in the problem at hand are discussed. In Sect. 3, an extension of SSA to fuzzy data is described, and illustrated with the help of synthetical data. In Sect. 4, the case study which motivated this paper, that of finding oscillations in the hydraulic pressure of the tip brake, and correlating them with long-term mechanical failures in the generator, is addressed. The paper is finished with some concluding remarks in Sect. 5.

2 Singular spectrum analysis for characterizing in frequency a mix of periodic signals and colored noise

In the following, we will consider that the time series

$$\{X(t) : t = 1, \ldots, N\} \tag{1}$$

is a sequence of samples of the variable of interest, whose mean value is zero. Let us embed this series in a vector space of dimension M, using lagged copies of the scalar data,

$$\mathbf{X}(t) = (X(t), X(t+1), \ldots, X(t+M-1)) \tag{2}$$

thus

$$\{\mathbf{X}(t) : t = 1, \ldots, N-M+1\}. \tag{3}$$

SSA consists in calculating the principal directions of the sequence of augmented vectors $\mathbf{X}(t)$ in phase space. In the first place, following (Ghil 2002), the covariance matrix C is estimated as

$$C_{ij} = \frac{1}{N-|i-j|} \sum_{t=1}^{N-|i-j|} X(t)X(t+|i-j|) \tag{4}$$

and the eigenelements $\{(\lambda_k, \rho_k) : k = 1, \ldots, M\}$ of C are obtained by solving

$$C\rho_k = \lambda_k \rho_k. \tag{5}$$

The eigenvalue λ_k equals the partial variance in the direction ρ_k. This is equivalent to form the matrix E whose columns are the eigenvectors and the diagonal matrix Λ whose elements are the eigenvalues λ_k in descending order:

$$E'CE = \Lambda. \tag{6}$$

The eigenvectors ρ_k of the lag covariance matrix C are also called empirical orthogonal functions (EOFs). It has been shown that pairs of EOFs in quadrature, when associated to eigenvalues with the same or similar modulus, are the nonlinear counterpart of a sine-cosine part in the standard Fourier analysis of linear problems (Ghil 2002).

The principal components (PCs) A_k are obtained by projecting the time series onto each EOF,

$$A_k(t) = \sum_{j=1}^{M} X(t-j+1)\rho_k(j). \tag{7}$$

In Fig. 2 a realistic function is displayed, that will be used for illustrating the algorithms in this paper. A 50 Hz sinusoidal signal was sampled at 100 KHz. Each 100 consecutive samples are averaged, giving an effective sample rate of 1 KHz. Between samples 100 and 200, a small signal in the 2nd harmonic is injected, and also between samples 400 and 500, where both 2nd and 3rd harmonics are present. Beginning in the 500th sample, pulses of 10 µs appear at 180 Hz. All these signals are corrupted by a strong AR(1) red noise, whose energy decays with frequency; this particular type of noise was chosen because it is more common in practice than the white noise assumed in many engineering models, besides it is much troublesome separating it from the signal. In Fig. 3, the amplitudes of the eigenvalues λ_k of the covariance matrix of this series are shown. The EOFs associated to the first six eigenvalues are displayed in Fig. 4, and the first six PCs are shown in Fig. 5. The embedding size is $M = 40$.

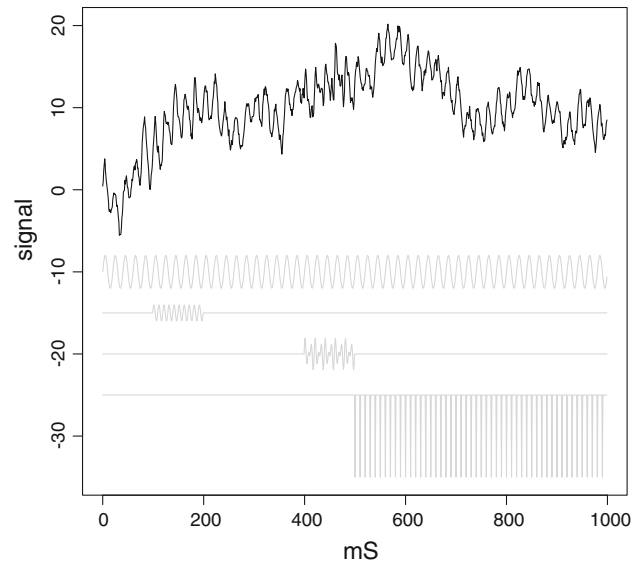

Fig. 2 Realistic synthetic data. The sample period is of 100 KHz, averaged in blocks of 100 samples. The time series is a combination of 50 Hz signal and AR(1) red noise, with a chirp of the 2nd harmonic between samples 100 and 200, another one of 2nd and 3rd harmonics between samples 400 and 500, and a sequence of pulses of 10 µs at 180 Hz between samples 500 and 1,000. These pulses do not apperar in the averaged signal (solid black) which mimics the appearance of the SCADA records of the hydraulic pressure in the tip brake circuit records in a windmill

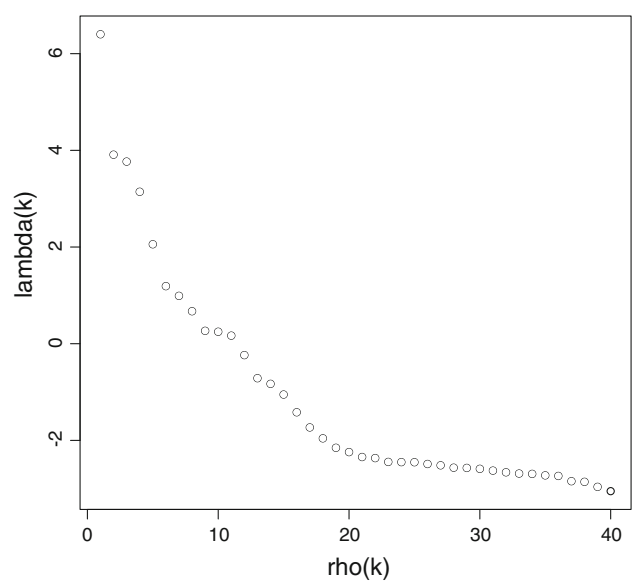

Fig. 3 SSA spectrum: logarithms of the ordered eigenvalues of the covariance matrix of the series in Fig. 2

2.1 SSA for characterizing noisy signals in frequency

It is clear from Fig. 5 that the relevance of the different components in the spectral transform does not only depend on the modulus of the corresponding eigenvalue. Observe

