Noname manuscript No.
(will be inserted by the editor)

Functional zoning for Air Quality

Rosaria Ignaccolo · Stefania Ghigo · Stefano Bande

Received: date / Accepted: date

Abstract This paper presents a land classification in zones featured by different criticality levels of atmospheric pollution, considering pollutant time series as functional data: we call this proposal "Functional Zoning". We aim to meet a request of European laws that impose to distinguish zones needing further actions from those needing only maintenance according to air quality status. To carry out zoning for Piemonte (northern Italy), we consider the hourly concentration fields of the main pollutants produced by a deterministic air quality model, and we preprocess them by assimilating observations gathered by monitoring networks. In order to consider administrative units which policy makers refer to, we present three different alternatives to upscale data to municipality scale. Then, to aggregate by pollutant, we evaluate two strategies

Work partially supported by Regione Piemonte.

R. Ignaccolo
Dipartimento di Statistica e Matematica Applicata,
Università di Torino,
Corso Unione Sovietica, 218/bis, 10134 Torino (TO), Italy
Tel.: +39-011-6705758
Fax: +30-011-6705783
E-mail: ignaccolo@econ.unito.it
R. Ignaccolo is also affiliated to Statistics Initiative, Collegio Carlo Alberto, Italy

S. Ghigo
Dipartimento di Statistica e Matematica Applicata,
Università di Torino,
Corso Unione Sovietica, 218/bis, 10134 Torino (TO), Italy
Tel.: +39-011-6705758
Fax: +30-011-6705783
E-mail: ghigo@econ.unito.it

S. Bande
Dipartimento Tematico Sistemi Previsionali, Qualità dell'aria,
ARPA Piemonte,
Via Pio VII 9, 10135 Torino (TO), Italy
email: s.bande@arpa.piemonte.it

to summarize time series: air quality index assessment, and use of the Multivariate Functional Principal Component Analysis (MFPCA), respectively. Therefore, we partition municipalities clustering air quality time series and MPFCA scores, and finally we illustrate a comparison study of the different strategies' results.

Keywords Functional Data · B-spline · Cluster Analysis · PAM (Partitioning Around Medoids) · Atmospheric Pollution

1 Introduction

European and Italian directives (Dir. 96/62/EC and D. Lgs. 351/99 art. 6) establish that Italian regions have to identify different land zones in connection to air quality status with the aim to plan suitable actions. In fact, in order to improve or preserve air quality conditions, policy makers define recovery, action or maintenance plans for the different areas.
To enforce these directives, it is necessary to find a classification strategy that takes jointly into account several critical pollutants. Regional or provincial environmental agencies deal with this issue to provide support to policy makers, and generally employ classical methodologies, as decision trees or cluster analysis. Environmental agencies consider pollutant concentration statistical summaries, limit value exceedances, and variables featuring land and municipalities to assign a municipality to a zone. Examples of such variables include pollutant emission density, inhabitant density, orography, and meteorological variables.

In literature, to our knowledge, the land classification problem is not approached in the air quality context, whereas it is discussed in several other applied sciences, sometimes with a different terminology. For instance, Haining (2003) calls "regionalization" a particular classification where spatial units are aggregated according to spatial contiguity or adjacency constraints. These constraints impose that the spatial units within a region must be geographically connected. In this context, Duque et al (2007) make an interesting review of different techniques, grouping them with respect to the strategy applied for satisfying the spatial contiguity constraints, directly or indirectly expressed. Also, Jacquez (2008) and Aldstadt (2010) describe different approaches for spatial clustering, whereas Guo (2008) and Duque et al (2010) employ them in geographical sciences for the 2004 US presidential election data and georeferenced socio-demographic data in Accra, Ghana, respectively.
In the agricultural framework, Wang et al (2010) carry out a regionalization of crop cultivation in China employing first a classical cluster analysis, and then modifying the resulting groups in order to preserve spatial contiguity, through a Geographical Information System (GIS) software.

Without considering spatial contiguity, Wang and Ni (2008) develop a Projection Pursuit Dynamic Clustering applied on a multivariate dataset, in order to classify China land according to water resources. Also, in the water quality context Robertson and Saad (2003), and Robertson et al (2005) present a regional classification scheme, meaning the partition of large areas into zones with similar environmental natural (not anthropic) factors affecting water quality: they develop an approach called SPARTA (SPAtial Regression-Tree Analysis) and its land-use-adjusted version. Note that this approach is similar to the strategies adopted by environmental local agencies to obtain zones in the air quality context. Analogously here, we do not take into account contiguity or adjacency in an explicit way, since our goal is not a spatial aggregation.

In this paper we propose a functional approach to partition a land in zones characterized by different criticality levels of atmospheric pollution, that we call "Functional Zoning". Specifically, we consider air pollutant time series provided by a deterministic air quality model on a regular grid, and preprocessed by assimilating observations, as functional data (Ramsay and Silverman 2005). We then classify them by using functional clustering, where the Partitioning Around Medoids (PAM) algorithm is embedded (as in Ignaccolo et al 2008) in place of the k-means one, as proposed by Abraham et al (2003). Thus the allocation to a specific zone preserves information about pollution temporal patterns and does not take into account any other information. By considering air pollutant time series exclusively, we do not include any other information about covariates, while this is necessary in model-based approaches (e.g. James and Sugar 2003, and Fruhwirth-Schnatter and Kaufmann 2008). Although our approach is proposed in the air quality context, it could be adopted whenever there are time series - that can be treated as functional data - observed in a spatial domain to be zoned.
Our proposal can be applied to time series observed on grid points, but since municipalities are the reference territorial administrative units for undertaking actions, we suggest three different algorithms to upscale data from a regular grid to the municipality scale. On the other hand, to support policy decisors it is not sufficient to apply our proposal separately per pollutant. Therefore, in order to have a multi-pollutant zoning, we propose at first a functional clustering of air quality index time series, calculated by aggregating by pollutant. Then we consider multivariate functional principal component analysis as an alternative technique to summarize the main pollutant time series (PAM algorithm is employed to cluster functional principal component scores). The different analysis strategies (three upscaling algorithms times two pollutants' aggregations) are applied on air quality data of Piemonte (Northern Italy) in 2005. Then, these strategies are compared by looking at differences between cluster labels in the zoning outcomes in order to suggest a final choice.

The paper is organized as follows. A description of the available dataset and of the convenient preprocessing is provided in Section 2. Section 3 reviews the functional clustering approach and explains the two pollutant aggregation strategies, while Section 4 presents the three different alternatives to upscale data to a municipality scale. In Section 5, we illustrate the results of the proposed techniques concerning the main critical air pollutants in Piemonte, both separately per pollutant, and considering them jointly. Section 6 includes comparison among zoning outcomes and discussion.

2 Data description and preprocessing

The information we are going to use consists of observed data gathered from irregularly spaced sites of monitoring networks in Piemonte and the surrounding area, and "simulated" data, given as output of a deterministic model on a regular thick grid. Note that the European law allows to use both observed and simulated data for the air quality assessment.

The simulated time series are output of a three-dimensional deterministic modeling system (C.T.M. F.A.R.M., Chemistry Transport Model Flexible Air Quality Regional Model) implemented by the environmental agency *ARPA Piemonte* (Bande et al 2007a). This model chain is capable to process meteorological observations and to simulate air pollutant transport, transformation and diffusion; moreover, it is effectively used in order to support the environmental department of Piemonte region in the annual evaluation of air quality status. Concentration fields of the main atmospheric pollutants (such as CO, SO_2, O_3, PM_{10} and NO_2), emission fields, principal meteorological and turbulence fields are produced on a hourly basis over a regular grid that has an horizontal resolution of 4 km and covers Piemonte, neighbor Italian regions and foreign countries (see Fig. 1). The total covered surface is 220×284 km^2. In this paper, we analyze concentration values of CO, SO_2, PM_{10} and NO_2 in the year 2005 and we do not include O_3 time series, even if O_3 is one of the main critical pollutants during the summer (Cocchi and Trivisano 2002, Bodnar et al 2008). Indeed, ozone is a by-product arising from the reaction between nitrogen oxides and volatile organic compounds, and thus it seems appropriate to consider in our analysis its precursors.

