Filling in long gap sequences by performing EOF and FDA jointly

Plaia Antonella, Di Salvo Francesca, Ruggieri Mariantonietta, Agró Gianna

Abstract In this paper the EOF methodology is performed jointly with the FDA approach on a spatiotemporal multivariate data set with the aim to fill in missing values as accurately as possible when long gap sequences occur. Simulated data sets, containing "artificial" gaps, are considered in order to test the performance of two proposed procedures; in the first one, observed data are reconstructed by EOF and then converted into functional ones; in the second one, observed data are transformed into functional ones and then EOF reconstruction is applied. By comparing some performance indicators computed for the two procedures, it is shown that a pre-processing of data by FDA, followed by the EOF, may result in a better reconstruction.

Key words: FDA, EOF, gap filling.

1 Introduction

Measurement errors and instruments misfunction are the main causes of unreliable or missing values in real data sets. During recent years a great numbers of methods have been developed for filling missing data by objective analysis or optimal interpolation; the problem with such an approach is that it needs information about the correlation function of the data [5]. Since the most commonly used approaches contain some subjective components, any method that would not need such additional choices could be of interest. One of the emerging approaches in this direction is the Empirical Orthogonal Function (EOF) methodology [1], which is a deterministic method allowing a linear, continuous projection to a high-dimensional space

Department of Statistical and Mathematical Sciences, University of Palermo. Viale delle Scienze, Ed.13, 90128 Palermo, Italy.
e-mail: antonella.plaia@unipa.it, francesca.disalvo@unipa.it, mariantonietta.ruggieri@unipa.it, gianna.agro@unipa.it

useful for doing continuous interpolation of missing values. The EOF analysis has been widely used for oceanographic and meteorological applications to fill in missing data in spatiotemporal univariate data sets. In this paper, the EOF is extended to spatiotemporal multivariate data set related to concentrations of four main pollutants hourly recorded in 9 monitoring stations in Palermo (Italy) during 2005. The method is applied jointly with the Functional Data Analysis (FDA) approach [2, 3], a useful denoising tool, allowing to convert time series gathered as discrete observations into functional data. Such a conversion has the advantage of reducing a great number of observations to few coefficients, by preserving their functional structure and their temporal pattern. Observed data are preliminarly standardized by linear interpolation, according to EU directives, to make comparisons among different pollutants possible [4]. The FDA approach and the EOF methodology are briefly described in Section 2 and in Section 3, respectively. When high percentages of missing values are considered, very little is specified about the pattern of missing data. To indagate what happens to an imputation method if missing data are consecutive, that is if long gap sequences (2 or 3 months long) occur, two different procedures are proposed and compared in Section 4. In spatiotemporal data sets on air pollution, long gaps may be caused by serious damages to monitoring stations or data from fixed and mobile stations that need to be integrated. Obtained results are reported and commented in Section 5.

2 The FDA approach

In the FDA approach [2, 3] observed data are assumed to be realizations of continuous functions, so the generic observation $x_{is}^{p_j}$, recorded at time s ($s = 1,...,T$) for the pollutant p_j ($j = 1,...,P$) at the station i ($i = 1,...,N$), is considered as the result of a signal $\tilde{x}_i^{p_j}(t)$ affected by a noise $\varepsilon_{is}^{p_j}$:

$$x_{is}^{p_j} = \tilde{x}_i^{p_j}(t) + \varepsilon_{is}^{p_j}. \tag{1}$$

The curves $\tilde{x}_i^{p_j}(t)$ may be fitted by spline functions with coefficients c_{i,k,p_j} and bases ϕ_k:

$$\tilde{x}_i^{p_j}(t) = \sum_k^K c_{i,k,p_j}\phi_k(t). \tag{2}$$

We choose the cubic B-spline basis for each of the considered pollutants and equally spaced knots. A smoothing parameter $\lambda \geq 0$ penalizes the curvature of the functions (2), depending also by the number of bases K. Details about smoothing strategies for functional data are reported in [2].

3 The EOF methodology

The classical EOFs [1] are computed by standard Singular Value Decomposition (SVD) technique, that uses only the first ν sorted to decreasing order singular values, and the corresponding vectors, to reconstruct the data matrix; in fact, it is assumed that the vectors corresponding to the largest singular values hold more signal than noise with respect to the ones corresponding to the smallest values. The EOF procedure starts after an initial rough missing values replacement; it may be, for example, a mean value of the entire data set or the row or column mean, according to the data structure. A linear or polynomial fitting may be also considered. Then, the EOFs are used to make the reconstruction and the process is repeated until a convergence criterion is fulfilled. At each step, in order not to lose any information, only missing values are replaced with the values from the reconstruction.

4 The proposed procedures

Two different procedures are here proposed and compared: in the first one, observed data are reconstructed by EOF and then converted into functional ones; in the second one, observed data are transformed into functional ones and then EOF reconstruction is applied. The aim is to investigate if a preliminary functional trasformation allows a better reconstruction by EOF in presence of long gap sequences.