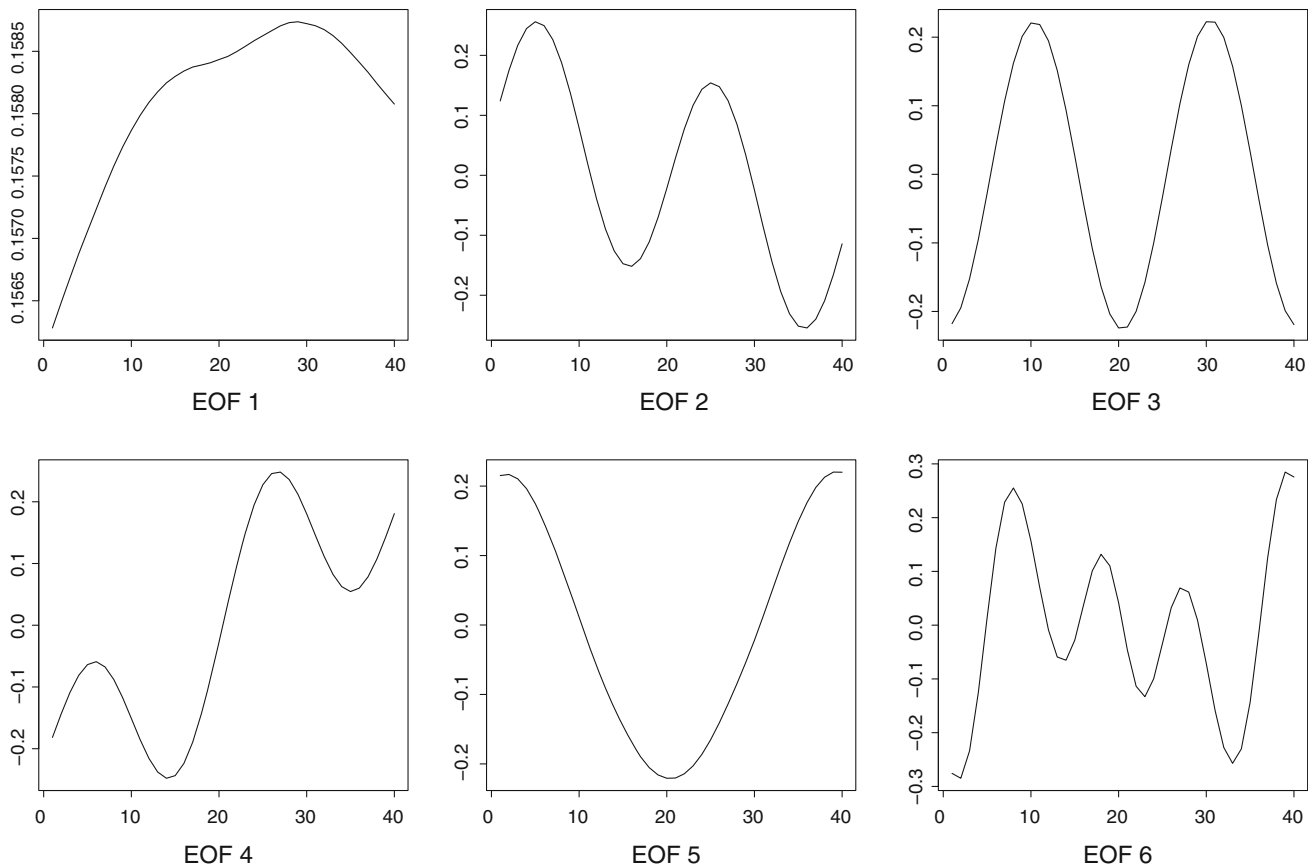

Fig. 4 SSA spectrum: empirical orthogonal functions associated to the six largest eigenvalues of the example problem described in Sect. 2

that the first PC in Fig. 5 closely matches the drift of the signal, that can be attributed to the frequency contents of the colored noise and thus the coefficient associated to this base function is not relevant for the diagnosis. Components 4 and 5 do not seem to include information relevant for the frequency characterization of the signal either. On the contrary, components 2 and 3 contain most of the 50 Hz signal, while component 6 shows traces of the higher harmonics, which we want to detect.

This separation between relevant and not relevant components cannot be achieved on the basis of their energies. The eigenvalues displayed in Fig. 3 reveal that the first five EOFs contain most of the energy of the signal. However, the information given by this selection would not be much different than that arising when filtering out those frequencies that do not match a peak in the power spectral density (PSD) of the series, as shown in Fig. 6, where the PSD was computed as the Fourier transform of the autocorrelation of the signal

$$\mathrm{acf}(\tau) = \sum_{t=\tau}^{N} X(t)X(t-\tau). \tag{8}$$

In order to compare the SSA spectrum with the PSD of the signal, it is useful to match each EOF with a dominant frequency, obtained by means of a reduced Fourier transform (Vautard et al. 1992), i.e. the maximum w.r.t. f of

$$e(\rho_k, f) = \left(\left(\sum_{j=1}^{M} \rho_{kj} \cos(2\pi f j) \right)^2 + \left(\sum_{j=1}^{M} \rho_{kj} \sin(2\pi f j) \right)^2 \right)^{\frac{1}{2}}. \tag{9}$$

As we have mentioned, a pair of almost identical eigenvalues associated to the same frequency signals a periodic oscillation at that frequency (Vautard and Ghil 1989); in this example, this has been signaled by the circles in Fig. 7. Observe also the similarities between this graph and the PSD shown in Fig. 6.

Albeit similar, the information provided by the PSD is not exactly the same as that provided by the SSA eigenvalues. On the one hand, the largest pairs of eigenvalues, marked in Fig. 7, match the 1st harmonic (ρ_2 and ρ_3) and the higher order chirps (ρ_6 and ρ_7, ρ_9 and ρ_{10}), but these last two groups are much more noticeable in the SSA transform than they were in the PSD (Fig. 6). On the other hand, the component with the highest energy (ρ_1) appears isolated, thus we know it is a trend and not an oscillation, something that we cannot tell from Fig. 6.

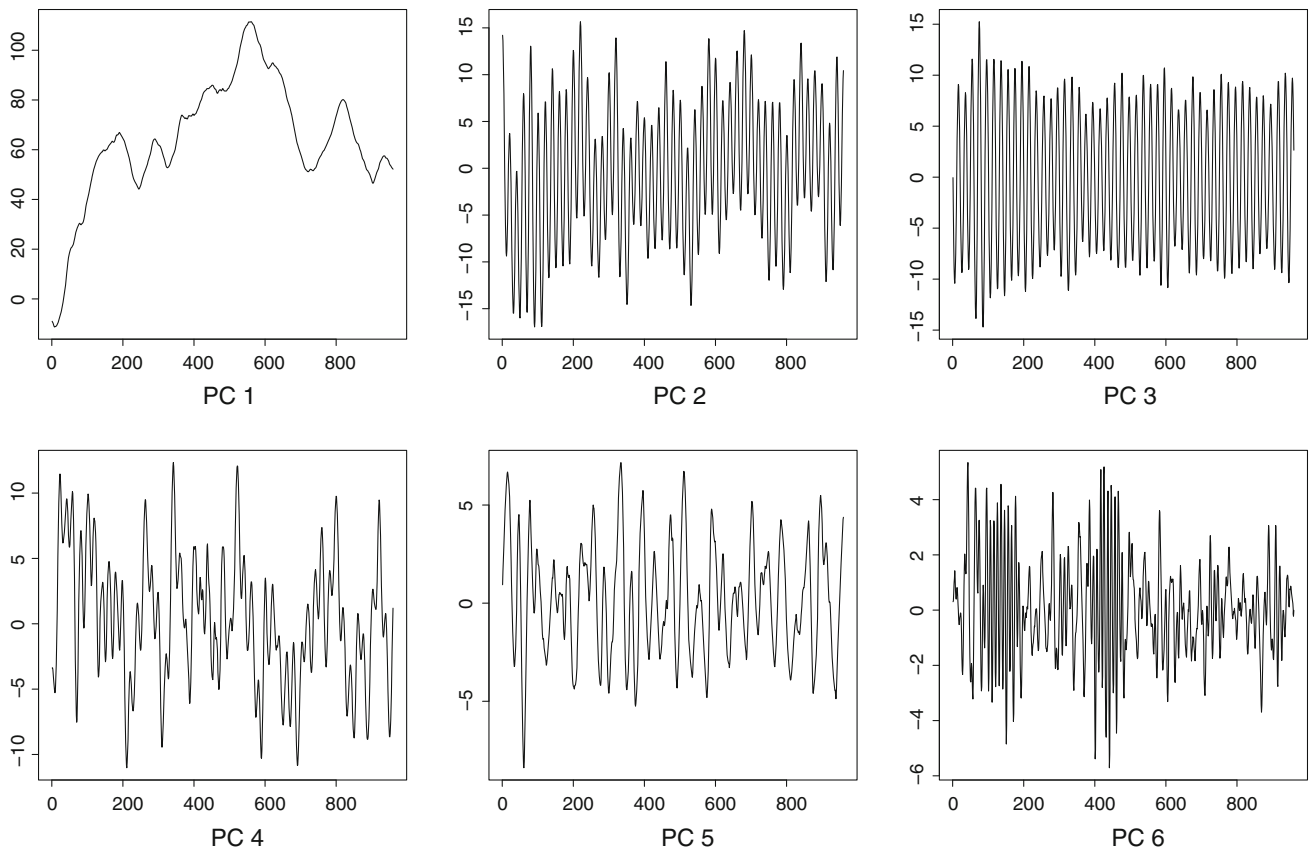

Fig. 5 SSA spectrum: principal components associated to the six largest eigenvalues of the example problem described in Sect. 2