The observed data are provided by the regional environmental agencies of Piemonte and its neighbor regions. There are 26 PM_{10} monitoring sites (Fig. 1) that record daily data, while CO, SO_2 and NO_2 are monitored by respectively 40, 12 and 61 instruments providing hourly data. Given the limited number of monitoring sites and their absence in a few areas, a "standard" kriging would not be able to provide a good pollutant prediction on the whole Piemonte region. The kriging variance would be larger in the areas uncovered by the monitoring network, and everywhere when the monitoring sites are only 12. Therefore, we take advantage of the availability of the data simulated on Piemonte and its neighborhood.

When comparing the output of the deterministic model with the observed data, it turns out that pollutant concentration is sometimes underevaluated or overevaluated (see Fig. 2 for an example). In fact, PM_{10} concentration levels are clearly underevaluated in the winter semester (from October to March), especially for monitoring sites out of Torino metropolitan area. The NO_2 output is characterized by underestimation of concentration peaks due to adverse weather conditions, while CO and SO_2 concentrations are overestimated. However, simulated levels of CO and SO_2 are below the law thresholds, as well as their observed levels. In order to improve pollutant model output, we preprocess the simulated data through Kriging with External Drift (KED, Wackernagel 2003), by employing the *geoR* package in R (Development Core Team

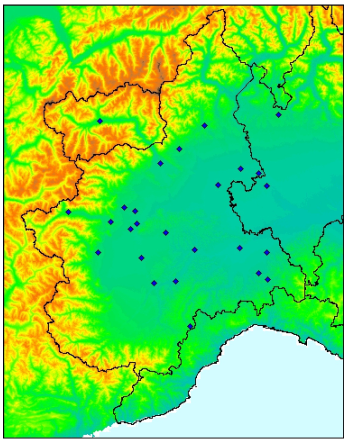


Fig. 1 Spatial domain of the deterministic model with orography, and overimposed PM_{10} monitoring sites

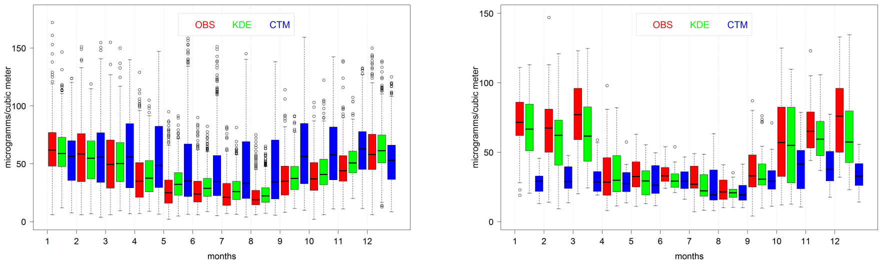
2010): concentration fields are "corrected" by assimilating observed data, before being used in the clustering procedure.

KED implementation follows Van de Kassteele et al (2009) that combine observations and deterministic dispersion model data of atmospheric NO_x concentration. Specifically, the kriging is applied on the observed data and the external drift is constituted by the deterministic model output. Exploratory analysis of pollutant data observed at the different sites showed skewed distributions and standard deviation correlated with mean. Therefore, in order to stabilize the variances and make the distributions approximately normal, a Box-Cox transformation (Box and Cox 1964) is applied to the original data separately per pollutant, and transformed observations are interpreted as realizations of a Gaussian process $Y(s)$ at spatial location s, in the domain S. This spatial process has the following structure

$$Y(s) = \mu(s) + w(s) + \varepsilon(s),$$

where $\mu(s) = \xi_0 + X(s)\xi_1$, $\forall s \in S$, is the spatial deterministic component (trend or drift), X is the deterministic variable that represents the model output and $\xi = \{\xi_0, \xi_1\}'$ is the parameter vector. The process $w(s)$ is stationary Gaussian with zero mean, sill σ^2, and spatial correlation function $\rho(\cdot)$ with range ϕ. Finally, $\varepsilon \sim N(0, \tau^2)$ is the measurement error field, where τ^2 is the nugget. For the definition of sill, range and nugget see Cressie (1993).
We consider an exponential spatial correlation function for each pollutant. In order to fit the model, we estimate first the parameters of the Box-Cox transformation, and then the parameters of the model, in both cases by the likelihood method, separately for each pollutant and time point. Then, concentration fields are corrected at each time.

The model output preprocessing performs well: the improvement is clear when we look simultaneously at simulated, observed and corrected distribution boxplots, especially during the summer for NO_2 and the winter for PM_{10},



(a) NO_2 in Orbassano

(b) PM_{10} in Borgaro

Fig. 2 Boxplots by month of observed (red), simulated (blue) and corrected by KED (green) distributions for two monitoring sites of NO_2 and PM_{10}. The time scale is month, from January (1) to December (12)

for all the sites. Figure 2 shows two examples in two different monitoring sites for NO_2 and PM_{10} distributions. Moreover, we carried out a cross-validation analysis in order to evaluate the KED performance, and it shows that kriging results are satisfactory. The results of this analysis are not reported here because they are beyond the goal of this paper (for further details see Ghigo 2009 - unpublished thesis).

3 Multi-pollutant functional zoning

Functional clustering methods (see Abraham et al 2003 and Ignaccolo et al 2008) could be directly employed on each pollutant: indeed considering grid points' pollutant time series as functional data and clustering estimated coefficients, we obtain several zoning outcomes of the same land. However, decisors need to jointly consider different pollutants in order to get a multi-pollutant zoning: for this reason we propose two strategies for pollutant aggregation.
We suggest to aggregate time series of different pollutants by using an air quality index - called BC index - and, alternatively, to summarize them by means of Functional Principal Component Analysis (FPCA). While Functional Cluster Analysis (FCA) is applied on air quality index time series, the PAM algorithm is employed to cluster functional principal component scores.
In order to make pollutants comparable, we standardize and temporally aggregate them with respect to the aggregation functions reported in Table 1, provided by EU directives (1999/30/EC and 2000/69/EC).

3.1 Functional clustering on BC index

Through synthetic environmental indices it is possible to reduce multivariate information to an univariate one. It is well-known that local agencies use indices as summary values for measuring pollutant effect on human health and

Table 1 Information about analyzed pollutants provided by European directives (sl_p = standard limit value for the pollutant p)

Pollutant	Temporal aggregation function	sl_p	Unit of measure
CO	Daily maximum of 8 hours moving average	10	mg/m^3
NO_2	Daily maximum	200	$\mu g/m^3$
PM_{10}	Daily mean	50	$\mu g/m^3$
SO_2	Daily mean	125	$\mu g/m^3$

on natural environment.

The family of air quality indices proposed by Bruno and Cocchi (2002) allows to aggregate different pollutants in space and time, using the standard limit values (sl_p in Table 1) to standardize and make data dimensionless; we refer to this family as BC indices. For our purposes, we aggregate in time and by type of pollutant in order to obtain daily time series of the air quality index for each site. Unlike classical air quality indices, the BC index is not featured by weights, since dividing each pollutant value by its standard limit we already take into account its higher or smaller criticality.

