This article was downloaded by: [University of Connecticut]
On: 28 October 2014, At: 10:12
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Functional principal component analysis for the explorative analysis of multisite-multivariate air pollution time series with long gaps

Mariantonietta Ruggieri [a], Antonella Plaia [a], Francesca Di Salvo [a] & Gianna Agró [a]

[a] Department of Statistical and Mathematical Sciences 'S. Vianelli', University of Palermo., Viale delle Scienze - Ed. 13, 90128, Palermo, Italy

Published online: 21 Dec 2012.

To cite this article: Mariantonietta Ruggieri, Antonella Plaia, Francesca Di Salvo & Gianna Agró (2013) Functional principal component analysis for the explorative analysis of multisite-multivariate air pollution time series with long gaps, Journal of Applied Statistics, 40:4, 795-807, DOI: 10.1080/02664763.2012.754852

To link to this article: http://dx.doi.org/10.1080/02664763.2012.754852

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &

Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [University of Connecticut] at 10:12 28 October 2014

Journal of Applied Statistics, 2013
Vol. 40, No. 4, 795–807, http://dx.doi.org/10.1080/02664763.2012.754852

Taylor & Francis
Taylor & Francis Group

Functional principal component analysis for the explorative analysis of multisite–multivariate air pollution time series with long gaps

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

Department of Statistical and Mathematical Sciences 'S. Vianelli', University of Palermo. Viale delle Scienze – Ed. 13, 90128 Palermo, Italy

(*Received 28 June 2012; final version received 28 November 2012*)

The knowledge of the urban air quality represents the first step to face air pollution issues. For the last decades many cities can rely on a network of monitoring stations recording concentration values for the main pollutants. This paper focuses on functional principal component analysis (FPCA) to investigate multiple pollutant datasets measured over time at multiple sites within a given urban area. Our purpose is to extend what has been proposed in the literature to data that are multisite and multivariate at the same time. The approach results to be effective to highlight some relevant statistical features of the time series, giving the opportunity to identify significant pollutants and to know the evolution of their variability along time. The paper also deals with missing value issue. As it is known, very long gap sequences can often occur in air quality datasets, due to long time failures not easily solvable or to data coming from a mobile monitoring station. In the considered dataset, large and continuous gaps are imputed by empirical orthogonal function procedure, after denoising raw data by functional data analysis and before performing FPCA, in order to further improve the reconstruction.

Keywords: air quality; functional data analysis; three-mode FPCA; EOF

Downloaded by [University of Connecticut] at 10:12 28 October 2014

1. Introduction

In recent years, quantifying air quality and, most of all, following the evolution of pollution, has been becoming a fundamental issue for local and central governments. The need to inform the population about the air quality and related health outcome has led to a proliferation of synthetic indicators, giving an idea of the state of pollution in a day [13]. Otherwise, the necessity to study the effects of pollution reduction policies, implemented by governments, asks for a study of space–time evolution of pollution. In this paper, air quality data for the city of Palermo (Italy)

*Corresponding author. Email: mariantonietta.ruggieri@unipa.it

© 2013 Taylor & Francis

are analyzed. Four pollutants are measured daily at nine monitoring sites over a period of 2 years (2005–2006). After a preliminary analysis, air pollutant concentrations are treated as functional data. Converting time series, recorded as discrete observations, into functional data preserves their functional structure and presents the advantage of reducing a great number of observations to a few coefficients and solving missing data issue at the same time. Actually, although functional data analysis (FDA) has a good performance in missing data reconstruction when short gaps occur, empirical orthogonal function (EOF) outperforms FDA in the presence of long gaps, as demonstrated in a previous work [15], especially when functional data are estimated with low smoothness. Therefore, the EOF procedure [1] is used here to fill in long gap sequences, while FPCA [14] allows to identify significant pollutants by tracking the evolution of their variability along time. FPCA proceeds in a manner analogous to the conventional PCA: the objective is to determine mutually orthogonal linear combinations of the original variables that maximize the explained variation; this is achieved by an eigenvector decomposition of the variance operator yielding, in the FPCA approach, eigenfunctions that vary with time.

The paper is organized as follows: Section 2 presents the air pollution dataset, its pretreatment and some preliminary analyses; Section 3 outlines the main characteristics of the FPCA and the computational aspects of the adopted methodologies; Section 4 describes the role of the EOF approach in reconstructing functional data, when long gaps occur; Section 5 shows and comments the obtained results and finally, Section 6 deals with some conclusions and further developments.

Downloaded by [University of Connecticut] at 10:12 28 October 2014

2. Air pollution data

In this paper air pollution-validated data, provided by AMIA (Azienda Municipalizzata Igiene Ambientale, http://www.amianet.it/), recorded in Palermo for 2 whole years, 2005 and 2006, are analyzed.

The city of Palermo stretches along the northwest coastline of Sicily, the island situated in the south of Italy, overlooking the Tyrrhenian Sea. It covers an area of about 160 km^2 with a population of approximately 700,000. The climate is typically Mediterranean, which is temperate with warm dry summers and rainy winters. Palermo is not an industrial city and air pollution is caused especially by the heavy traffic.

The concentrations of four main pollutants (CO, NO_2, PM_{10} and SO_2), recorded hourly (or two-hourly) at nine monitoring stations (Figure 1), are considered here.

To aggregate hourly (or bihourly) values in order to obtain a daily synthesis at each site for each pollutant, we use the functions suggested by EU guidelines [6–8], applied if at least 75% of values is available on each day.

To compare pollutants with different measurement units or order of magnitude we adopt, as the US Environmental Protection Agency does [5], the standardization by linear interpolation [12], with breakpoints modified according to EU standards and directives [11]. This choice is dictated by many epidemiological studies, showing that air pollution cause long-term as much as short-term adverse health response; therefore, low pollutant concentrations should be also taken into account.

Missing values are, at this stage, replaced with the station mean. Then, before performing FPCA, a more sophisticated imputation technique, based on the EOF methodology (cf. Section 3), is considered.