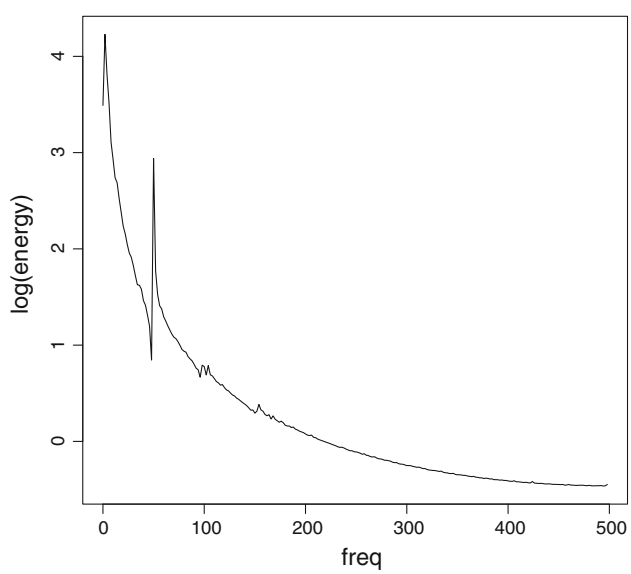

Fig. 6 Power spectral density of the example function, showing spikes at 1 and 50 Hz, and also small peaks at 100 and 150 Hz

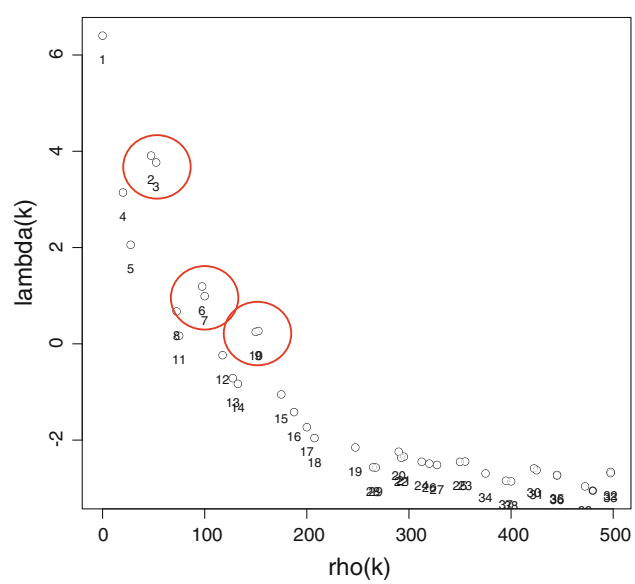

Fig. 7 SSA spectrum: eigenvalues versus dominant frequencies

If a procedure for separating a group $\mathscr{E}$ of relevant EOFs was available, then a filtered series could be reconstructed by combining the PCs associated to the EOFs in $\mathscr{E}$, as follows (Vautard et al. 1992):

$$R_{\mathscr{E}}(t) = \frac{1}{M_t} \sum_{k \in \mathscr{E}} \sum_{j=L_t}^{U_t} A_k(t-j+1)\rho_k(j). \tag{10}$$

The values of M_t, L_t and U_t are

$$(M_t, L_t, U_t) = \begin{cases} (t, 1, t) & 1 \le t \le M-1, \\ (M, 1, M), & M \le t \le N-M+1, \\ (N-t+1, & \\ t-N+M, M), & N-M+1 \le t \le N. \end{cases} \tag{11}$$

In Fig. 8, the results of this reconstruction, applied to the subset of EOFs marked with circles in Fig. 7, is displayed. Observe that the original signal was recovered with a fidelity that cannot be obtained with standard filtering techniques, and this proofs that the information in an small subset of six EOFs is enough for summarizing the properties of the signal.

In the next subsection we will describe a procedure for selecting a subset of EOFs that are statistically relevant for the problem, according to certain hypotheses about the properties of the noise.

2.1.1 *Determining the statistically relevant components of the SSA transform*

For determining whether a given eigenvalue in the SSA is relevant in a reconstruction of the signal, we will use statistical tests. These tests will determine whether the energy of an isolated component is compatible or not with pure AR red noise.

Let us assume a first order model, thus our null hypothesis is that the signal is pure AR(1) red noise,

$$X(t) = a_1[X(t-1) - X_0] + \sigma\xi(t) + X_0, \tag{12}$$

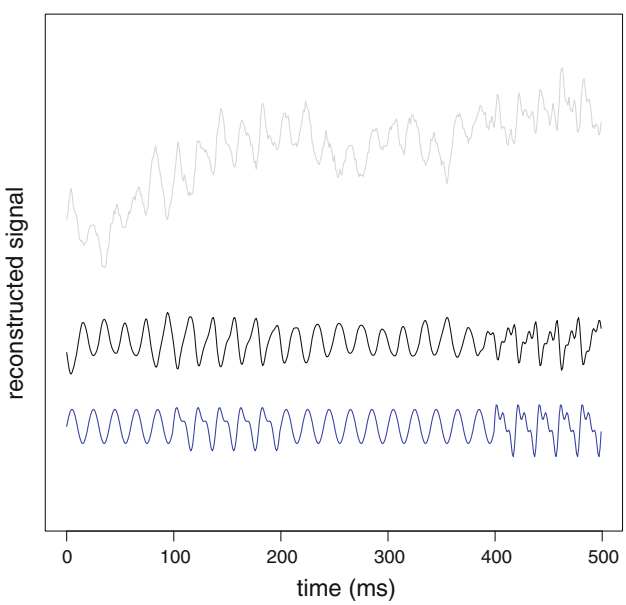

Fig. 8 *Upper part* synthetic SCADA data with noise. *Center* SSA-based filter. *Lower part* synthetic noise-free data, showing the quality of the reconstruction

where $\xi(t)$ is Gaussian-distributed white noise with zero mean and unit variance.

The first step consists in estimating the parameters $\hat{X}_0, \hat{\sigma}$ and $\hat{a}_1$ from the time series $X(t)$, using a maximum likelihood criterion (Akaike 1969). In addition to this, $N \cdot S$ independent realizations of a Gaussian distributed random variable, also with zero mean and unit variance, are drawn:

$$\{\xi^s(t), s = 1, \ldots, S, t = 1, \ldots, N\}. \tag{13}$$

On the basis of these calculations, a list comprising S *surrogate* noise series is generated as follows:

$$\hat{X}^s(t) = \begin{cases} \hat{a}_1[\hat{X}^s(t-1) - \hat{X}_0] + \hat{\sigma}\xi^s(t) + \hat{X}_0 & \text{if } t > 2 \\ \hat{a}_1[\hat{X}(1) - \hat{X}_0] + \hat{\sigma}\xi^s(1) + \hat{X}_0 & \text{if } t = 1 \end{cases} \tag{14}$$

A covariance matrix C^s is evaluated then for each of the X^s

$$\hat{C}^s_{ij} = \frac{1}{N - |i-j|} \sum_{t=1}^{N-|i-j|} \hat{X}_s(t)\hat{X}_s(t + |i-j|). \tag{15}$$

These matrices are projected onto the same basis E of the original data (Eq. 6), defining in turn S approximately diagonal matrices $\hat{\Lambda}^s$,

$$\hat{\Lambda}^s = E'\hat{C}^s E, \ s = 1, \ldots, S. \tag{16}$$

Let us group the diagonal elements of these matrices $\hat{\Lambda}^s$ into M sets

$$L_k = \{\hat{\lambda}^s_k\}_{s=1,\ldots,S}. \tag{17}$$

Each of these sets contains S independent realizations of an estimator of λ_k (assumming that the null hypothesis is true). At this point, a confidence interval for each of the M eigenvalues $\lambda_k, k = 1, \ldots, M$, can be produced by bootstrap estimation. It is remarked that, in this paper, a simple percentile interval will be used, as bias corrected and accelerated estimators such as BC_α (Efron and Tibshirani 1993) do not alter the selection of EOFs noticeably, however the numerical procedure would not be substantially altered if a different estimator is chosen. Therefore, the $1 - 2\alpha$ percentile interval is defined by the interval $L_{k\alpha}$ that spans the values between the α and $1 - \alpha$ percentiles of the set L_k defined in Eq. 17.

Lastly, the associated statistical test, for a level $(1 - 2\alpha)^M$, consists in rejecting that the series is pure noise when some of the λ_k are not in the mentioned confidence intervals L_α. If the null hypothesis is rejected, the set $\mathcal{E}$ of EOFs used for separating signal and noise comprise all the columns ρ_k of the matrix E whose eigenvalues λ_k are not between the α and $1 - \alpha$ percentiles of the bootstrap distribution of $\{\hat{\lambda}^s_k, \ s = 1, \ldots, S\}$.