For a fixed site, let x_{pdh} be the elementary measurement where $p = 1, \ldots, P$ indexes the pollutants, $d = 1, \ldots, 365$ the days and $h = 1, \ldots, 24$ the hours. First, we temporally aggregate data according to the aggregation functions in Table 1, obtaining a daily value $x_{pd} = f(x_{pdh})$. Then, we choose to aggregate over pollutants by the maximum function, in order to keep information about critical cases, obtaining a BC index time series for the fixed site. Therefore the index is defined as

$$I_d = \max_p \left(\frac{x_{pd}}{sl_p} \right) \qquad d = 1, \ldots, 365, \tag{1}$$

where, as said before, sl_p represents the pollutant standard limit value.

For all the sites, we consider these time series $\{I_d\}_{d=1,\ldots,365}$ as observed functional data and assume the existence of a continuous function underlying the data (Ramsay and Silverman 2005): each curve is summarized by a vector of B-spline coefficients in $\mathbb{R}^{g+K+1}$, where K is the knot number and g is the degree of B-splines. Then, we cluster the spline coefficients and obtain groups of sites, through a functional cluster analysis where PAM algorithm is embedded (Kaufman and Rousseeuw 2005). The choice falls on this non-hierarchical algorithm because it provides an object - the socalled "medoid" - representing the cluster, which in this case is a fitted curve showing the temporal evolution of the air quality index at a certain site. Moreover, this algorithm suggests the suitable clusters' number to split the objects and provides useful cluster features by the socalled "silhouette plot". Indeed, the silhouette plot shows a "silhouette width" for each object that represents a belonging measure of the object to a cluster and could have a negative value if misclassification happens. Thus, by averaging silhouette widths over a cluster, we can compare the quality of different groups; similarly, by averaging silhouette widths over all objects we have an index $\bar{s}_k$, changing with the number of groups k. PAM suggests to choose k such that $\bar{s}_k$ reaches the maximum (for further details see Kaufman and Rousseeuw 2005).

3.2 Clustering of Multivariate FPCA scores

In order to aggregate pollutants, as an alternative technique in the functional context, we explore the Functional Principal Component Analysis in its multivariate version (MFPCA, Ramsay and Silverman 2005, p. 167, and Henderson 2006), that clearly takes into account pollutant interactions. We review here some MFPCA theory and set the notation.

Let $G_{pi}(t)$, $p = 1, \ldots, P$ and $i = 1, \ldots, n$, be a functional object for the p-th pollutant and i-th site, assumed to be a smooth function underlying the pi-th time series, that is the datum y_{pij}, $j = 1, \ldots, m$, is associated with the curve value $G_{pi}(t_j)$ by the model

$$y_{pij} = G_{pi}(t_j) + \epsilon_{pij}$$

where ϵ_{pij} are independent random errors.

In the multivariate functional context, the weight functions subject to the orthonormality constraints ($||\alpha_r||^2 = \int_{\mathcal{T}} \alpha_r(t)^2 dt = 1$ and $\int_{\mathcal{T}} \alpha_r(t)\alpha_s(t)dt = 0$ for $r \neq s$) that maximize the sample variance of the principal component scores are P - dimensional vectors $\alpha_r(t) = \{\alpha_r^1(t), \ldots, \alpha_r^P(t)\}$, where r indicates the r-th component. The functions $\alpha_r^p(t)$ are solutions of the eigenequation

$$\sum_{p^*=1}^{P} \int_{\mathcal{T}} c_{pp^*}(t, t^\star)\alpha^{p^*}(t^\star)dt^\star = \lambda\alpha^p(t) \qquad p = 1, \ldots, P,$$

where $c_{pp^*}(t, t^\star)$ denotes the covariance function of $G_{pi}(t)$ when $p = p^*$, while c_{pp^*} is the cross-covariance function when $p \neq p^*$. Note that the eigenfunction components $\alpha_r^p(t)$ are defined over the same time range $\mathcal{T}$ of the functional data.

Therefore, the r-th principal component scores are

$$z_{ir} = \sum_{p=1}^{P} \int_{\mathcal{T}} \alpha_r^p(t) G_{pi}(t) dt, \qquad i = 1, \ldots, n. \tag{2}$$

MFPCA is employed on standardized (divided by sl_p) time series of the analyzed pollutants and its implementation takes place estimating G_{pi} by means of B-splines using the *fda* package in R environment (*pca.fd* with centered curves). After we have obtained the principal component scores, we choose the suitable number of functional principal components (FPCs). Sites are then grouped by clustering the first few FPC scores by PAM algorithm. Note that now the medoids are not functions since the temporal component is integrated out in Eq. (2).

The choice of B-splines is classic for nonperiodic data, but recently Kayano and Konishi (2009) have proposed to use Gaussian basis expansion instead of

B-splines since they deal with unbalanced data (time series observed at possibly different time points). Our pollutant data are already balanced because, as we said before, through the aggregation functions (provided by air quality EU directives) we obtain daily values.

4 Upscaling

In order to enforce air quality legislation and therefore perform actions to improve air quality, policy makers have to refer to administrative units, that are municipalities. So, it is necessary to upscale the "corrected" air quality time series from the grid point scale to the municipality one and, for this purpose, we consider three alternative procedures.

The aggregation at municipality scale could be realized solving the so-called "Change Of Support Problem" (COSP), that is "concerned with inference about the values of the variable at points or blocks different from those at which it has been observed" (Gelfand 2010, chap. 29, p. 522). With this meaning, a solution of COSP presumes to fit a model or, at least, to implement an universal block kriging. Moreover, with a more complex model it is possible to combine data at different scales provided by numerical models and monitoring networks, realizing a data fusion and solving COSP at the same time. However, the goal of this paper is land classification and we propose solutions to carry out data fusion and upscaling at a very reasonable computational cost and at a moderate complexity such that practitioners could be encouraged to implement them.

To retrieve a block value we can consider the integral over the area that provides a block average (BA), that is

$$Z(B) = \frac{1}{|B|}\int_B Z(s)ds, \tag{3}$$

where $B \subseteq S \subseteq \mathbb{R}^2$, Z is a spatial stochastic process defined by a random variable family $\{Z(s) : s \in S\}$ and s a generic spatial location in the domain S. Since our spatial domain is discretized through grid points, after we have identified the c_i points belonging to a municipality, Equation (3) becomes

$$Cc_i = \sum_{l=1}^{c_i} \gamma_{il} z_{il},$$

where Cc_i represents the mean pollutant concentration in a municipality, z_{il} is the value of the stochastic process and γ_{il} is the weight for the l-th point of the i-th municipality (it will be referred to as Municipality Block Average or *Munic. BA*). More specifically, weights are calculated as the ratio between municipality area belonging to the l-th cell (a point in the grid represents a cell) and the total municipality area.

Moreover, as alternative weights, we propose to consider the built-up surface percentages (referred to as *Built BA*). Indeed, the built-up surface percentage

is an important indicator of the anthropic activity which could be associated with more pollution in a municipality, whereas a wide country area could not contribute at all. Now, weights are the ratio between municipality built-up area belonging to the l-th cell and the total municipality built-up area.
Further, we propose to upscale taking the 90-th percentile (referred to as *90th perc.*) of the values observed in the set of cells belonging to the considered municipality. This is a measure of extreme cases in a municipality, and it is taken in a precautionary perspective. Although from this point of view the maximum could seem the natural statistic to consider, it could be an outlier, either too extreme or too rare. Therefore, the choice of the 90-th percentile guarantees a better robustness.