Time series of pollutant standardized concentrations aggregated by months are reported in Figure 2. Such a synthesis allows to better distinguish the behaviour of each station along time and makes the comparison easier. In Figure 2, DB appears to be the most polluted station, while Bo and Cep seem to be the least polluted, for almost all the considered periods, with regard to

Downloaded by [University of Connecticut] at 10:12 28 October 2014

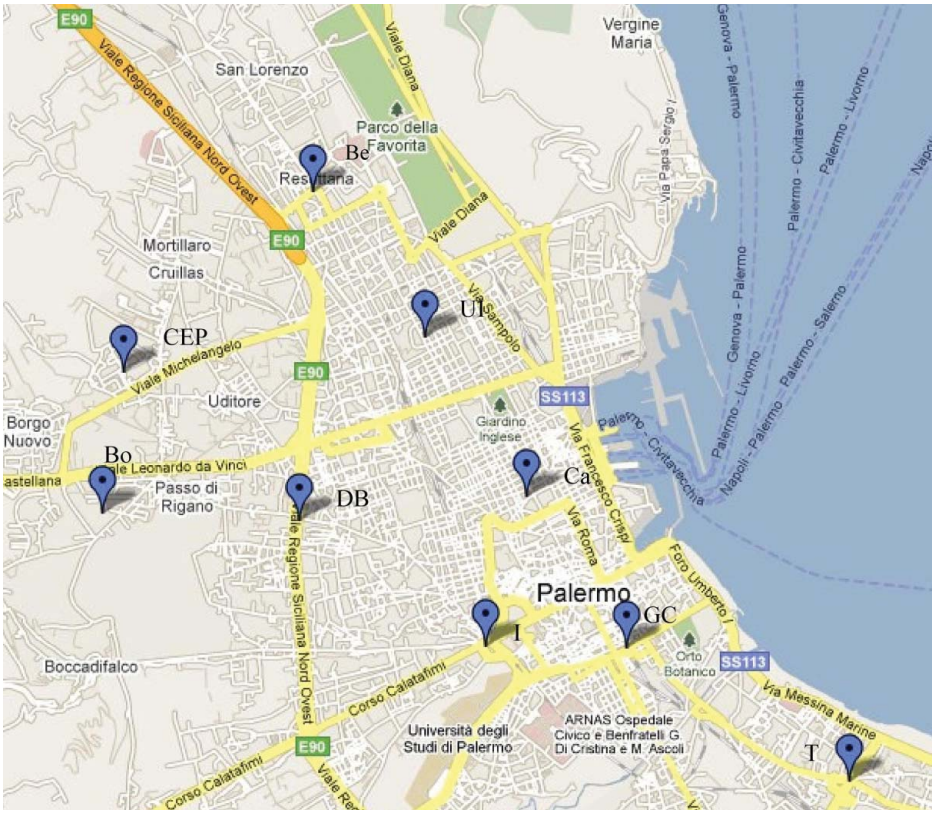

Figure 1. Air monitoring sites.

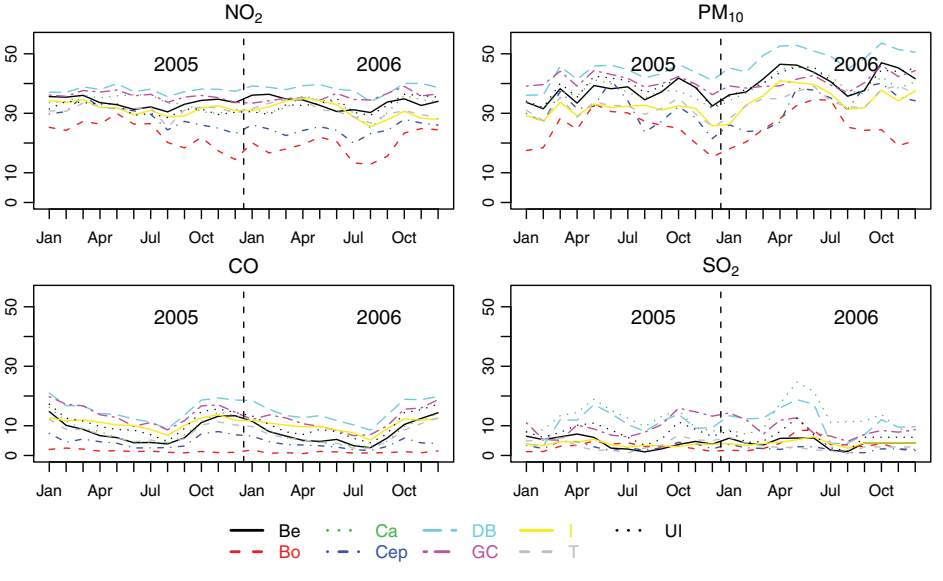

Figure 2. Standardized daily concentrations aggregated by month.

PM_{10}, NO_2 and CO. SO_2 reaches the highest values in Ca and the lowest in Bo, Cep and T. Spatial differences may be due to local traffic conditions as much as emission points of a particular pollutant and possible urban canyon effects. PM_{10} creates the most severe problems in Palermo:

Table 1. Correlation matrices between pollutants by station.