In Fig. 9, the confidence intervals arising from the example problem ($(1-2\alpha)^{40} = 0.95$) have been plotted in grey, along with the eigenvalues λ_k of the series. Observe that the significant EOFs are the pairs $(\rho_2, \rho_3), (\rho_6, \rho_7)$ and (ρ_9, ρ_{10}), as expected, along with the isolated components ρ_8, ρ_{11}, which can be discarded, since they are not associated to pairs of conjugated EOFs.

Observe also that the spikes displayed in Fig. 2 are undetected in the spectral transform, since we have not yet included the information about the maximum and the minimum and these short pulses do not have enough energy for appearing in the averaged signal. The generalization of SSA to fuzzy signals, able to exploit the imprecise frequency contents in min–avg–max signals, will be introduced in the next section.

3 SSA for fuzzy-valued series

Consider that the elements of the time series are not accurately perceived but all the information about the value of $X(t)$ is given by a nested family of confidence intervals,

$$\{[X^l_\alpha(t), X^u_\alpha(t)]\}_{\alpha \in A} \tag{18}$$

where

$$\alpha_1 \le \alpha_2 \Rightarrow [X^l_{\alpha_1}(t), X^u_{\alpha_1}(t)] \supseteq [X^l_{\alpha_2}(t), X^u_{\alpha_2}(t)] \quad \text{for all } \alpha_1, \alpha_2 \in A, \tag{19}$$

and

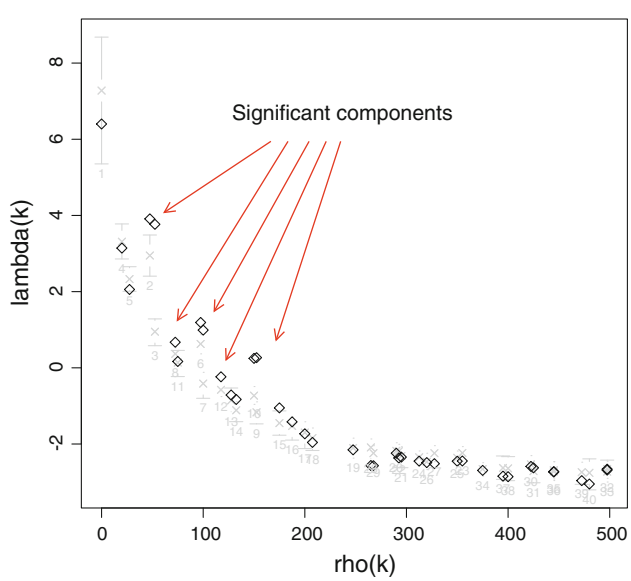

Fig. 9 SSA and Monte-Carlo generated confidence intervals for AR(1) red noise

$$P(X(t) \in [X^l_\alpha(t), X^u_\alpha(t)]) \ge 1 - \alpha \quad \text{for all } t = 1, \ldots, N. \tag{20}$$

Let the possibilistic representation of this information (Couso and Sánchez 2008a) be the fuzzy-valued time series

$$\{\widetilde{X}(t) : t = 1, \ldots, N\}, \tag{21}$$

where

$$\widetilde{X}(t)(x) = \sup\{\alpha \in A \mid x \in [X^l_\alpha(t), X^u_\alpha(t)]\}. \tag{22}$$

Lastly, let the *degree of compatibility* between a crisp series $\{S(t) : t = 1, \ldots, N\}$ and the fuzzy series $\{\widetilde{X}(t)\}$ be defined as

$$\mu_X(S) = \min_t \widetilde{X}(t)(S(t)). \tag{23}$$

$C(S)$ is the lag covariance matrix of the series S:

$$C(S)_{ij} = \frac{1}{N - |i-j|} \sum_{t=1}^{N-|i-j|} S(t)S(t + |i-j|). \tag{24}$$

With the help of the preceding definitions, it is proposed that the extension to fuzzy data of the SSA technique is a numerical procedure that inputs a fuzzy times series $\widetilde{X}(t)$ and outputs two results:

1. An orthonormal basis
$$(\rho_1, \ldots, \rho_M). \tag{25}$$
2. A vector of fuzzy values
$$(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_M). \tag{26}$$

These vectors fulfill the following properties:

– The set comprising all the products between the covariance matrices of the series that are compatible with $\widetilde{X}$, and the elements of the mentioned orthonormal basis, is contained in the set defined by the products of the fuzzy eigenvalues and the same elements of the basis,

$$\int_{C(S)\rho_j} (\mu_X(S) \mid C(S)\rho_j) \subset \widetilde{\lambda}_j \odot \rho_j, \quad j = 1, \ldots, M. \tag{27}$$

where

$$\widetilde{A} \odot (v_1, \ldots, v_m) = (\widetilde{A} \odot v_1, \ldots, \widetilde{A} \odot v_M) \tag{28}$$

and

$$(\widetilde{A} \odot k)(x) = \begin{cases} \widetilde{A}(x/k) & k \ne 0 \\ 0 & \text{else.} \end{cases} \tag{29}$$

– The fuzzy sets $\widetilde{\lambda}_j$ are the most specific sets fulfilling Eq. 27. The nonspecificity at level α is measured by the volume

$$\text{nonspec}(\alpha) = \prod_{j=1}^{M} (\sup[\widetilde{\lambda}_j]_\alpha - \inf[\widetilde{\lambda}_j]_\alpha), \tag{30}$$

thus the sets $\widetilde{\lambda}_j$ have to chosen so that nonspec (α) is minimized for all $\alpha \in A$.

The elements ρ_k of the orthogonal basis are therefore approximations to the EOFs of the unknown lag covariance matrix $C(X)$, and the sets $\widetilde{\lambda}_j$ are fuzzy restrictions of the eigenvalues of the same matrix. Consequently, the principal components (PCs) are fuzzy time series $\widetilde{A}_k$, obtained by projecting the time series onto each EOF,

$$\widetilde{A}_k(t) = \bigoplus_{j=1}^{M} \widetilde{X}(t-j+1) \odot \rho_k(j) \tag{31}$$

where

$$(\widetilde{A} \oplus \widetilde{B})(x) = \sup\{\min(\widetilde{A}(a), \widetilde{B}(b)) : a+b = x\}. \tag{32}$$

Observe also that the fuzzy extension of the SSA here proposed reduces itself to the crisp version if the elements of $\widetilde{X}(t)$ are singletons.

In the next subsections, we will describe two different numerical procedures for estimating these approximations to the eigenvalues and eigenvectors of ill-defined time series. The first procedure is based on the definition of a fuzzy-valued lag covariance matrix, and the second one on preprocessing the series with a Karhunen–Loewe transform (Rao and Yip 2001).

3.1 Fuzzy lag covariance matrix

A simple approach to solve this problem consists in extending Vautard and Ghil's definition (1989) by means of fuzzy arithmetic operators, and define a fuzzy-valued lag covariance matrix,

$$\widetilde{C}_{ij} = \frac{1}{N-|i-j|} \bigoplus_{t=1}^{N-|i-j|} \widetilde{X}(t) \odot \widetilde{X}(t+|i-j|) \tag{33}$$

where the fuzzy addition has been defined before, and

$$(\widetilde{A} \odot \widetilde{B})(x) = \sup\{\min(\widetilde{A}(a), \widetilde{B}(b)) : ab = x\}. \tag{34}$$

For diagonalizing this fuzzy matrix, it will be assumed that the EOFs of the unknown time series are comparable to those of the crisp time series whose degree of compatibility with $\widetilde{X}$ is the highest. In other words, let

$$S^*(t) = \arg\max_x \widetilde{X}(t)(x) \tag{35}$$

be the crisp series formed by the modal points of the fuzzy time series $\widetilde{X}$, and let us admit in the first place that

$$C(X)\rho = \lambda\rho \Rightarrow C(S^*)\rho \approx \eta\rho \tag{36}$$

for suitable real numbers $\lambda(\rho)$ and $\eta(\rho)$. Let E_S^* be the matrix whose columns are the eigenvectors of S^*. For computing the fuzzy bounds of the eigenvalues, the proposed approximation is:

$$(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_M) \approx \text{diag}(E'_{S^*} \widetilde{C} E_{S^*}), \tag{37}$$

where the product between matrices with fuzzy terms is understood as a fuzzy arithmetic-based extension of the matrix product.