5 Piemonte zoning

We employ the methodologies introduced above on Piemonte preprocessed pollutants' datasets (1763 grid point time series per pollutant), and then on upscaled datasets (1206 municipality time series per pollutant). In the following, results for the main critical pollutants are presented, first separately and then jointly. In order to better highlight the criticality of the resulting zones, all the maps are characterized by a color gradation (as traffic lights) changing with cluster mean values (green, orange, red and purple) that provides a sort of group ranking.
For land classification we analyze NO_2, PM_{10}, CO and SO_2. Since previous exploratory analysis show that NO_2 and PM_{10} are more critical than CO and SO_2, in Subsection 5.1 only NO_2 and PM_{10} results are discussed. A first analysis on single pollutants is illustrated in Bande et al (2007b), even if in this previous work simulated data are not preprocessed and we classify all the grid points of the model output (the whole rectangular region of Figure 1). The following subsections show results of multi-pollutant zoning carried out through the proposed strategies: while Subsection 5.2 introduces the results on regular grid time series, Subsection 5.3 presents findings on the municipality ones.

5.1 Zoning for single pollutants

In order to provide a land classification featured by each pollutant concentration, we carry out a Functional Cluster Analysis on single pollutant time series. This step of our analysis is important for two reasons: on one hand decisors are interested in the critical pollutants and consequently in zoning outcomes based on them (despite the need to base decisions on global air quality status); on the other hand, the resulting maps for single pollutants can be helpful in the interpretation of the following multi-pollutant zoning outcomes, looking at possible similarities between single- and multi-pollutant maps.

First, we produce a land classification on 1763 Piemonte grid points for single pollutants and then we refer to municipalities using the three upscaling approaches described in Section 4. Moreover, to select the “best” number k of clusters to partition all the curves, we look at the average silhouette width over all objects $\bar{s}_k$ and generally we choose the k that yields the highest $\bar{s}_k$; sometimes we make an exception, opportunely described when necessary.

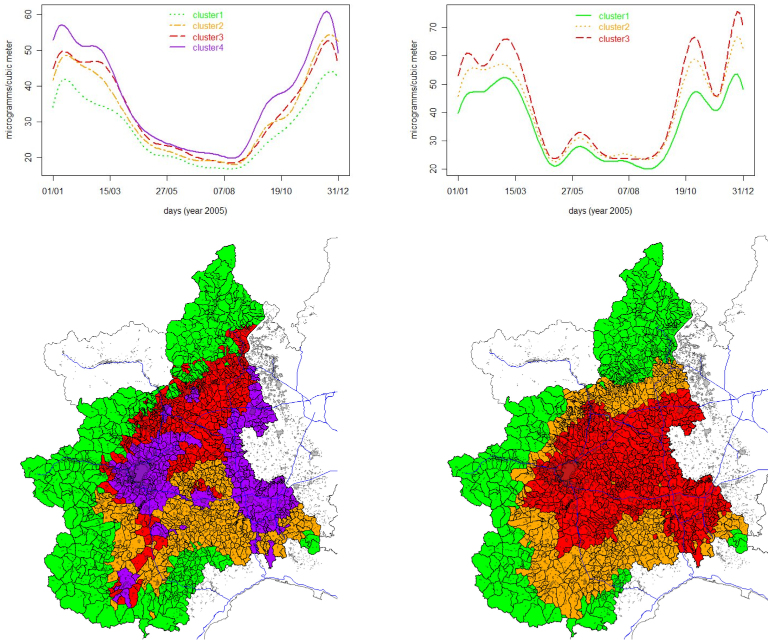
Table 2 Average silhouette width over the cluster ($\bar{s}_c$), number of grid points or municipalities (n_c) and *color* featuring each cluster in the grid point and municipality classifications (in the three upscaling cases)

	NO_2				**PM_{10}**		
cluster	C_1	C_2	C_3	C_4	C_1	C_2	C_3
color	green	orange	red	purple	green	orange	red
Grid points							
$\bar{s}_c$	0.61	0.14	0.26	0.10	0.54	0.24	0.08
n_c	635	347	537	244	805	460	498
Munic. BA							
$\bar{s}_c$	0.52	0.21	0.17	0.12	0.45	0.24	0.13
n_c	387	325	321	173	328	365	513
Built BA							
$\bar{s}_c$	0.53	0.11	0.18	0.14	0.44	0.24	0.12
n_c	333	428	300	145	342	355	509
90th perc.							
$\bar{s}_c$	0.47	0.18	0.20	0.05	0.49	0.26	0.07
n_c	377	304	309	216	311	393	502

In Table 2 we show results for each classification obtained for NO_2 and PM_{10}, while Figure 3 focuses on the upscaling case *90th perc.*.
NO_2 cluster structure is quite similar when zoning results concern the grid points and when they are related to the municipality time series obtained after the three upscaling techniques. Indeed, these outcomes are characterized by four clusters stretching in according to roadway networks and main cities (see Fig. 3(a)). Even if $k = 4$ corresponds to the third $\bar{s}_k$ suggested by PAM, our choice falls upon it because it makes easier to distinguish emission sources. These are concentrated in biggest conurbations and industrial areas of Piemonte (Torino, Alessandria, Novara and their suburbs) that belong to the cluster 4 (C_4). C_3 includes zones around highways, main connection roads and northern plain and piedmont areas, while the cluster 2 is formed by remaining plain and piedmont areas. Mountains belong to C_1. The medoids in Figure 3(a) highlight the big variation that exists, especially in the winter months, among sites belonging to different groups: it reaches 30 $\mu g/m^3$ for some time points.

As for PM_{10}, $k = 3$ gives the best $\bar{s}_k$ for the grid points case, and this k is also kept for the other cases because it corresponds to the three directive plans (recovery, action, and maintenance) and it makes easier to compare all the resulting PM_{10} maps. Figure 3(b) shows the three clusters in the *90th perc.* upscaling case, but the other zoning outcomes are very similar. Indeed they reflect morphology: the flat country belongs to C_3, piedmont areas to C_2 and mountains to C_1.

Values of $\bar{s}_c$ (that represent the average silhouette widths of objects belonging to a cluster) in Table 2 highlight that grid points and municipalities



(a) NO_2 (b) PM_{10}

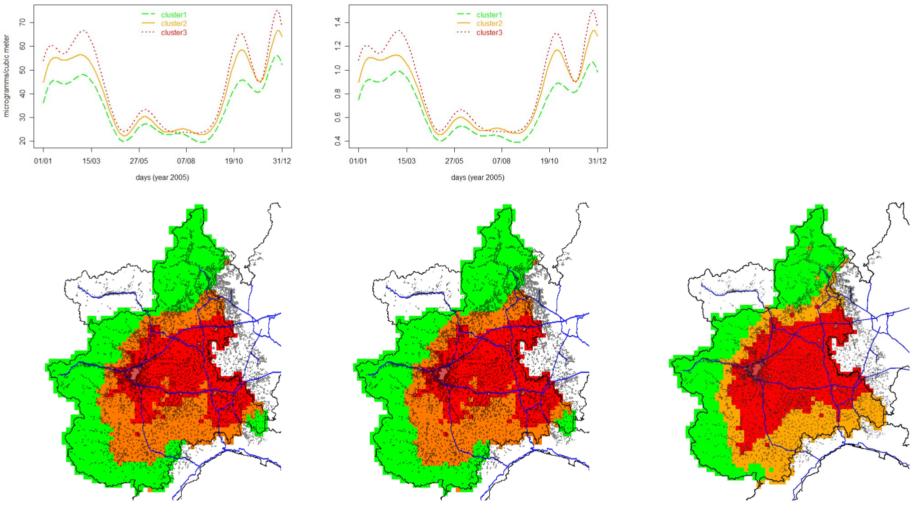
Fig. 3 Medoids (top) and municipality zoning outcome maps (bottom) for NO_2 and PM_{10} in the *90th perc.* case. The background layers show the regional boundaries, the highway network and the municipality built-up area

of the NO_2 cluster 1, C_{1,NO_2}, - and also $C_{1,PM_{10}}$ - lie well within their cluster although some of them are spatially very far away. C_{4,NO_2} and $C_{3,PM_{10}}$ have a low $\bar{s}_c$ instead: since they group cities having similar features of industrialization and morphology but not comparable for road conditions and population, low $\bar{s}_c$ values look proper. Moreover, Table 2 provides the numerousness n_c of each cluster. Looking at n_c in the three different upscaling cases, we can observe that some units "migrate" from one group to another one, even if cluster structures are quite similar.