	NO_2	PM_{10}	CO	SO_2	NO_2	PM_{10}	CO	SO_2	NO_2	PM_{10}	CO	SO_2
	Be				Bo				Ca			
NO_2	1.00	0.33	0.58	0.50	1.00	0.38	0.58	0.40	1.00	0.37	0.24	0.48
PM_{10}	0.33	1.00	0.34	0.33	0.38	1.00	0.30	0.48	0.37	1.00	0.08	0.44
CO	0.58	0.34	1.00	0.39	0.58	0.30	1.00	0.24	0.24	0.08	1.00	−0.05
SO_2	0.50	0.33	0.39	1.00	0.40	0.48	0.24	1.00	0.48	0.44	−0.05	1.00
	Cep				DB				GC			
NO_2	1.00	0.43	0.48	0.48	1.00	0.47	0.58	0.57	1.00	0.46	0.47	0.33
PM_{10}	0.43	1.00	0.37	0.29	0.47	1.00	0.34	0.47	0.46	1.00	0.43	0.31
CO	0.48	0.37	1.00	0.59	0.58	0.34	1.00	0.36	0.47	0.43	1.00	0.46
SO_2	0.48	0.29	0.59	1.00	0.57	0.47	0.36	1.00	0.33	0.31	0.46	1.00
	I				T				UI			
NO_2	1.00	0.38	0.53	0.38	1.00	0.29	0.39	0.37	1.00	0.44	0.49	0.37
PM_{10}	0.38	1.00	0.32	0.33	0.29	1.00	0.34	0.16	0.44	1.00	0.39	0.37
CO	0.53	0.32	1.00	0.27	0.39	0.34	1.00	0.51	0.49	0.39	1.00	0.43
SO_2	0.38	0.33	0.27	1.00	0.37	0.16	0.51	1.00	0.37	0.37	0.43	1.00

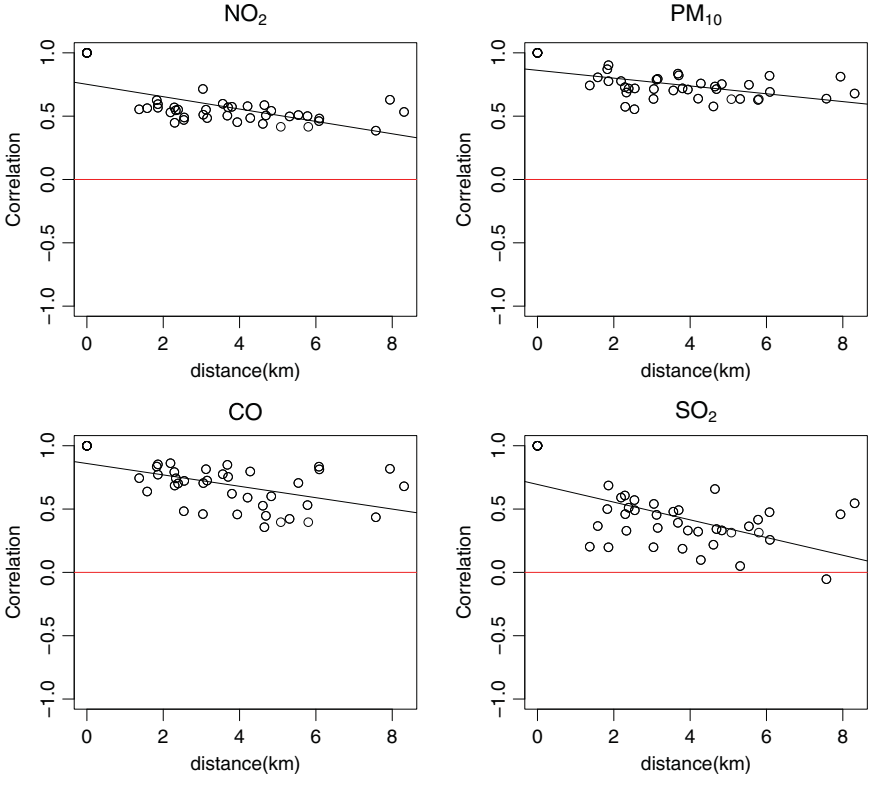

Figure 3. Spatial correlogram by pollutant.

Downloaded by [University of Connecticut] at 10:12 28 October 2014

it reached concentrations greater than 50 (threshold value) in DB in 2006. Also NO_2 presents higher concentrations than other pollutants for almost all the stations during 2 years.

Table 1 reports the correlation matrices between pollutants at the nine stations. These correlations may be due to the processes by which they are formed, as combustion processes almost entirely come from road traffic emissions. High correlations depend also on the strong relationship existing among concentrations of pollutants and some meteorological variables, such as temperature, rain and wind (speed and direction) that may influence the production of secondary pollutants.

It may be noted that the correlations between pollutants are almost always positive and roughly constant across the stations. There is also a very high spatial correlation between the readings for each of the considered pollutants. As we would expect, the closer the sites are, the higher the correlation between them is, as shown in Figure 3, where correlations versus distances between sites are plotted for each pollutant. Such a relationship appears to be quite similar for all the pollutants, but it is more evident for NO_2 and PM_{10}.

Downloaded by [University of Connecticut] at 10:12 28 October 2014

3. The FPCA approach

Our main purpose is to investigate temporal variation by means of FPCA and provide a comprehensible representation of the variability of the monitored pollutants.

> Repeated measurements taken over time on an object are often analyzed using longitudinal data analysis. FDA to some extent represents an alternative that is of most interest when the measurements taken are fairly accurate and when there are generally more data available, so fewer assumptions and constraints on the nature of the profile or curve are needed. The functional data analytic approach is to treat the whole curve as a single entity and not be concerned about correlations between the repeated measurements. (Henderson [9])

According to FDA, the observed concentration, $y_{ir}^{p_j}$, for the pollutant p_j ($j = 1, \ldots, P$) and the station i ($i = 1, \ldots, N$), is considered as a discrete realization, recorded at time r ($r = 1, \ldots, T$) of a continuous process [14]. To transform discrete data into curves, a smoothing procedure is required, so the existence of a 'signal plus noise' is assumed in the observed data:

$$y_{ir}^{p_j} = x_i^{p_j}(t) + \varepsilon_{ir}^{p_j}, \tag{1}$$

where $x_i^{p_j}(t)$ is the functional datum and $\varepsilon_{ir}^{p_j}$ is a random error. As it is known, the functional dataset is represented by a linear expansion in terms of a basis system ϕ_k (in the paper we use cubic B-splines, as described in Section 4) that we assume to be unique for all the considered pollutants:

$$x_i^{p_j}(t) = \sum_{k=1}^{K} c_{i,k,p_j} \phi_k(t). \tag{2}$$

The estimates $\tilde{x}_i^{p_j}(t)$ of $x_i^{p_j}(t)$ are computed through least-squares smoothing with a roughness penalty defined in terms of integrated squared second derivatives, a scalar $\lambda \geq 0$ controls the amount of smoothing.