Observe that this technique can be regarded as a generalization of the method for obtaining spectrum of the surrogate series, seen in the preceding section. It is remarked that the success of this procedure depends on the elements $\widetilde{X}$ of the fuzzy time series being specific enough for Eq. 36 being admissible. Otherwise the calculations do not produce specific enough results for many practical purposes. Following with the example time series described in the preceding section, this problem will be shown later in this paper, in the left part of Fig. 12, where the computations defined in this section have been applied to a fuzzy series where the support of their terms is defined by the minimum and maximum of each group of samples, and whose modal points are the average values of the data (see a detail of this series in Fig. 10).

3.2 Karhunen–Loewe transform, maximal specificity of the fuzzy eigenvalues

The second procedure is computationally harder than the preceding one, but narrower bands for the eigenvalues can be estimated than those obtained by the fuzzy covariance matrix

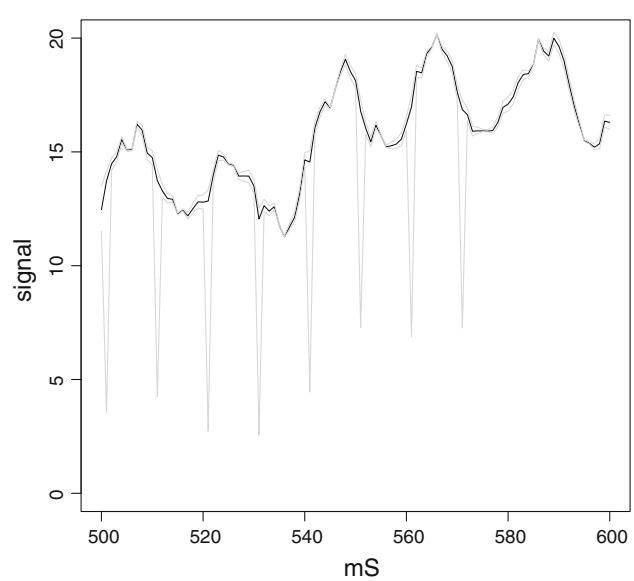

Fig. 10 Time Series. Sample period = 100 KHz, minimum, maximum and average of the blocks shown in Fig. 2. Detail of samples between 50,000 and 60,000. The pulses can only be perceived in the minimum part

based method. The algorithm that will be described in this section is based on a transform of the augmented series, so that its covariance matrix is approximately diagonal.

Let $\mathbf{S}$ be the augmentation of the selection S,

$$\mathbf{S}(t) = (S(t), S(t+1), \ldots, S(t+M-1)), \tag{38}$$

and let $E(S)$ be an orthogonal matrix and $\Lambda(S)$ a diagonal matrix, such that

$$C(S) = E'(S) \times \Lambda(S) \times E(S) \tag{39}$$

for all S, and let the Karhunen–Loewe (KH) transform of $\mathbf{S}$ be (Rao and Yip 2001)

$$\mathbf{Z}(S,t) = E(S) \times \Lambda(S)^{-\frac{1}{2}} \times \mathbf{S}(t), \tag{40}$$

thus the sample covariance of $\mathbf{Z}$ is the identity matrix. Lastly, let the augmented fuzzy time series $\widetilde{\mathbf{X}}(t)$ be defined

$$\widetilde{\mathbf{X}}(t) = (\widetilde{X}(t), \widetilde{X}(t+1), \ldots, \widetilde{X}(t+M-1)) \tag{41}$$

with

$$\{\widetilde{\mathbf{X}}(t) : t = 1, \ldots, N-M+1\} \tag{42}$$

and the following transform:

$$\widetilde{\mathbf{Z}}(S,t) = E(S) \times \Lambda(S)^{-\frac{1}{2}} \odot \widetilde{\mathbf{X}}(t), \tag{43}$$

where the product between a crisp matrix A and a fuzzy vector $\widetilde{V} = (\widetilde{V}_1, \ldots, \widetilde{V}_M)$ is a fuzzy subset of $\mathbf{R}^M$ whose membership function is

$$\begin{aligned}[A \odot \widetilde{V}](V) = \max\{\min\{\widetilde{V}_1(v_1), \ldots, \widetilde{V}_M(v_M)\} : \\ (v_1, \ldots, v_M) \in \mathbf{R}^M,\ V = A \cdot (v_1, \ldots, v_M)'\}.\end{aligned} \tag{44}$$

Observe that the matrix operation $E(S) \times \Lambda(S)^{-\frac{1}{2}}$ is a rotation followed by a scaling. Each α-cut of the product $E(S) \times \Lambda(S)^{-\frac{1}{2}} \odot \widetilde{\mathbf{X}}(t)$ can be efficiently computed by applying these rotation and scaling operators to the vertices of the same α-cut of $\widetilde{\mathbf{X}}(t)$, the result being defined as the convex hull of the transformed vertexes.

The lag covariance matrix of $\mathbf{Z}(S, t)$ will be the identity matrix, and the fuzzy extension (seen in the preceding section) of Vautard and Ghil's definition of the covariance of $\widetilde{\mathbf{Z}}(S,t)$, will be nearly diagonal. This allows for a more efficient numerical approximation of the covariance of the fuzzy time series, where the fuzzy multiplication is avoided in favor of the square operator. Let $\widetilde{C}_{jj}(\widetilde{\mathbf{Z}}(S)), j = 1 \ldots, M$ be the jth term of the diagonal of the covariance of $\widetilde{\mathbf{Z}}(S,t)$; assuming that this matrix is diagonal, the definition in Eq. 33 can be approximated by

$$\widetilde{C}_{jj}(\widetilde{\mathbf{Z}}(S)) \approx \frac{1}{N-M+1} \bigoplus_{t=1}^{N-M+1} \left(\mathrm{Proj_j}(\widetilde{\mathbf{Z}}(\mathrm{S,t}))\right)^2 \tag{45}$$

where the projection and square operators are defined as follows:

$$(\mathrm{Proj_j}\widetilde{\mathbf{Z}})(\mathrm{u}) = \max\{\widetilde{\mathbf{Z}}(\mathrm{x}) \mid \mathrm{x} \in \mathbf{R}^{\mathrm{M}}, \mathrm{u} = \mathrm{x_j}\} \tag{46}$$

$$(\widetilde{A}^2)(x) = \max\{\widetilde{A}(a) \mid x = a^2\} \tag{47}$$

thus

$$(\widetilde{\lambda}_1(S), \ldots, \widetilde{\lambda}_M(S)) = \Lambda(S) \odot (\widetilde{C}_1(\widetilde{\mathbf{Z}}(S)), \ldots, \widetilde{C}_M(\widetilde{\mathbf{Z}}(S)))'. \tag{48}$$

The requisite that the nonspecificity of the fuzzy eigenvalues is minimized is achieved by finding the crisp time series S, $\mu_X(S) > 0$, such that the value defined in Eq. 30 is minimized for each value of α. The optimization algorithm used for finding this series will be detailed later.

3.2.1 Graphical example

In Fig. 11, there is a graphical explanation of the steps taken for the computation of the nonspecificity of a fuzzy time series, given the eigenvectors matrix. For an easier interpretation of the figures, an embedding $M = 2$ has chosen, thus the α-cuts of the elements of the augmented series $\widetilde{\mathbf{X}}(t)$ are rectangles. For making a simpler representation, only one of these cuts is displayed.

The upper left part of the figure shows the terms of the augmented series (rectangle shaped cuts, are mentioned). In the first place, the covariance matrix of the centerpoints of these terms is diagonalized. The directions of its eigenvectors are represented by the black perpendicular axis in this upper left part. It is remarked that the algorithm proposed in this paper will ultimately produce a different set of eigenvectors for this problem; the directions of these last eigenvectors are shown by the red axes.