5.2 Multi-pollutant zoning on grid points

As we already mentioned, to look at the regional air quality status and to provide one zoning outcome for all pollutants, two time series summary methods are proposed, through BC air quality index and MFPCA (Section 3). First of all, we implement these two strategies on 1763 grid point time series in order to have a benchmark to evaluate the persistence of the resulting clusters when we move at municipality scale (in Section 5.3).

During the PAM step we choose the second best value, that is $k = 3$ for both the strategies: the "best" is obtained with $k = 2$ that provides a meaningless partition from the policy point of view, because it splits Torino and the metropolitan area from the rest of Piemonte, in a too simplistic way.



(a) PM_{10} (b) BC index (c) MFPCA

Fig. 4 PM_{10}, BC and MFPCA zoning outcome maps (bottom) and corresponding medoids (top) of Piemonte grid points. The background layers show the regional boundaries, the highway network and the municipality built-up area

As illustrated in Section 3.1, we classify functional BC indices by using the PAM algorithm. This multi-pollutant zoning outcome is characterized by $\bar{s}_3$ equal to 0.33. Figure 4(b) shows that the metropolitan area and almost the whole Po valley belong to C_3, piedmont regions to C_2 and mountains to C_1 again. Moreover Table 3 provides average silhouette width and numerousness for every cluster. Note that these values are very similar to those obtained when we zone PM_{10} grid points (Tab. 2): the same number of grid points belongs to cluster 3 in both the classifications, and a few grid points move from $C_{1,BC}$ to $C_{2,PM_{10}}$, keeping about the same $\bar{s}_c$. Also PM_{10} and BC index maps and medoids are very similar, as it can be seen in Figure 4. Therefore, we can gather that the PM_{10} drives the construction of BC index defined in Eq. (1).

When we summarize pollutant time series through MFPCA(see Section 3.2) we cluster the functional principal component scores by PAM. Figure 4(c) displays the zoning outcome and the right side of Table 3 shows numerousness and average silhouette width for every cluster, whereas the overall average silhouette width $\bar{s}_3$ is 0.46. In this analysis we take into account the first two principal components that explain 90.89% of the total variability. The

Table 3 Average silhouette width over the cluster ($\bar{s}_c$), number of grid points (n_c) and *color* featuring each cluster

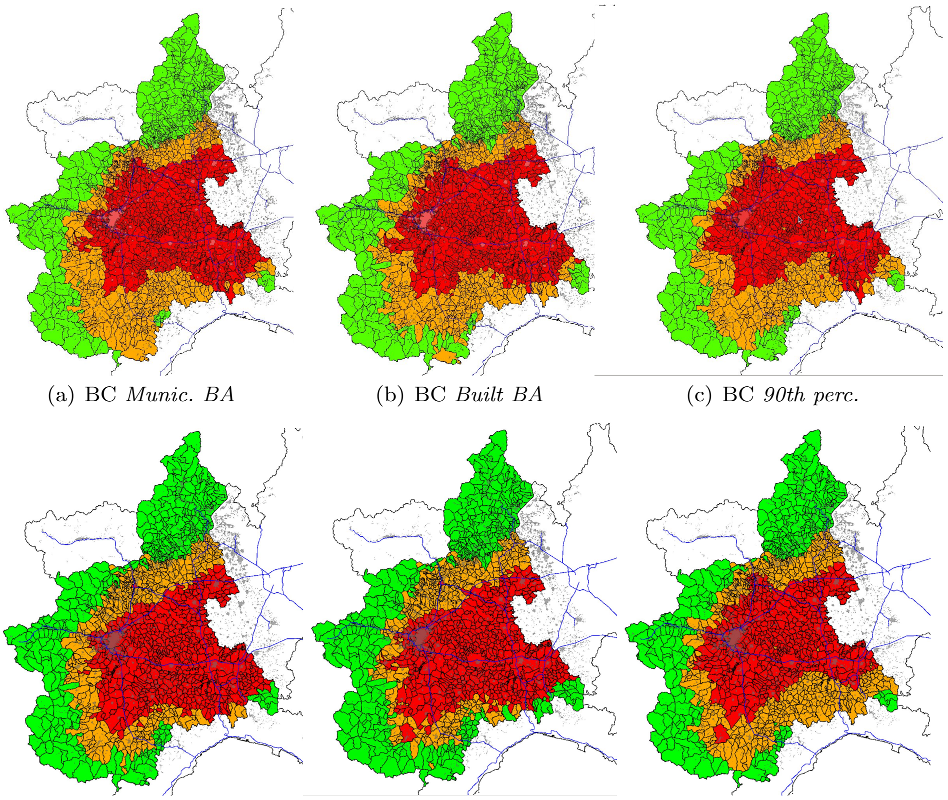
	BC index			**MFPCA scores**		
cluster	C_1	C_2	C_3	C_1	C_2	C_3
color	green	orange	red	green	orange	red
$\bar{s}_c$	0.55	0.23	0.09	0.29	0.69	0.32
n_c	784	481	498	473	710	580

most critical cluster C_3 includes main connection roads and industrial areas, grouping the biggest towns. C_2 is characterized by piedmont areas, and C_1 by mountains. Since $\sum_{p=1}^{P} ||\alpha_r^p||^2 = 1$ by construction, $||\alpha_r^p||^2$ provides the proportion of the variability in the r-th component ascribable to variation in the p-th pollutant curves. For the first principal component, 91.97% of variation is attributable to PM_{10} curves, that is $||\alpha_1^{PM_{10}}||^2 = 0.9197$; 7.69% is associated with NO_2 while the other pollutants play no substantial role (SO_2: 0.11%, CO: 0.23%). Also in the second principal component, the higher variability proportion concerns PM_{10} ($||\alpha_2^{PM_{10}}||^2 = 0.9808$); the other contributions are 0.59%, 1.12% and 0.21% for NO_2, SO_2 and CO, respectively. Then, it is possible to say that fine particulate matter explains almost all the variability of the first two principal components, or rather they are essentially affected by PM_{10}.
When comparing the two outcomes, Table 3 highlights that $C_{2,MFPCA}$ is featured by n_c and $\bar{s}_c$ bigger than $C_{2,BC}$ ones, whereas conversely n_c and $\bar{s}_c$ of $C_{1,MFPCA}$ are smaller than those of $C_{1,BC}$, as displayed by the different orange and green zones in Fig. 4(b) and (c). As for C_3, the value of $\bar{s}_{C_{3,MFPCA}}$ is bigger than $\bar{s}_{C_{3,BC}}$ meaning that $C_{3,MFPCA}$ is more homogeneous than $C_{3,BC}$.