FPCA is the most widely used technique to decompose variation in functional data by exploring the space of the variables and finding directions in the observation space along which data have the highest variability. The term *mode of variation* identifies some directions of interest. In [14], this goal is achieved by an eigenanalysis of the variance operator yielding eigenfunctions $\xi_m(t)$ that vary with time.

The mth eigenvalue ρ_m quantifies the variation along the mth principal direction and the ratio $\rho_m / \sum_m \rho_m$ measures the proportion of variability explained by the mth principal component (PC).

The resulting mth eigenfunction $\xi_m(t)$, named *harmonic*, helps to find interesting components of variability in the functional data. As it is known, the eigenfunctions are orthonormal:

$$\langle \xi_m\, \xi_n \rangle = \int \xi_m(t)\, \xi_n(t)\, \mathrm{d}t = 0 \quad \text{and} \quad ||\xi_m||^2 = \int \xi_m^2(t)\, \mathrm{d}t = 1,$$

so the mth one catches the most relevant mode of variation along a direction orthogonal to all the other modes. The harmonics can be splitted into P sub-eigenfunctions, one for each pollutant:

$$\xi_m(t) = [\xi_m^{p_1}(t), \xi_m^{p_2}(t), \xi_m^{p_3}(t), \xi_m^{p_4}(t)]. \tag{3}$$

For the element $\xi_m^{p_j}$ of the mth eigenfunction, the norm $||\xi_m^{p_j}||^2$ is the proportion of the variability in the mth PC ascribable to pollutant p_j.

The score for station i on the mth PC is

$$f_{im} = \sum_j f_{im}^{p_j} \quad \text{with } f_{im}^{p_j} = \int \xi_m^{p_j}(t) \tilde{x}_i^{p_j}(t)\, \mathrm{d}t. \tag{4}$$

3.1 *Computational aspects for three-mode functional PCs*

The computational procedure for a three-mode FPCA is based on the approximation of the functional eigenanalysis and this goal may be achieved by defining the continuous eigenfunction in terms of the centered smoothed functional data.

The computational aspects, described for the univariate case in the literature [14, Section 8.4], are here generalized for the three-mode analysis. The procedure is outlined as follows:

(a) let us define $x_i^{p_j}(t)$ in terms of its linear expansion (2);
(b) let us define the covariance functions between pollutants in matrix terms:

$$v^{p_j,p_l}(t,t') = N^{-1}\Phi(t)'\mathbf{C}'_{p_j}\mathbf{C}_{p_l}\Phi(t'), \tag{5}$$

where Φ is the vector with elements ϕ_k, and $\mathbf{C}_{p_j}$ is the bidimensional slice extracted, by fixing p_j, from the three-dimensional array $\mathbf{C}$ ($N \times K \times P$) with elements c_{i,k,p_j};
(c) by expressing the eigenfunction $\xi^{p_j}(t)$ in terms of the basis functions system

$$\xi^{p_j}(t) = \Phi(t)'\mathbf{b}^{p_j} \tag{6}$$

and defining the order K symmetric matrix

$$\mathbf{W} = \int \Phi(t)\Phi(t)'\, \mathrm{d}t \tag{7}$$

that can be computed by numerical integration, the eigenequation system can be written out as

$$\begin{aligned}
&\int v^{p_1,p_1}(t,t')\xi^{p_1}(t)\,\mathrm{d}t + \cdots \int v^{p_1,p_j}(t,t')\xi^{p_j}(t)\,\mathrm{d}t \cdots + \int v^{p_1,p_P}(t,t')\xi^{p_P}(t)\,\mathrm{d}t = \rho\xi^{p_1}(t),\\
&\int v^{p_2,p_1}(t,t')\xi^{p_1}(t)\,\mathrm{d}t + \cdots \int v^{p_2,p_j}(t,t')\xi^{p_j}(t)\,\mathrm{d}t \cdots + \int v^{p_2,p_P}(t,t')\xi^{p_P}(t)\,\mathrm{d}t = \rho\xi^{p_2}(t),\\
&\int v^{p_3,p_1}(t,t')\xi^{p_1}(t)\,\mathrm{d}t + \cdots \int v^{p_3,p_j}(t,t')\xi^{p_2}(t)\,\mathrm{d}t \cdots + \int v^{p_3,p_P}(t,t')\xi^{p_P}(t)\,\mathrm{d}t = \rho\xi^{p_3}(t),\\
&\vdots\\
&\int v^{p_P,p_1}(t,t')\xi^{p_1}(t)\,\mathrm{d}t + \cdots \int v^{p_P,p_2}(t,t')\xi^{p_P}(t)\,\mathrm{d}t \cdots + \int v^{p_P,p_P}(t,t')\xi^{p_P}(t)\,\mathrm{d}t = \rho\xi^{p_P}(t).
\end{aligned}$$

Downloaded by [University of Connecticut] at 10:12 28 October 2014

In the previous system, by substituting Equations (5) and (7) each integral becomes

$$\int v^{p_j,p_l}(t,t')\xi^{p_j}(t)\,\mathrm{d}t = N^{-1}\Phi(t)'\mathbf{C}'_{p_j}\mathbf{C}_{p_l}\mathbf{W}\mathbf{b}^{p_j} \tag{8}$$

and with Equation (6), the generic equation in the system becomes

$$N^{-1}\mathbf{\Phi}(t)'\mathbf{C}'_{p_j}\,\mathbf{C}_{p_1}\,\mathbf{W}\mathbf{b}^{p_j} + \cdots + \cdots + N^{-1}\mathbf{\Phi}(t)'\mathbf{C}'_{p_j}\mathbf{C}_{p_P}\,\mathbf{W}\mathbf{b}^{p_j} = \rho\mathbf{\Phi}(t)'\mathbf{b}^{p_j} \tag{9}$$

for $j, l = 1, \ldots, P$, where the solution must hold for each s;