If the eigenvectors of the ill-defined series were those obtained by diagonalization of the covariance of the centerpoints, then the result of applying the KH transform to the imprecise data would be the graph shown in the upper center part, where each rectangle is transformed into a rhomboid (see Eq. 44). For obtaining the α-cut of the approximately diagonal fuzzy covariance of this data (Eq. 45), the interval-valued variances of the projections of these rhomboids are computed. This amounts to enclosing each rhomboid in a rectangle (upper part, right) and computing the pair or variances of the sets of the least favorable corners (the most distant points to the origin, as shown in the lower left part of the same figure) and the nearest points to the origin (not marked in the figure). These interval-valued variances are computed as follows:

$$\overline{\sigma}_j^2 = \frac{1}{N} \bigoplus_{t=1}^{N} \{x^2 : x \in \mathrm{proj_j}\widetilde{\mathbf{Z}}_\alpha\} \tag{49}$$

(see also Eq. 45) and the covariance matrix of the transformed augmented series is approximated by the diagonal matrix $C(\widetilde{\mathbf{Z}})$, where

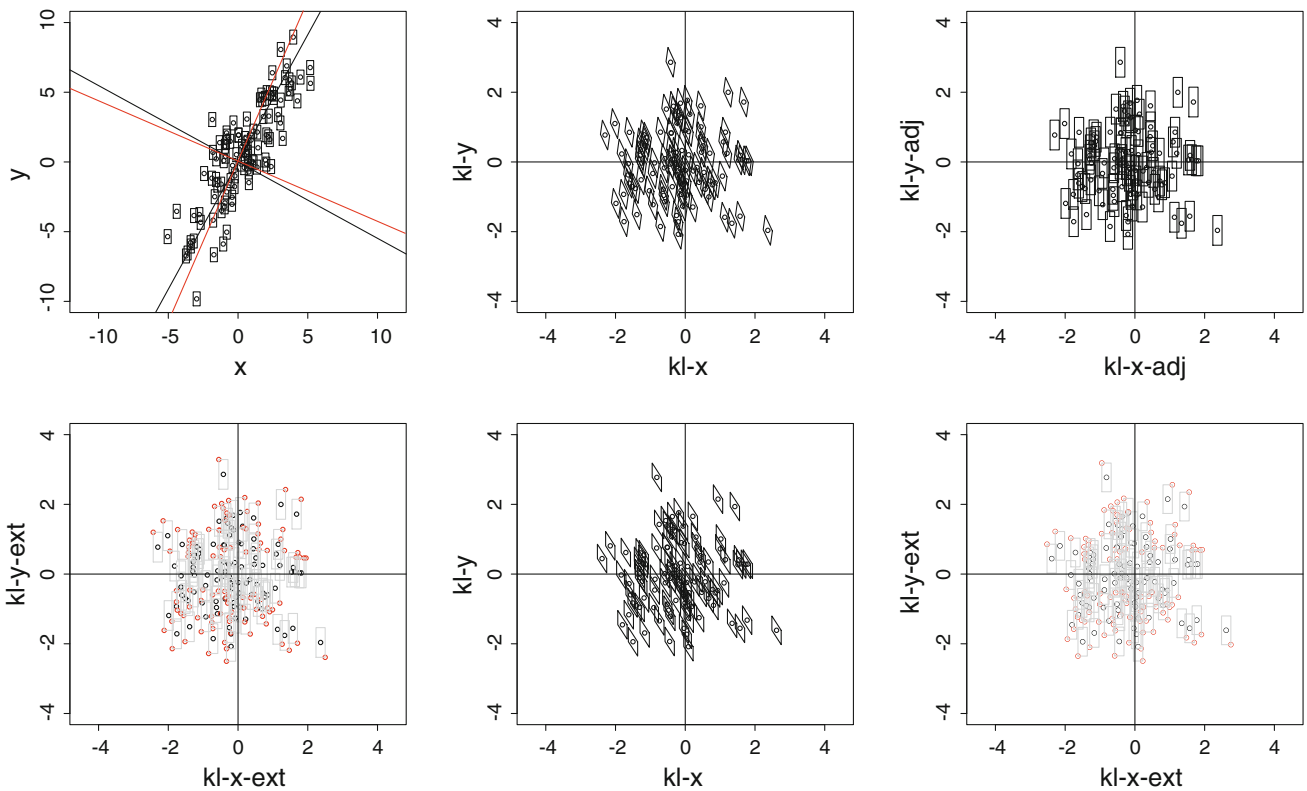

Fig. 11 Steps in the determination of the eigenvectors producing the lowest nonspecificity

$$[C_j(\widetilde{\mathbf{Z}})]_\alpha = \overline{\sigma}_j^2. \tag{50}$$

In the lower, center part of the same figure the KH transform with respect to the optimal selection (those eigenvectors for which the specificity of the fuzzy eigenvalues is maximum) is plotted, which in this case amounts to a clock-counterwise rotation of the eigenvectors of the centerpoints (upper left part, axis colored in red, as mentioned before) so that the romboids in the KH transform of the data are aligned with the vertical axis, therefore the projection on the vertical axis is shorter, and the increase in the horizontal projection is balanced by the vertical reduction, making for an higher specificity, implicit in the lower, right part of the figure.

3.2.2 *Numerical search of the eigenvectors maximizing the specificity*

The method described in this section inputs an orthonormal matrix E and a fuzzy time series $\widetilde{X}$, and outputs both a set of fuzzy eigenvalues $\widetilde{\lambda}_j$ and the nonspecificity nonspec(α) of each α-cut. The proposed extension of SSA to fuzzy data requires determining a matrix E^* for which nonspec(α) is minimum for all α, as mentioned.

E^* is not an arbitrary orthogonal matrix, but it must be the eigenvector matrix of a crisp series S^* fulfilling $\mu_X(S^*) > 0$. Therefore, in this paper the optimization is intended to find a crisp series, contained in the support of $\widetilde{X}$, such that the vector of nonspecificities for all the considered values of α is not worse than that of a different series. Multicriteria Genetic algorithms are well suited for this task, and therefore the well-known NSGA-II algorithm (Deb et al. 2002) was chosen. The decisions taken are summarized in the list that follows:

1. Coding: Each individual represents a time series $S(t)$ with N' terms, where $N' \leq N$. Real coding is used, where the allelles $\beta_t, t = 1, \ldots, N'$, are real numbers between 0 and 1, and
$$S(t) = \min \operatorname{supp}(\widetilde{X}(t)) + \beta_t(\max \operatorname{supp}(\widetilde{X}(t)) - \min \operatorname{supp}(\widetilde{X}(t))) \tag{51}$$
2. Genetic operators: Standard arithmetic crossover and mutation (Michaelewicz 1994).
3. Fitness function: A vector comprising the list of the values "nonspec(α)", $\alpha \in A$ obtained when the eigenvector matrix that diagonalizes the lag covariance matrix of the series S which is represented by the individual is used. The dominance relation between fitness vectors is as follows:

$$\{\eta_\alpha\}_{\alpha\in A} \preceq \{\gamma_\alpha\}_{\alpha\in A} \Longleftrightarrow \eta_\alpha \le \gamma_\alpha \quad \text{for all } \alpha \in A, \text{ and } \eta_{\alpha^*} < \gamma_{\alpha^*} \text{ for some } \alpha^* \in A \tag{52}$$

3.3 Determining the statistically relevant components of the fuzzy SSA transform

The bootstrap tests defined in Sect. 2.1.1 are now extended to imprecise data. The null hypothesis is the same as before: the signal is pure AR(1) red noise (see Eq. 12) and their parameters $\hat{X}_0, \hat{\sigma}$ and $\hat{a}_1$ are estimated from the modal points of the fuzzy time series $\widetilde{X}(t)$. Again, S surrogate series are generated and a confidence interval for each of the M eigenvalues $\lambda_k, k = 1, \ldots, M$, is produced by bootstrap estimation.

The statistical test associated to these confidence intervals has to take into account that the values λ_k are not precisely known, but they are perceived through the fuzzy estimations $\widetilde{\lambda}_k$ described in the preceding section. Therefore, an statistical test for low quality data is needed (Couso and Sánchez 2011a, b).