It is known that the accuracy and reliability of an imputation procedure depends on the pattern of the gaps as well as on their relative size with respect to the considered data set. In order to test the proposed procedures, we simulate 100 missing data indicator arrays $\mathbf{M}$, containing very long gap sequences (of about 3 months), randomly positioned according to pollutant and station, as they may frequently occur in real cases.

We start from the observed array $\mathbf{X}$, where actual missing values are filled by the annual mean, fixing the station and the pollutant. Data are purified from noise by converting $\mathbf{X}$ into the functional array $\tilde{\mathbf{X}}$ then, from now on, $\tilde{\mathbf{X}}$ will be used as a reference against which we will compare missing values imputed by the two different procedures below described. Before applying the two different procedures, each of the simulated array $\mathbf{M}$ is laid upon $\mathbf{X}$ by creating "artificial" missing values and obtaining new arrays $\mathbf{X^M}$. Artificial missing in $\mathbf{X^M}$ are initially filled by the annual mean, fixing the station and the pollutant, then:

- the EOF procedure is applyed on each array $\mathbf{X^M}$ and the reconstructed arrays $\hat{\mathbf{X}}^{\mathbf{M}}_{EOF}$ are converted into functional obtaining arrays $\hat{\mathbf{X}}^{\mathbf{M}}_{P1}$ (**Procedure 1**);
- each array $\mathbf{X^M}$ is converted into the functional array $\tilde{\mathbf{X}}^{\mathbf{M}}$ and the EOF procedure is applyed on $\tilde{\mathbf{X}}^{\mathbf{M}}$ obtaining reconstructed arrays $\hat{\mathbf{X}}^{\mathbf{M}}_{P2}$ (**Procedure 2**).

Arrays $\hat{\mathbf{X}}^{\mathbf{M}}_{P1}$ and $\hat{\mathbf{X}}^{\mathbf{M}}_{P2}$, obtained by the Procedure 1 and 2, respectively, are compared with $\tilde{\mathbf{X}}$ by means of the following distance between curves:

$$D_{i,p_j} = (\mathbf{c}_{i,p_j} - \mathbf{c}^{P.}_{i,p_j})' \mathbf{W} (\mathbf{c}_{i,p_j} - \mathbf{c}^{P.}_{i,p_j}), \tag{3}$$

where $\mathbf{c}_{i,p_j}$ is the vector of the K basis function coefficients of $\tilde{\mathbf{X}}$ and $\mathbf{c}^{P.}_{i,p_j}$ that one of $\hat{\mathbf{X}}^{\mathbf{M}}_{P1}$ or $\hat{\mathbf{X}}^{\mathbf{M}}_{P2}$; $\mathbf{W}$ has elements $W_{l,m} = \int \phi_l(t)\phi_m(t)dt$.

Moreover, by considering only imputed values from each matrix $\hat{\mathbf{X}}^{\mathbf{M}}_{P1}$ and $\hat{\mathbf{X}}^{\mathbf{M}}_{P2}$, these are compared with the corresponding values in $\tilde{\mathbf{X}}$ by means of the coefficient of correlation (ρ) and the root mean square deviation (*RMSD*, the lower the better).

5 First results

On the basis of partial results, Procedure 2 (P2) seems to outperform Procedure 1 (P1); as an example, on a single matrix, we have obtained $\rho_{P1} = 0.81$, $\rho_{P2} = 0.86$ and $RMSD_{P1} = 7.41$, $RMSD_{P2} = 2.43$. More complete results for each of the considered procedures, including the distributions of the distances (3) and of the two performance indicators (ρ and *RMSD*), over the 100 matrices $\mathbf{M}$, summarized by their means $\hat{\mu}$ and standard deviations $\hat{\sigma}$, will be reported in the final version of the paper. We will expect that Procedure 2 provides a better reconstruction when long gap sequences are observed in a multivariate space-time data set. Such a situation is quite frequent in air pollution data as long time failures may occur for some monitoring instruments or data from a mobile station have to be integrated. It is worth highlighting that if an imputation procedure is retained valid for time series presenting consecutive gaps, then it may also be used to solve forecasting problems.

References

1. Beckers, J.M. and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. J. Atmos. Oceanic Technol. **20**, 12, 1839–1856 (2003)
2. Ramsay, J.O. and Silverman, B.W.: Applied Functional Data Analysis. Springer-Verlag (2002)
3. Ramsay, J.O. and Silverman, B.W.: Functional Data Analysis. Second Edition. Springer-Verlag (2005)
4. Ruggieri, M. and Plaia, A.: An aggregate AQI: comparing different standardizations and introducing a variability index. Sci total environ. (2012) doi: 10.1016/j.scitotenv.2011.09.019
5. Sorjamaa, A., Lendasse, A., Cornet, Y., Deleersnijder, E.: An improved methodology for filling missing values in spatiotemporal climate data set. Comput Geosci. **14**, 1, 55–64 (2009)