5.3 Multi-pollutant zoning on municipalities

The three upscaling techniques illustrated in Section 4 are now implemented to refer to municipalities. We decided to keep $k = 3$ that makes easier to compare new zoning results with the previous ones, and corresponds to the number of plans that currently policy makers have to define. The resulting maps are shown in Figure 5. Generally, the most critical group C_3 lengthens over the main road network, including main conurbations (Torino, Alessandria and Novara) and their industrialized suburbs. The intermediate C_2 is formed by piedmont municipalities. Mountain municipalities are grouped in the less critical C_1 that is featured by the highest average silhouette width for every obtained classification (Table 4). Overall, the municipalities' partition resembles the grid points' one (see Fig. 4(b)-(c) and Fig. 5), meaning that the three upscaling techniques do not shuffle the zoning outcomes. Moreover, at first glance, it seems that in all the municipalities' maps PM_{10} overbears the other pollutants as it can be seen by comparing Figure 5(a)-(f) with Figure 3(b).

Specifically, with regard to BC index zoning on municipalities, we have to emphasize that Figure 6(a) is associated with Figure 5(b) and it is a helpful instrument to look at the temporal evolution of the BC index in the three

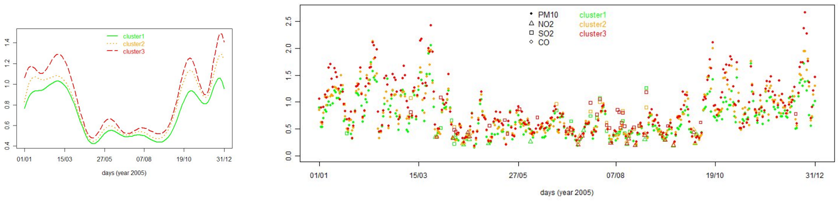


(a) BC *Munic. BA* (b) BC *Built BA* (c) BC *90th perc.*

(d) MFPCA *Munic. BA* (e) MFPCA *Built BA* (f) MFPCA *90th perc.*

Fig. 5 BC index (top) and MFPCA (bottom) multi-pollutant zoning outcome maps in the three upscaling cases. The background layers show the regional boundaries, the highway network and the municipality built-up area

zones. The left side of Table 4 shows that a small number of municipalities migrates from a group to another one when we change the upscaling algorithm, supporting the fact that maps are quite similar. Also the overall average silhouette widths ($\bar{s}_3 = 0.252; 0.245; 0.241$ for *Munic. BA*, *Built BA* and *90th perc.* respectively) and the medoids are comparable (the interested reader can see Fig. 3.15 in Ghigo 2009 - unpublished thesis). Moreover, *90th perc.* map (see Fig. 5(c)) seems to better reflect Piemonte land use: for instance the orange zone in southern Piemonte adheres to main road connections, while the red one in the south-east is due to the closeness to Genova metropolis. *Munic. BA* outcome seems to present some misclassifications instead, for instance most southern mountain municipalities belong to C_2 instead of C_1.
To understand which pollutants have a part in the BC index construction, we look at the medoid time series composition, as we are used in the classical PCA. Thus in Figure 6(b) for the *Built BA* case, medoid values at each time instant are drawn with a different shape depending on which pollutant gives



(a) Smoothed medoids

(b) Medoids time series

Fig. 6 Medoids and their time series for BC index zoning in *Built BA* case. The point shape in time series changes with the pollutant giving the maximum in Eq. (1)

Table 4 Average silhouette width over the cluster ($\bar{s}_c$), number of municipalities (n_c) and *color* featuring each cluster in the three upscaling cases

	BC index			**MFPCA scores**		
cluster *color*	C_1 green	C_2 orange	C_3 red	C_1 green	C_2 orange	C_3 red
			Munic. BA			
$\bar{s}_c$	0.48	0.23	0.13	0.55	0.32	0.20
n_c	297	389	520	336	355	515
			Built BA			
$\bar{s}_c$	0.45	0.23	0.13	0.50	0.34	0.18
n_c	317	361	528	378	331	497
			90th perc.			
$\bar{s}_c$	0.50	0.25	0.07	0.59	0.29	0.24
n_c	309	403	494	303	447	456

the maximum in Eq. (1). We can observe that the predominant shape is the PM_{10} one, although SO_2 prevails on a few summer days. So, as we supposed before, PM_{10} seems to be the driving pollutant. For the *Munic. BA* case the analogous plot looks very similar to Figure 6(b), whereas in the *90th perc.* one the SO_2 shape never appears.

When clustering MFPCA scores, we take into account the first three functional principal components: indeed the percentages of explained variability are 91.5%, 91.3% and 92.5%, as Table 5 shows. The summary values obtained by applying the PAM on the MPCA scores are reported in the right side of Table 4, and the overall average silhouette widths are $\bar{s}_3 = 0.331; 0.328; 0.344$ for *Munic. BA*, *Built BA* and *90th perc.*, respectively. Even if the maps in Fig. 5(d)-(f) seem quite similar to the BC index ones, there are maybe some misclassifications. For instance, *Built BA* and *90th perc.* place Cuneo (the red isolated unit in the south-west) in the red zone, while a preliminary data analysis suggests that it is not so critical; indeed the negative silhouette widths warn about a possible misclassification. Moreover, based on knowledge of land use we would expect that some municipalities were in the critical C_3. Instead, as for *Munic. BA* and *Built BA*, some municipalities in the north-west (near Valle D'Aosta, see Fig. 5(d) and (e)) featured by well-travelled connection roads belong to C_2 instead of C_3. Analogously, for *90th perc.* in Figure 5(f), there are some municipalities close to the industrialized Genova in the south-east belonging to C_2.

Table 5 Explained variabilities and variability percentages ($||\alpha_r^p||^2$) for each pollutant of the first three FPCs, in the three upscaling cases