(d) in order to solve the eigenequation system, we define:

$$\tilde{\mathbf{V}}^{p_j p_l} = N^{-1}\mathbf{C}'_{p_j}\mathbf{C}_{p_l}\mathbf{W} \tag{10}$$

and the block matrix $\tilde{\mathbf{V}}_{[(K\times P),(K\times P)]}$:

$$\tilde{\mathbf{V}} = \begin{vmatrix} \tilde{\mathbf{V}}^{p_1p_1} & \tilde{\mathbf{V}}^{p_1p_2} & \cdots & \tilde{\mathbf{V}}^{p_1p_P} \\ \tilde{\mathbf{V}}^{p_2p_1} & \tilde{\mathbf{V}}^{p_2p_2} & \cdots & \tilde{\mathbf{V}}^{p_2p_P} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{\mathbf{V}}^{p_Pp_1} & \tilde{\mathbf{V}}^{p_Pp_2} & \cdots & \tilde{\mathbf{V}}^{p_Pp_P} \end{vmatrix}.$$

(e) the $h \leq N$ solutions are found by solving the linear system:

$$\tilde{\mathbf{V}}[\mathbf{b}^{p_1}, \mathbf{b}^{p_2}, \ldots, \mathbf{b}^{p_P}]^{\mathrm{T}} = \rho[\mathbf{b}^{p_1}, \mathbf{b}^{p_2}, \ldots, \mathbf{b}^{p_P}]^{\mathrm{T}}. \tag{11}$$

The mth resulting $(K \times P)$ vector of coefficients $\mathbf{b}_m$ can be splitted into P parts and the eigenfunctions are computed as

$$\xi_m^{p_j}(t) = \Phi(t)'\mathbf{b}_m^{p_j}. \tag{12}$$

3.2 *Computational aspects in deriving EOFs by FPCA and FSVD*

After obtaining the functional PCs and the harmonics, it is straightforward to compute an approximation of the functional data, as an expansion in terms of these EOFs. Choosing the number $h \leq N$ of PCs, that explain an appropriate portion of variance of the functional data, the approximation is

$$\widehat{x}_i^{p_j}(t) = \sum_{m=1}^{h} f_{im}\xi_m^{p_j}(t), \tag{13}$$

where f_{im} are the scores (4) and the functions $\xi_m^{p_j}(t)$ are the EOFs (12).

It is worth noting that the same results of the FPCA can be achieved through a functional singular value decomposition (FSVD).

Here, we derive the computational method for FSVD in the three-mode analysis.

In the classical SVD, we look at the decomposition of a matrix data $\mathbf{Z}$, $\mathbf{Z} = \mathbf{U\Gamma A}$, and this leads to results similar to the ones from the eigenanalysis of the variance matrix V, such that $\mathbf{Z}'\mathbf{Z} = N^{-1}\mathbf{V}$. The steps for the three-mode FPCA, formalized in Section 3.1, allow a straightforward definition of the correspondent matrix $\mathbf{Z}$ in FSVD.

Downloaded by [University of Connecticut] at 10:12 28 October 2014

Referring to Equation (10), let us define

$$\mathbf{Z}^{p_j} = N^{-1}\mathbf{C}_{p_j}\mathbf{W}\mathbf{L},$$

where $\mathbf{L}$ is the inverse matrix of the Cholesky decomposition of $\mathbf{W}$, i.e. $\mathbf{L} = (W^{1/2})^{-1}$.

With this definition of $\mathbf{Z}^{p_j}$, in terms of the basis system and the coefficients related to the pollutant p_j, we have $\tilde{\mathbf{V}}^{p_j p_l} = \mathbf{Z}^{p_j}\mathbf{Z}^{p_l}$ so that a factorization of $\tilde{\mathbf{V}}$, i.e. $\mathbf{Z}'\mathbf{Z} = N^{-1}\tilde{\mathbf{V}}$, is obtained with $\mathbf{Z}$:

$$\mathbf{Z} = |\mathbf{Z}^{p_1}, \mathbf{Z}^{p_2}, \mathbf{Z}^{p_3}, \ldots, \mathbf{Z}^{p_P}|.$$

FSVD in this formulation is the singular value decomposition of an appropriate matrix defined in terms of the basis system and the coefficients of the functional data.

By selecting the first $\nu \leq \text{rank}(\mathbf{Z})$ eigenvalues of the matrix $\mathbf{\Gamma}$, we can obtain an approximation of $\mathbf{Z}$:

$$\mathbf{Z}_{N\times(K\times P)} \approx \mathbf{U}_{N\times\nu}\mathbf{\Gamma}_{\nu\times\nu}\mathbf{A}_{\nu\times(K\times P)}. \tag{14}$$

As in the non-functional context, from the relationship between FPCA and FSVD, it follows that the standardized principal scores (4) are the columns of the matrix $\mathbf{U}$; the coefficients of the harmonics (12) can be obtained by premultiplying $\mathbf{A}'$ by $\mathbf{L}^{-1} = \mathbf{W}^{1/2}$.

The singular values $\sqrt{\rho_m}$ are the elements of the diagonal matrix $\mathbf{\Gamma}$.

4. EOF reconstruction for long gaps

The presence of missing values, mainly due to the jam of the stations in monitoring networks, is a crucial concern in air pollution datasets: data need to be adequately preprocessed before performing any analysis in order to lose as less as possible meaningful information about their temporal and spatial characteristics.

Imputing missing data by a simple average has the advantage of not needing any a priori assumption or complex calculations, but presents the drawback of assuming all of the observations as equally important and does not consider their patterns. Although FDA may solve the problem of missing values when short gaps occur, the reconstruction by EOFs is a more valid approach, especially for time series exhibiting long gap sequences, as demonstrated in a previous work [15]. This is our case, where SO_2 concentrations are missing on four stations for four consecutive months (from September to December) in 2006.