Recall that $L_{k\beta}$ is the interval spanning the values between the β and $1-\beta$ percentiles of the set L_k defined in Eq. 17. For a given α-cut and a level $(1-2\beta)^M$, it is not rejected that the series is pure noise if the α-cuts of all the $\widetilde{\lambda}_k$ are completely contained in the mentioned confidence intervals, i.e.

$$[\widetilde{\lambda}_k]_\alpha \subset L_{k\beta} \quad \text{for all } k. \tag{53}$$

Otherwise, the null hypothesis is rejected if

$$[\widetilde{\lambda}_k]_\alpha \cap L_{k\beta} = \emptyset \quad \text{for some } k, \tag{54}$$

and the test is not conclusive if neither Eqs. 53 or 53 are met. This can also be expressed by means of a fuzzy p-value (Couso and Sánchez 2008b; Couso and Sánchez 2011b; Denœux et al. 2005); consider that the null hypothesis is rejected if the p-value is lower than a given bound, and this p-value is a fuzzy set whose α-cut is an interval $[\beta_*, \beta^*]$ where

$$\beta_* = \sup\{\beta \mid [\widetilde{\lambda}_k]_\alpha \subset L_{k\beta} \quad \text{for all } k\}, \tag{55}$$

$$\beta^* = \sup\{\beta \mid [\widetilde{\lambda}_k]_\alpha \cap L_{k\beta} \neq \emptyset \quad \text{for all } k\}. \tag{56}$$

If the null hypothesis is rejected, or the test is inconclusive, the most conservative set $\mathcal{E}$ of EOFs used for separating signal and noise comprise all the columns ρ_k of the matrix E whose eigenvalues $\widetilde{\lambda}_k$ fulfill

$$\text{supp}(\widetilde{\lambda}_k) \not\subset L_{k\beta}. \tag{57}$$

In turn, for evaluating the criterion "similar amplitudes at the same frequency", used for detecting oscillations, the set of possible distances between two fuzzy eigenvalues $\widetilde{\lambda}_k$ and $\widetilde{\lambda}_r$ is the fuzzy set

$$d(\widetilde{\lambda}_k, \widetilde{\lambda}_r)(t) = \max\{\min(\widetilde{\lambda}_k(\lambda), \widetilde{\lambda}_r(\mu)) : d(\lambda, \mu) = t\} \tag{58}$$

and in the particular case where the supports of these sets are considered, then (a) we discard that two amplitudes are similar if the minimum of the preceding set is higher than a given bound, (b) we assume that they are the same if the maximum is smaller than the bound and (c) we cannot decide otherwise.

Summarizing, the outline of the whole procedure is as follows:

1. Estimate a set of values for the noise parameters which is compatible with the modal points of the fuzzy data.
2. Generate N surrogate time series according to these different noise parameters.
3. Initialize M set-valued counters.
4. Determine the confidence intervals $L_{k\beta}$, with $(1-2\beta)^M = 0.95$.
5. Determine whether the spectrum of the experimental data is contained, intersects with, or has disjoint sets of values for each frequency under study. Add 0, $\{0, \frac{1}{N}\}$ or $\frac{1}{N}$ to the kth counter, respectively.
6. Go to step 4 and repeat N times.
7. If the product of the maximum values of the counters is lower than 0.05, reject the hypothesis (i.e. assume that the time series is not noise). If the product of the minimum values of all the counters is higher than 0.05, do not reject that the signal is AR noise. In other cases, the test is not decisive.
8. Find the EOFs whose counters are different than zero and reconstruct the signal with this information.

3.4 Realistic example

The SSA spectrum plotted in Fig. 9 for a crisp time series is easily generalized to a fuzzy time series. In the proposed approach, the EOFs are crisp, but the eigenvalues are fuzzy, thus each point in these figures can be replaced by a vertical bar spanning the range of the corresponding component.

According to this, the SSA spectrum of the ill-defined signal plotted in Fig. 2 is shown in Fig. 12. In the left part of this figure, the fuzzy lag-covariance matrix-based approximation was used. In the right part, the Karhunen–Loewe approximation, with genetic reduction of the nonspecificity, was applied instead. Observe that the uncertainty of the eigenvalues is noticeably reduced in this last approach, with respect to the first approximation. There is still information enough about the oscillations (as marked by the circles), and the pulse of width 20 μs at 180 Hz appears for the first time (marked by the arrow).

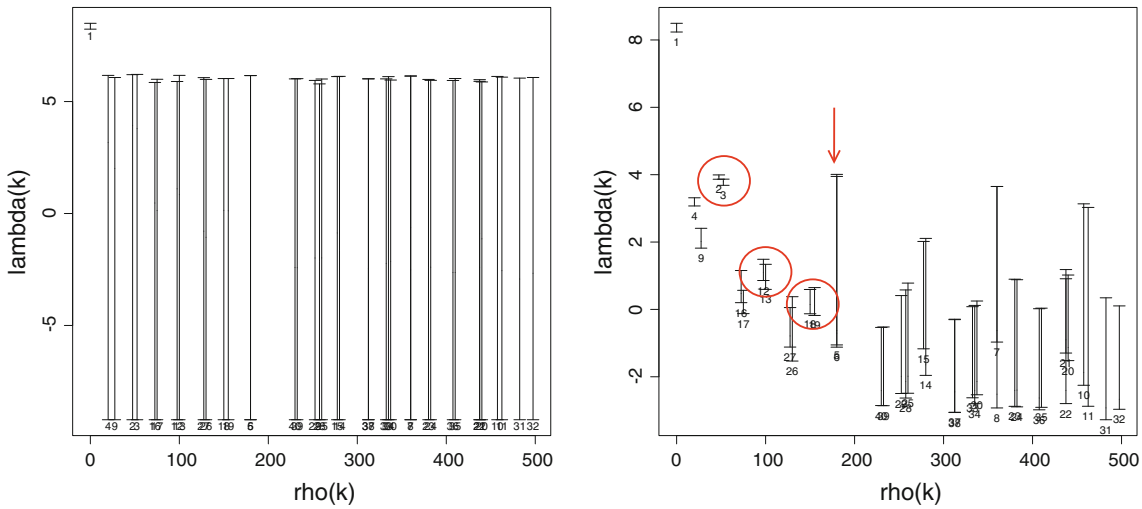

Fig. 12 Fuzzy-SSA based in the use of fuzzy covariance matrices (*left part*) and the iterative reduction of the nonspecificity (*right part*), applied to the same fuzzy time series

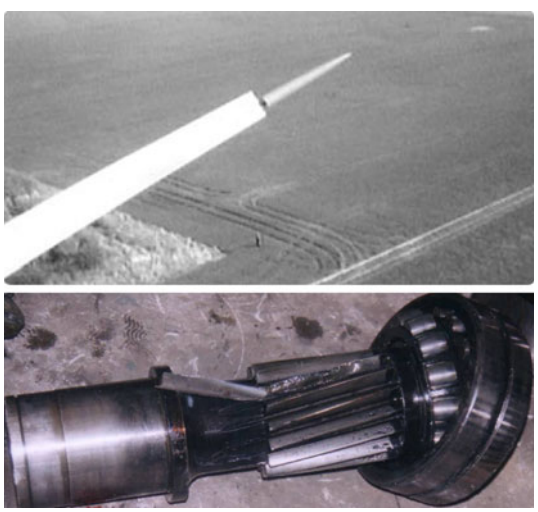

Fig. 13 *Upper part* Tip brake. *Lower part* Teeth of the Intermediate pinion broken as a result of frequent stoppage and starting of the windmill [images taken from (Stiesdal 1999) (Ramesh and Jithesh 2008)]

Observe also that the imprecision grows with the frequency, as commented in the introduction, and there are also aliases of the 180 Hz component. In short, fuzzy SSA has been able to detect traces of signals located over the Nyquist frequency of the equivalent sample rate, as the mentioned pulse of width 20 μs.

In the following section, we show with the help of a practical application how this procedure can be used for characterizing the frequency content of data given by triplets (minimum, average, maximum), as happens with the practical problem that inspired this transform.

4 Case study

Assessing the breakdown risks of windmills from SCADA data is a cost effective measure, nonetheless limited by the low dynamic quality of the available data. In the case at hand, we are interested in discovering evidence that helps to predict mechanical failures (bearings and gear box damage). We analyzed data from 40 fixed pitch windmills, collected between the years 2006 and 2009.

The interval rate of the SCADA data of the studied wind farm is of 10 min, and minimum, average and maximum of different electrical variables, rotor speed, temperatures and hydraulic pressures are logged. In this study, we are interested in the hydraulic pressure of the tip brake circuit (see Fig. 13). This signal is relevant to our diagnostic because frequent stoppage and starting of the windmill may result in broken teeth in gear wheels and pinions (Ramesh and Jithesh 2008), and thus an anomalous frequency content of the tip brake pressure might be related to future breakdowns. Having said that, braking produces a sudden descent in the hydraulic pressure, that quickly recovers its steady value. Narrow pulses are filtered out when 10 min of data are averaged, and therefore there is not information enough in the average pressure for characterizing the tip brake dynamics. It is needed to complement the data with the "minimum" signal, which detects whether there has been at least one brake action in the last scan interval. This fact justify the use of SSA for imprecise data in this context.