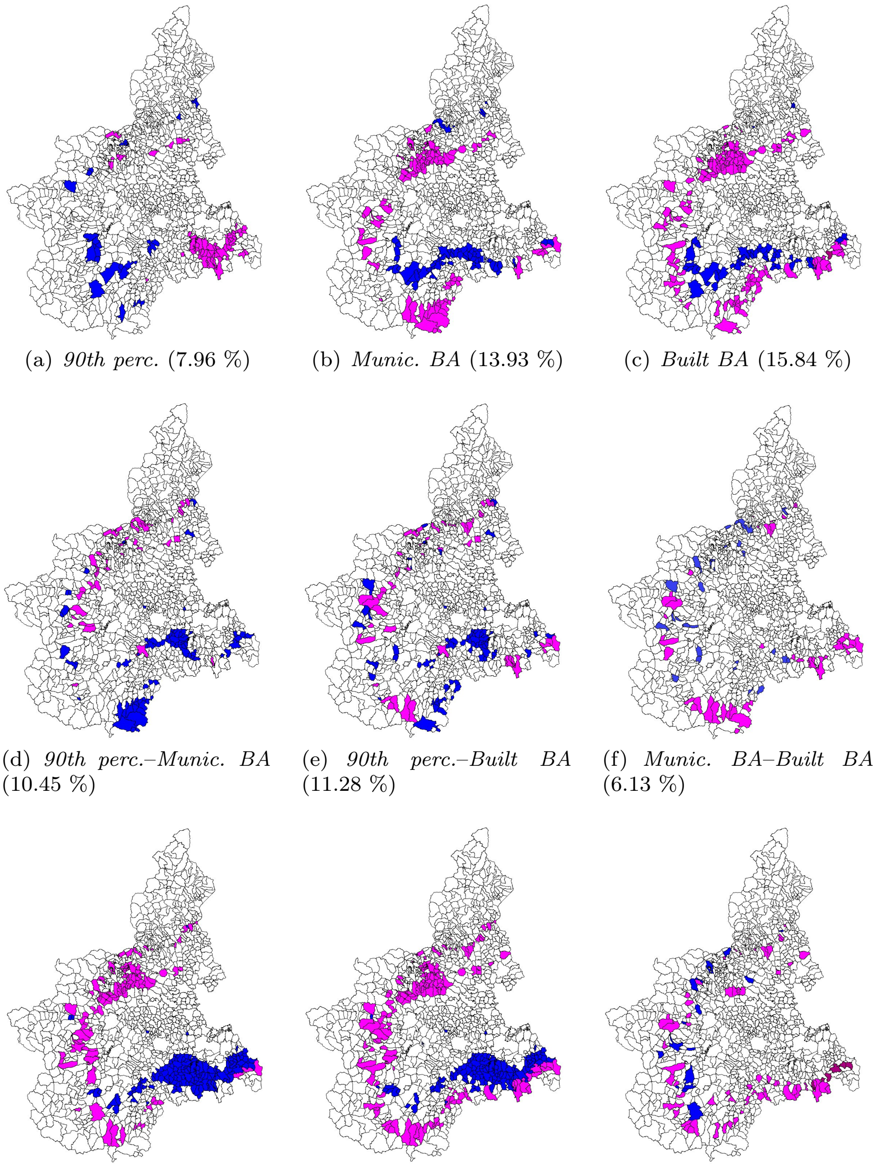
Explained Variability		$||\alpha_r^p||^2$			
		PM$_{10}$	**NO$_2$**	**SO$_2$**	**CO**
		Munic. BA			
65.6	**1st FPC**	0.9095	0.0503	0.0381	0.0021
20.5	**2nd FPC**	0.5740	0.0119	**0.4134**	0.0007
5.4	**3rd FPC**	0.8813	0.0686	0.0491	0.0010
91.5	**Total**				
		Built BA			
66.4	**1st FPC**	0.9127	0.0532	0.0320	0.0021
19.5	**2nd FPC**	0.5968	0.0111	**0.3913**	0.0008
5.4	**3rd FPC**	0.8738	0.0789	0.0463	0.0010
91.3	**Total**				
		90th perc.			
75.8	**1st FPC**	0.9116	0.0821	0.0030	0.0033
12	**2nd FPC**	0.9690	0.0012	0.0280	0.0018
4.7	**3rd FPC**	0.7616	**0.2363**	0.0006	0.0015
92.5	**Total**				

Now, we can look at the proportion of variability of the components attributable to each pollutant curves (remember that $\sum_{p=1}^{P} ||\alpha_r||^2 = 1$). Table 5 shows that FPCs are composed almost completely by PM$_{10}$. Only for the second FPC of the two *BA* upscaling cases, some variability is explained by SO$_2$ ($||\alpha_2^{SO_2}|| = 41.34\%$ and 39.13% respectively), while in the third FPC of the *90th perc.* case 23.63% of variability is associated to NO$_2$. Thus, PM$_{10}$ looks like the driving pollutant of the functional principal components, as well as of the BC index.

6 Discussion

In order to support policy makers in enforcing European laws, environmental local agencies have to provide a land classification in relationship with air quality status. In this paper, we propose a functional approach to zone, applied on corrected air pollutant time series that are provided by a deterministic model. The preprocessing of simulated data through a KED procedure is carried out separately for each pollutant.

There are two convenient “constraints” to consider. First: recovery, action or maintenance plans have to be determined on the basis of the air quality status, implying the need for an aggregation of critical pollutants. Second: plans have to be applied to administrative units where responsibilities are well defined, so that any proposal is feasible in practice if zoning results concern municipalities. To satisfy the first need, we suggest an aggregation of several pollutants in the BC air quality index (that is the input of the functional clustering) and, alternatively, a summary of data by MFPCA (employing PAM algorithm on



(a) *90th perc.* (7.96 %)

(b) *Munic. BA* (13.93 %)

(c) *Built BA* (15.84 %)

(d) *90th perc.–Munic. BA* (10.45 %)

(e) *90th perc.–Built BA* (11.28 %)

(f) *Munic. BA–Built BA* (6.13 %)

(g) *90th perc.–Munic. BA* (21.23 %)

(h) *90th perc.–Built BA* (23.38 %)

(i) *Munic. BA–Built BA* (7.30 %)

Fig. 7 Maps of the differences between cluster labels in zoning outcomes: BC versus MFPCA (top), BC versus BC (center), MFPCA versus MFPCA (bottom) in the three upscaling cases. The municipality color reflects the belonging, or not, to the same cluster in the compared classifications: blue corresponds to "-1" in Table 6, white to "0", magenta to "+1" and purple to "+2". The number in round brackets represents the percentage of municipalities with different cluster labels in the two compared outcomes

Table 6 Frequency distributions of the differences between cluster labels in two compared zoning outcomes (changing pollutants' aggregation and upscaling algorithm)

-1	**0**	**1**	**-1**	**0**	**1**	**-1**	**0**	**1**	**2**
BC-MFPCA									
90th perc.			***Munic. BA***			***Built BA***			
32	1110	64	62	1038	106	50	1015	140	1
BC–BC									
90th perc.–Munic. BA			***90th perc.–Built BA***			***Munic. BA–Built BA***			
82	1080	44	81	1070	55	31	1132	43	0
MFPCA–MFPCA									
90th perc.–Munic. BA			***90th perc.–Built BA***			***Munic. BA–Built BA***			
141	950	115	124	924	158	17	1118	65	6

functional principal component scores). Then, in order to refer to municipalities, we propose to upscale data from a regular grid to municipality scale by means of three different methods: i) averaging the field values weighted by areas over the cells belonging to a certain municipality; ii) averaging using as weight the built-up percentage for every cell; iii) taking the 90-th percentile over the cell values in a municipality.

Crossing the proposed two pollutant aggregations and three upscaling solutions, we obtain six classifications of municipalities in Piemonte for the year 2005, as shown through six maps in Figure 5. Obviously, a choice is necessary. To compare the zoning outcomes and realize a sensitivity analysis, we map the differences between cluster labels (1, 2 and 3) in Figure 7(a)-(i) and we present their frequency distributions in Table 6. The municipalities that belong to the same group in the two compared outcomes - white in the maps of Figure 7 - vary from 924 to 1132 (see the frequencies of "0"). The "migrating" municipalities, from a cluster to another one, are colored: blue corresponds to "-1", magenta to "+1" and purple to "+2". Since the clusters are labeled from 1 to 3 according to a criticality ranking (based on the average air quality index), we can read a difference "-1" as a shift towards a more critical zone, and interpret a large frequency of "-1" as indication of a more preventive zoning outcome: when we compare two zoning outcomes a large number of "-1" suggests that the second zoning strategy is more preventive. Indeed, overall the strategies, the frequency of "-1" varies from 17 to 141. Instead a difference equal to "+1" indicates a shift of a municipality towards a less critical zone and the frequency of "+1" changes from 43 to 158. Finally, there are some purple municipalities (7 in total) that move from cluster C_3 in the first zoning outcome to C_1 in the second one.

The colored municipalities in the difference maps in Figure 7 do not appear randomly distributed, and a randomness test based on joint counts (Cliff 1970) confirms that. If the colored municipalities were randomly distributed on the difference maps, the two compared zoning outcomes could be considered the same. Instead, they are grouped and located between mountain and piedmont zones, as well as piedmont and plain zones, that roughly coincide with boundaries between two different clusters.

When comparing the two pollutant aggregation strategies (BC versus MFPCA) with the same upscaling algorithm, through the difference maps in Figure 7(a)-(c), we can see that *90th perc.* outcomes are very similar (differences are mostly focused in the south-east of the region). In fact, the percentage of municipalities with different cluster labels is only 7.96% (corresponding to 96 out of 1206). This percentage is at most 15.84% in the case of *Built BA* upscaling.