We start from the observed multivariate space–time array $\mathbf{Y}$, where missing values are first filled by the annual mean, fixing the station and the pollutant. Since correlations among pollutants have been observed in each station (cf. Table 1), in order to take into account the simultaneous variability of all the pollutants, we have used a multivariate approach. Then, data are purified from noise by converting $\mathbf{Y}$ into the functional array $\tilde{\mathbf{X}}$; the cubic B-spline basis system is chosen with 179 equally spaced interior knots, in order to have intervals of a few days, obtaining $K = 183$, number of basis functions.

Since a multivariate approach is considered, a unique basis system is necessary for all the pollutants; anyway, the high number $K = 183$ of parameters allows to characterize the different behavior of each pollutant.

Choosing equally spaced knots seems appropriate as pollutant time series are gathered regularly and are characterized by seasonal, monthly, weekly and daily variations [10]; the high number of knots chosen aims at fitting almost daily variations.

The chosen value for the smoothing parameter ($\lambda = 2$) seems to be a right compromise between what is suggested by an automatic method, here the generalized cross-validation is considered [3], and a subjective choice that aims at smoothing rough data without losing too much variability,

Downloaded by [University of Connecticut] at 10:12 28 October 2014

in order to take into account the peaks recorded in pollutant concentrations. Indeed, here the EOF method gives better results with $K = 183$ and $\lambda = 2$.

Once obtaining $\tilde{\mathbf{X}}$, $\nu = 3$ EOFs are extracted from $\tilde{\mathbf{X}}$ on the basis of the total variability explained (more than 90%). Then, EOFs are used to reconstruct $\tilde{\mathbf{X}}$, according to

$$\widehat{x}_i^{p_j}(t) = \sum_{m=1}^{3} f_{im}\xi_m^{p_j}(t), \tag{15}$$

where the functions $\xi_m^{p_j}(t)$ are the EOFs (3) and f_{im} are the scores (4).

In order not to lose any information, only missing values are replaced with the values from the reconstruction.

Finally, FPCA is performed on the new reconstructed array.

The entire analysis is implemented in R by using also *fda* package (http://cran.r-project.org).

Downloaded by [University of Connecticut] at 10:12 28 October 2014

5. Results and comments

Some recent contributions are intended to extend the classical methods for spatial–temporal analysis to a functional context [16]; in this framework, the proposed procedure can be considered as a first step to introduce multidimensionality. In fact, we deal with multisite and multivariate data at the same time, while in the literature only multisite or only multivariate data are considered. FDA takes the advantage to work with few coefficients, rather than a large amount of data, and to solve missing values issues as well. Actually, although FDA is a valid approach for this aim when short gaps occur, here we carry out the EOF procedure after denoising raw data by FDA, obtaining an improved reconstruction, especially in the presence of long gap sequences. It is worth noting that such an approach can be considered a useful tool to pre-process data before any spatio-temporal analysis, even when data should not be analyzed as functional.

As for the conventional PCA, the interpretation of the resulting functional PCs is not always a straightforward matter. Usually the first step is looking at the proportion of total variability explained by the PCs (Table 2). The first three PCs explain 90% of the total variability among stations; in particular, the first PC (PC1) explains 74% of the variability, while the second mode of variation (PC2) explains 11% and the third (PC3) only 5%. A deeper understanding about the meaning of the PCs is obtained by looking at the proportion of variability accounted for by variation of each pollutant. As it is highlighted in bold in Table 2, the most significant pollutants for PC1 are PM_{10} and NO_2, since they explain 43% and 28% of variability, respectively, while the major contribution to PC2 is given by SO_2 (68%). Finally, the pollutants PM_{10} (58%) and CO (22%) give the major contribution to PC3.

The second step is looking at the first three harmonics $\xi_m^{p_j}(t)$, for $m = 1, 2, 3$, displayed in Figure 4 for each pollutant p_j.

Table 2. Main results from Three-mode FPCA.

$\frac{\rho_m}{\sum_m^{\nu}\rho_m}$		$\|\|\xi_m^{p_1}\|\|^2$	$\|\|\xi_m^{p_2}\|\|^2$	$\|\|\xi_m^{p_3}\|\|^2$	$\|\|\xi_m^{p_4}\|\|^2$	$\sum_{j=1}^{4}\|\|\xi_m^{p_j}\|\|^2$
PC1	0.74	**0.28**	**0.43**	0.19	0.10	1
PC2	0.11	0.17	0.11	0.04	**0.68**	1
PC3	0.05	0.16	**0.58**	0.22	0.05	1
	0.90	$p_1 = NO_2$	$p_2 = PM_{10}$	$p_3 = CO$	$p_4 = SO_2$	

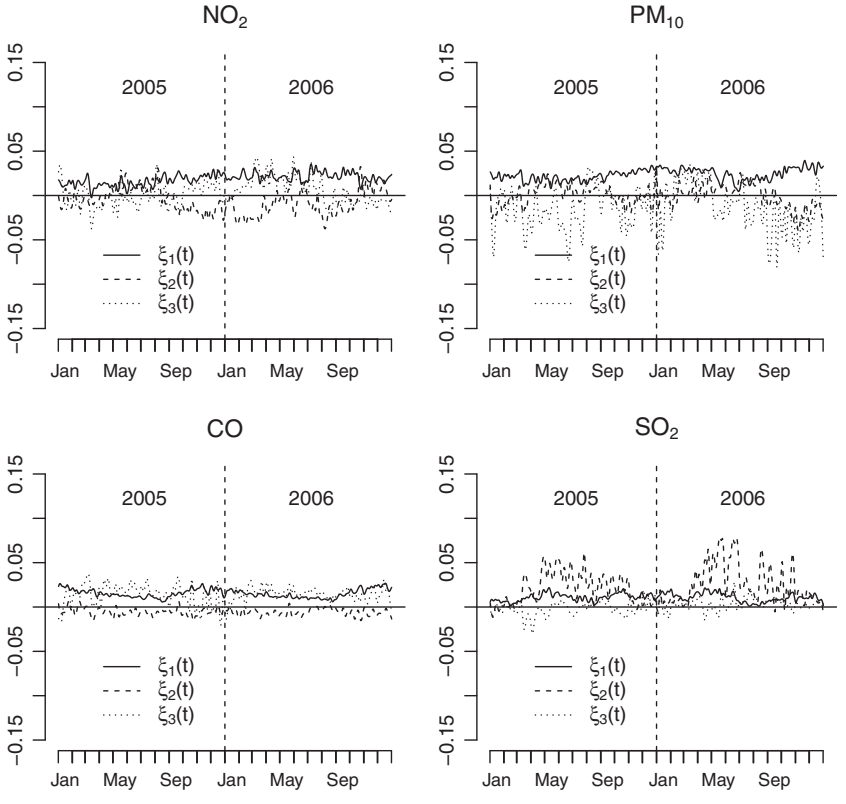

Figure 4. Plots of the first three harmonics by pollutant.