In Fig. 14 we have displayed the bounds of the eigenvalues found with interval SSA, as proposed in this paper, applied to the SCADA data of the brake pressure, with respect to the prevalent frequency of the EOFs. These

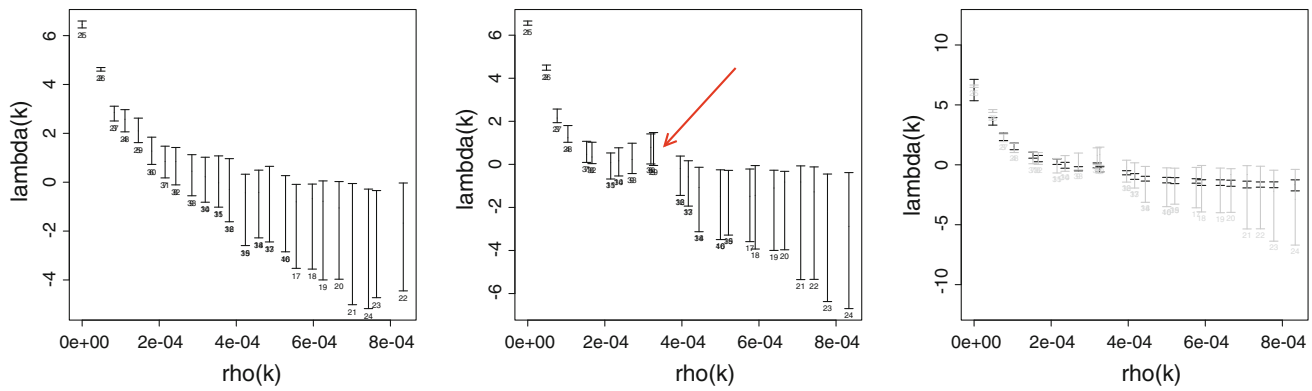

Fig. 14 SSA of tip brake pressure data in 2006 for windmills that were without mechanical problems between 2006 and 2009 (*left*) and for one whose gear box had to be replaced in 2009 (*center*). *Right* surrogate AR(1) data used for judging the relevance of the differences of energy between the signal and red noise

Table 1 Confusion matrix of a classifier based in the comparison between the tip brake pressure and red noise

	Mechanical failures	Normal
Predicted "risk of failure"	67%	9%
Predicted "normal"	34%	91%
Number of windmills	6	34

signals were captured in 2006, shortly after the windmills were first put into service. The left part corresponds to a windmill that has not suffered relevant mechanical problems to date. The center signal was taken in a windmill whose gear box had to be replaced in 2009; observe that there is a noticeable increase of the energy at mid-frequencies.

In order to determine whether there is a significant correlation between the SSA spectrum and the breakdowns of the generators, we have used the statistical test mentioned in the preceding section for deciding whether the energy at the band of interest is higher or not than that of AR(1) red noise. We have generated 100 surrogate series for each generator, with parameters estimated from the average signals. Periods comprising 144 samples of continuous coupling of the generator and the electrical network (1 day) were evaluated, and the size of the augmented data is $M = 24$ (4 h). The graph corresponding to the the second windmill is displayed in the right part of Fig. 14. The results of the test have been used to design a simple classifier: if the energy between 0.00025 and 0.0004 Hz is significantly higher than that of AR(1) noise (with parameters estimated from the average data), then the windmill is marked as "risk of failure", otherwise the windmill is marked as "normal". The confusion matrix of the resulting classifier is in Table 1; observe that the whole dataset has been used to compute the confusion matrix, as no parameters were learned.

The results suggest that some of the generators in the farm there might have undergone damage in 2006 when first put into service, perhaps because of wind turbulences or an incorrect functioning of the tip brakes; the gear boxes of the affected generators failed three years later. However, the set of data is too small and new studies are needed for further supporting this hypothesis.

5 Concluding remarks

When the interval rate of SCADA data is large, the frequency contents of the signal is limited and quick changes cannot be detected. Adding the minimum and the maximum values of the variable during each interval is an alternative to increasing the sample rate, but conventional spectral techniques cannot exploit this triple signal.

In this paper an extension of SSA to fuzzy-valued time series was proposed, that allows to recover an imprecise spectrum from compound signals, as well as others that can be given a possibilistic interpretation as a nested family of confidence intervals. This spectrum can be compared with those of different kinds of noise, and the relevant components of the signal isolated by means of statistical tests for imprecise data. These significant components can in turn be used for improving the signal to noise ratio, or for describing the series, as done in the practical application that motivated this paper, where different sequences of pulses in the tip brake pressure of a windmill have been characterized and then classified, allowing the diagnosis of the gearbox by examining the frequency spectrum of a sequence of braking pulses, in a case where the length of these pulses was shorter than the resolution of the recorded data.

Acknowledgments This work was funded by Spanish M. of Education, under the grant TIN2008-06681-C06-04, and INDRA S.A., contract FUO-EM-024-11.

References

Akaike H (1969) Fitting autoregressive models for prediction. Ann Inst Stat Math 21:243–247

Allen M (1992) Interactions between the atmosphere and oceans on time scales of weeks to years. PhD thesis. University of Oxford, Oxford, England

Allen MR, Robertson AW (1996) Distinguishing modulated oscillations from coloured noise in multivariate datasets. Clim Dyn 12:775–784

Allen MR, Smith LA, Monte Carlo SSA (1996) Detecting irregular oscillations in the presence of coloured noise. J Clim 9:3373–3404

Butkiewicz BS (2006) Fuzzy approach to Fourier transform. In: Proceedings of SPIE, 6159-II, Article number 615947

Couso I, Sánchez L (2008a) Higher order models for fuzzy random variables. Fuzzy Sets Syst 159:237–258

Couso I, Sánchez L (2008b) Defuzzification of fuzzy *p*-values. In: Proceedings of soft methods in probability and statisticas SMPS, pp 126–132

Couso I, Sánchez L (2011a) Mark-recapture techniques in statistical tests for imprecise data. Int J Approx Reason 52(2):240–260

Couso I, Sánchez L (2011b) Inner and outer fuzzy approximations of confidence intervals. Fuzzy Sets Syst 184:68–83

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comp 6(2):182–197

Denœux T, Masson MH, Hébert PA (2005) Nonparametric rank-based statistics and significance test for fuzzy data. Fuzzy Sets Syst 153:1–28

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Monographs on statistics and applied probability, vol 57. Chapman and Hall, London

Fang LQ, Mi D, Xu XS, Chen DG (1993) Fault diagnosis of fuel injection system of engine based on the singular spectrum analysis. J Combust Sci Technol 9(2):108–111

Ghil et al (2002) Advanced spectral methods for climatic time series. Rev Geophys 40(1):1–41

Ho DWC, Zhang Pm, Xu J (2001) Fuzzy wavelet networks for function learning. IEEE Trans Fuzzy Syst 9(1):200–211

Michaelewicz Z (1994) Genetic algorithms + data structures = evolution programs. 2nd edn edn. Springer, Berlin

Martínez-Rego D, Fontela-Romero O, Alonso A (2011) Power Wind Mill Fault Detection via one-class *v*-SVM Vibration Signal Analysis. Proc. IJCNN 2011. Accepted

Ramesh J, Jithesh SV (2008) Breakdown risk in wind energy turbines. Pravartark III(3):9–10

Rao KR, Yip PC (2001) The transform and data compression handbook. CRC Press, Boca Raton

Stiesdal H (1999) The wind turbine: components and operation. Bonus-Info Special Issue, Bonus Energy A/S, Autumn

Vautard R, Ghil M (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Phys D 35:395–424

Vautard R, Yiou P, Ghil M (1992) Singular spectrum analysis: a toolkit for short noisy chaotic signals. Phys D 58:95–126

Zaher A, McArthur SDJ, Infield DG (2009) Online wind turbine fault detection through automated SCADA data analysis. Wind Energy 12(6):574–593