In order to carry out a sensitivity analysis with respect to the upscaling procedure, we plot the difference maps in Figure 7(d)-(f) for the BC index and in Figure 7(g)-(i) for the MFPCA. In both the cases, the smallest percentage of migrating municipalities is observed in the comparison of the two weighted block average methods (6.13% for BC and 7.30% for MFPCA). Instead, the highest percentages of migrating municipalities occur when we compare *90th perc.* with a block average method in the case of MFPCA pollutant aggregation: 21.23% and 23.38% in Figure 7(g) and (h). Therefore MFPCA approach turns out to be less robust than the BC index with respect to the upscaling method.

In addition to the above better robustness to the upscaling procedures, we have other reasons to prefer FCA on BC index rather than PAM on MFPCA scores. First of all, the construction of the BC index time series makes them easily interpretable for policy makers who can also look at the medoids to have an idea of the temporal evolution of the BC index in the corresponding zone. Then, since the frequency of "-1" is smaller than the frequency of "+1" in the three upscaling cases at the first row of Table 6, the BC zoning outcome appears more preventive than the MFPCA one. Finally, a further "merit" of the FCA on BC procedure is that it does not take into account either spatial correlation or spatial contiguity among municipalities (spatial correlation is implicitly considered in MFPCA instead) so that it does not force neighbors to be in the same zone and provides a zoning outcome based exclusively on the similarity of the air quality status.

As for the choice of an upscaling method, we can see from Table 6 that zoning outcomes are more cautionary with the *Munic. BA* procedure: the frequency of "-1" is greater than that of "+1" when we compare *Munic. BA* with *90th perc.*, and vice versa in the comparison *Munic. BA* versus *Built BA* in both the pollutant aggregation methods (see the second and the third row of Table 6). However, discussing the choice from an interpretive point of view with policy makers, it seems that BC *90th perc.* approach has more consistent results with their previous knowledge about Piemonte land and air quality status (as already said for Fig. 5(c)).

Very recently Kayano et al (2010) have proposed a multivariate FCA that allows to consider all pollutants without synthesizing them. Their proposal is an alternative to the use of FCA on the BC index that we could take into account in the future. Nevertheless, first of all an assessment of the computational cost of that methodology would be necessary, whereas we have a low cost for the proposed FCA on BC index. Moreover we stress again that the use of PAM provides information about the air quality temporal evolution in

the zones and warning about a possible misclassification of a municipality. Another issue that could be worthy of attention is the data assimilation and upscaling by means of fitting a hierarchical model, as suggested in Gelfand (2010, chap. 29, p. 536). However, the related computational costs are surely higher than those necessary to implement the KED preprocessing and an upscaling algorithm.
Therefore, our proposal provides environmental agencies and policy makers with a useful and easy-to-read instrument to prepare recovery, action, and maintenance plans for the different zones at a very reasonable computational cost. Moreover, the outline of the comparison study could be interesting for more general purposes, for example to compare municipalities' classifications year by year.

Acknowledgements The authors wish to thank Giorgio Arduino, Michela Cameletti, and Luigi Ippoliti for useful discussion and comments.

References

Abraham C, Cornillon PA, Matzner-Løber E, Molinari N (2003) Unsupervised curve clustering using B-splines. Scandinavian Journal of Statistics 30:581–595

Aldstadt J (2010) Spatial clustering. In: Fischer MM, A G (eds) Handbook of applied spatial analysis, Springer, pp 279–300

Bande S, Clemente M, De Maria R, Muraro M, Picollo ME, Arduino G, Calori G, Finardi S, Radice P, Silibello C, Brusasca G (2007a) The modelling system supporting Piemonte region yearly air quality assessment. UAQ 2007, Cipro, March 2007

Bande S, Ghigo S, Ignaccolo R (2007b) Functional zoning with respect to air quality assessment. In: Proceedings ISI Conference 2007

Bodnar O, Cameletti M, Fassò A, Schmid W (2008) Comparing air quality in Italy, Germany and Poland using BC indexes. Atmospheric Environment 42:8412–8421

Box GEP, Cox DR (1964) An analysis of transformations. Journal of the Royal Statistical Society: Series B 26:211–246

Bruno F, Cocchi D (2002) A unified strategy for building simple air quality indexes. Environmetrics 13:243–261

Cliff AD (1970) Computing the spatial correspondence between geographical patterns. Transactions, Institute of British Geographers 50:143–154

Cocchi D, Trivisano C (2002) Ozone. In: El-Sharaawi A, Piegorsch W (eds) Encyclopedia of environmetrics, Wiley, New York, pp 1518–1523

Cressie N (1993) Statistics for Spatial Data. John Wiley and Sons

Development Core Team (2010) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Duque JC, Ramos R, Suriñach J (2007) Supervised regionalization methods: a survey. International Regional Science Review 30 (3):195–220

Duque JC, Aldstadt J, Velasquez E, Franco JL, Betancourt A (2010) A computationally efficient method for delineating irregularly shaped spatial clusters. Journal of Geographical Systems DOI 10.1007/s10109-010-0137-1

Fruhwirth-Schnatter S, Kaufmann S (2008) Model-based clustering of multiple time series. Journal of Business and Economic Statistics 26 (1):78–89

Gelfand AE (2010) Misaligned spatial data: The change of support problem. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P (eds) Handbook of Spatial Statistics, Chapman & Hall/CRC, pp 517–539

Guo D (2008) Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science 22 (7):801–823

Haining R (2003) Spatial Data Analysis - Theory and Practice. Cambrige University Press

Henderson B (2006) Exploring between site differences in water quality trends: a functional data analysis approach. Environmetrics 17:65–80

Ignaccolo R, Ghigo S, Giovenali E (2008) Analysis of air quality monitoring networks by functional clustering. Environmetrics 19:672–686

Jacquez GM (2008) Spatial cluster analysis. In: Fotheringham S, Wilson J (eds) The Handbook of Geographic Information Science, Blackwell Publishing, pp 395–416

James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98:397–408

Kaufman L, Rousseeuw PJ (2005) Finding Groups in Data. An introduction to cluster analysis. Wiley

Kayano M, Konishi S (2009) Functional principal component analysis via regularized gaussian basis expansions and its application to unbalanced data. Journal of Statistical Planning and Inference 139:2388–2398

Kayano M, Dozono K, Konishi S (2010) Functional cluster analysis via orthonormalized gaussian basis expansion and its application. Journal of Classification 27:211–230

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer

Robertson DM, Saad DA (2003) Environmental water-quality zones for streams: A regional classification scheme. Environmental Management 31 (5):581–602

Robertson DM, Saad DA, Heisey DM (2005) A regional classification scheme for estimating reference water quality in streams using land-use-adjusted spatial regression-tree analysis. Environmental Management 37 (2):209–229

Van de Kassteele J, Stein A, Dekkers ALM, Velders GJM (2009) External drift kriging of NO_x concentrations with dispersion model output in a reduced air quality monitoring network. Environmental and Ecological Statistics 16(3):321–339

Wackernagel H (2003) Multivariate geostatistics: an introduction with applications. Springer

Wang H, Zhang X, Li S, Song X (2010) Spatial clustering and outlier analysis for the regionalization of maize cultivation in China. WSEAS Transactions on Information Sciences and Applications 7 (6):860–869

Wang SJ, Ni CJ (2008) Application of projection pursuit dynamic cluster model in regional partition of water resources in China. Water Resources Manage 22:1421–1429