The harmonics, arising from the procedure, are very interesting from different points of view; they represent temporal evolution of the pollutant contributions to PCs and are a good synthesis of the variability among stations. Decreasing (increasing) weight functions $\xi(t)$ means decreasing (increasing) variability of pollutant concentrations and consequently homogeneous (non-homogeneous) air pollution.

The weight functions $\xi_1^{p_j}(t)$ (Figure 4), positive for each pollutant, have higher values with higher variability along time for PM_{10} and NO_2. This mode of variation for PM_{10} is high from September to April with lower values in May and June, especially for 2006, when high concentrations of PM_{10} occur and the variability among stations reduces dramatically, revealing high pollution for the whole urban area. Similar interpretation may be given to the positive contributes of the harmonics $\xi_1(t)^{NO_2}$, having the lowest contribution in March and at the end of May for 2005 and at the end of June and in October for 2006. Contributes of the other pollutants to PC1 are similar, but low and therefore of little interest. Thus, the first PC quantifies the general level of pollution.

The major contribution to PC2 is given by SO_2 (68%), and the weight function $\xi_2^{SO_2}(t)$ shows a great variability and positive values along time, with high level in the warmer period from March to October for both 2005 and 2006. In other words, the second mode catches increasing variability, during spring and summer, due primarily to high concentrations of SO_2 in some sites (Figure 2).

As already stated, the weight function $\xi_3(t)$ is relevant for PM_{10}. Interpreting this weight function as not straightforward, but looking at Figure 5 we can state that it accounts for the variability determined by occasional departures of some monitoring sites from their PM_{10} average

Downloaded by [University of Connecticut] at 10:12 28 October 2014

Table 3. PC scores.

	PC1	PC2	PC3
Be	67.86	−85.19	41.17
Bo	−484.57	109.16	−31.49
Ca	104.44	192.24	7.60
Cep	−250.92	−50.59	118.61
DB	393.59	45.20	8.19
GC	220.24	−25.63	−21.71
I	−57.99	−66.46	−127.32
T	−101.13	−108.69	−21.74
UI	108.48	−10.05	26.69

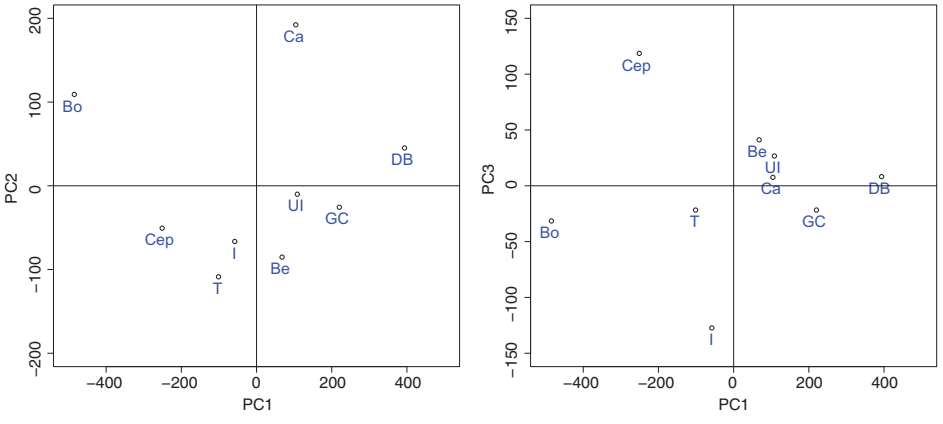

Figure 5. Plots of the first three PC scores.

level. This additional detail is obtained as a result of the choice $\lambda = 2$; a greater value of λ would determine a very flat weight function $\xi_3(t)$, hiding the variability linked to very high concentrations of PM_{10}.

Figure 4 shows a certain amount of roughness in the curves; alternatively, the regularized PCs analysis could be adopted by incorporating the smoothing in PCA itself. In general, this kind of procedure makes clearer the interpretation of the harmonic plots, but in our case it is more important to preserve the amount of variation in $\tilde{X}$ accounted for by the weight functions.

The scores of the nine monitoring stations, on the first three PCs, obtained by integrating the weight function against the functional datum (4), are presented in Table 3 and plotted in Figure 5.

As for conventional PCA, the score plot gives information about how much each monitoring station contributes to the whole variability: the further a point lies from the origin, the greater its contribution to the variation. Additionally, close stations tend to be correlated or have similar behavior.

Remembering that most of the variation is explained by the first PC, which represents the level of air pollution accounted for PM_{10} and NO_2, then the distribution of monitoring stations along the axis PC1 is of most interest. DB has the most polluted air with the highest score on PC1, while Bo has the lowest one. The remaining stations record a level of pollution near the mean of the town pollution (origin of axes).

Downloaded by [University of Connecticut] at 10:12 28 October 2014

High scores on the axis PC2 (Figure 5, left side) are assigned to stations with high SO_2 pollutant concentration in the warmer period: Ca station scored with the higher value while T with the lower one.

On the axis PC3 (Figure 5, right side), characterized mainly by PM_{10} variability, Cep is in the upper left corner because, despite its low yearly average value of PM_{10}, it has more occasional overcomings.

From a practical point of view, the obtained results could be useful in the management of air pollution avoiding cumbersome syntheses of multiple time series. In particular, the harmonic trends in Figure 4 allow to knows levels and variations of each pollutant along the two considered years, while the score plots in Figure 5 make the most polluted stations immediately evident and highlight those ones with similar behavior compared with the whole set of pollutants. This kind of information can help policy-makers to take common and proper action measures and to plan abatement strategies for the protection of environment and human health.

Downloaded by [University of Connecticut] at 10:12 28 October 2014

6. Conclusions and further developments

The application of FPCA, for studying the peculiarity of the evolution of recorded pollutant concentrations, is here extended to a multisite/multivariate case, using a dataset of four pollutants collected at nine monitoring sites in Palermo during 2 years. 'FDA appears to be a powerful exploratory technique for understanding and visualizing differences in non-linear air quality trends, and more widely other functional curves' [9].

The obtained results state that FPCA, applied to a large matrix of daily pollutant concentrations at several sites, can be used to quickly identify the key pollutants for air pollution, without doing hard syntheses through multiple time series plots. In particular, by means of the analysis of resulting harmonics it is easy to know the mode of variation of a pollutant and derive pollution levels during 2 years. The score plot, as in conventional PCA, allows to identify areas where higher pollution levels occur along the considered period, seasonally or only occasionally.

The advantage of using FPCA, instead of the conventional PCA, depends on the nature of the considered data; such an approach is just suggested by the functional structure of our data. FPCA preserves the functional structure of data, purifying them by random errors and, moreover, allows to deal with a few coefficients rather than a great number of observations without losing too much variability.

In the paper, the problem of missing data is also faced by considering the projection of $\tilde{X}$ onto the subspace spanned by $\xi_1(t)$, $\xi_2(t)$, $\xi_3(t)$, referred as EOFs; such a procedure allows an accurate reconstruction when long gap sequences occur as in our data.

As it is known, the spatial misalignment of data is a very common problem in environmental studies; in our case, for example, ozone (O_3) is recorded at two monitoring sites only and, consequently, it was not included in the analysis. The prediction (imputation) of pollutant concentrations, where they are not observed, would enable to realign data and consider a more exhaustive dataset on which, considering the seasonal behavior of O_3, we expect different conclusions. The application of a kriging algorithm [4] to impute O_3 concentrations, before performing FPCA, will be the object of a subsequent paper. Further developments will involve in the analysis also some climatic variables, such as wind and rain, as well as temperature and solar radiation that, as it is known [2], can influence air pollution.

Acknowledgements

The research was supported by a 2007 grant for Cooperation and International Relationships (CORI) by the University of Palermo ('Un indicatore aggregato della qualitá dell'aria'). We would like to thank the anonymous reviewer and the editor for their useful suggestions.

References

[1] J.M. Beckers and M. Rixen, *EOF calculations and data filling from incomplete oceanographic datasets*, J. Atmos. Ocean. Technol. 20 (2003), pp. 1839–1856.

[2] L.M. Caligiuri, G.D. Costanzo, and A. Reda, *The study of ground Ozone concentration levels : A functional analysis approach based on Principal Components Analysis, in Air pollution XIII: Thirteenth International Conference on Modelling, Monitoring and Management of Air Pollution*, Cordoba, Spain, WIT Transactions on Ecology and the Environment, Vol 82, WIT Press, Southampton, 2005, pp. 59–67. Available at http://www.witpress.com/contents/c0144.pdf

[3] P. Craven and G. Wahba, *Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math. 31 (1979), pp. 377–403.

[4] P. Delicado, R. Giraldo, C. Comas, and J. Mateu, *Statistics for spatial functional data: Some recent contributions*, Environmetrics 21 (2010), pp. 224–239.

[5] Environmental Protection Agency, *Guideline for reporting of daily air quality: Air quality index (AQI)*, United States Environmental Protection Agency, EPA-454/B-06-001, 2006.

[6] European Community, *CouncilDirective 1999/30/EC of 22 April 1999 relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air*, Official Journal, L 163, 29/06/1999 (1999), pp. 41–60.

[7] European Community, *Directive 2000/69/EC of the European Parliament and of the Council of 16 November 2000 relating to limit values for benzene and carbon monoxide in ambient air*, Official Journal L 313, 13/12/2000 (2000), pp. 12–21.

[8] European Community, *Directive 2002/3/EC of the European Parliament and of the Council of 12 February 2002 relating to ozone in ambient air*, Official Journal L 67, 9/3/2002 (2002), pp. 14–30.

[9] B. Henderson, *Exploring between site differences in water quality trends: A functional data analysis approach*, Environmetrics 17 (2006), pp. 65–80.

[10] R. Ignaccolo, S. Ghigo, and E. Giovenali, *Analysis of air quality monitoring networks by functional clustering*, Environmetrics 19 (2008), pp. 672–686.

[11] F. Murena, *Measuring air quality over large urban areas: Development and application of an air pollution index at the urban area of Naples*, Atmos. Environ. 38 (2004), pp. 6195–6202.

[12] W.R. Ott and W.F. Hunt, *A quantitative evaluation of the pollutant standards index*, J. Air Pollut. Control Assoc. 26 (1976), pp. 1051–1054.

[13] A. Plaia and M. Ruggieri, *Air quality indices: A review*, Rev. Environ. Sci. Biotechnol. 10 (2011), pp. 165–179.

[14] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, 2nd ed., Springer-Verlag, New York, 2005.

[15] M. Ruggieri, F. Di Salvo, A. Plaia, G. Agró, *EOFs for gap filling in multivariate air quality data: A FDA approach*, 19th International Conference on Computational Statistics, Paris, France, August 22–27, 2010.

[16] M.D. Ruiz-Medina, *New challenges in spatial and spatiotemporal functional statistics for high-dimensional data*, Spatial Stat. 1 (2012), pp. 82–91.

Downloaded by [University of Connecticut] at 10:12 28 October 2014