Theor Appl Climatol (2013) 114:39–54
DOI 10.1007/s00704-012-0801-0

ORIGINAL PAPER

Functional clustering for Italian climate zones identification

E. Di Giuseppe · G. Jona Lasinio · S. Esposito · M. Pasqui

Received: 28 December 2011 / Accepted: 14 November 2012 / Published online: 28 December 2012

Abstract This work presents a functional clustering procedure applied to meteorological time series. Our proposal combines time series interpolation with smoothing penalized B-spline and the partitioning around medoids clustering algorithm. Our final goal is to obtain homogeneous climate zones of Italy. We compare this approach to standard methods based on a combination of principal component analysis and Cluster Analysis (CA) and we discuss it in relation to other functional clustering approaches based on Fourier analysis and CA. We show that a functional approach is simpler than the standard methods from a methodological and interpretability point of view. Indeed, it becomes natural to find a clear connection between mathematical results and physical variability mechanisms. We discuss how the choice of the basis expansion (splines, Fourier) affects the analysis and propose some comments on their use. The basis for classification is formed by monthly values of temperature and precipitation recorded during the period 1971–2000 over 95 and 94 Italian monitoring stations, respectively. An assessment based on climatic patterns is presented to prove the consistency of the clustering and a comparison of results obtained with different methods is used to judge the functional data approach.

Keywords Climate zones · Functional clustering · Physical connection

E. Di Giuseppe (✉) · S. Esposito
Research Unit for Climatology and Meteorology Applied to Agriculture (CRA-CMA), Consiglio per la ricerca e la sperimentazione in agricoltura, Via del Caravita 7/a, 00186 Rome, Italy
e-mail: edmondo.digiuseppe@entecra.it; edmondo.digiuseppe@uniroma1.it

E. Di Giuseppe · G. Jona Lasinio
Department of Statistics (Uniroma1-Dss), University of Rome La Sapienza, P.le A. Moro 5, 00185 Rome, Italy

M. Pasqui
Institute of Biometeorology (CNR-Ibimet), National Research Council, Via dei Taurini, 19, 00185, Rome, Italy

1 Introduction

A key issue in meteorological field analysis is played by the study of their spatiotemporal variability. There exists a structural variability which describes the nature of a phenomenon both to intra-annual (*seasonality*) and long-term variability (*climate trend*) and it is relevant to be able to analyze them over homogeneous climate areas. A set of different methods are used for climate zone determination, typically a combination of principal component analysis (PCA) and cluster analysis (CA). Guidelines on the use of PCA in meteorology and climatology have been set in the work of Preisendorfer and Mobley (1988). A theoretical and applied framework of the principal component analyses of climate-related fields is given in Chapter 13 of Von Storch and Zwiers (1999). The spatial domain PCA (S-mode) is a reduction of the information related to the temporal patterns of the locations (Ehrendorfer 1987). Thus, each component generates a mapping of mixed physical features. On the other hand, the temporal domain PCA (T-mode), by reducing the information seen from the time series point of view, attempts to describe climate regime (Richman 1986). Finally, the R-mode approach points at local similarity in mean and variances of meteorological fields across a fixed time by means of CA (Fovell and Fovell 1993). The main drawback of PCA-based techniques is that the reduced space they return as output does not have an immediate connection with the physical one.

In this work, a combination of functional data analysis (FDA) and partitioning around medoids (PAM) clustering technique is applied in Italy to monthly surface temperature and precipitation fields in order to delineate local climate zones. FDA is a collection of techniques to model data from dynamic systems in terms of some set of basis functions, which are a linear combination of known functions. FDA consists of converting observations gathered at discrete time into functional data. The choice of the basis to implement this conversion is crucial. The functional data approach is typically used in genetic (Kim et al. 2008) and pollution's diffusion analysis (Ignaccolo et al. 2008) and only very recently in climate studies (Laguardia 2011). Kim et al. (2008) used functional data approach for modeling the time-dependent expression value of genes in the genome of yeast and they found that the features of those genes are properly modeled by a 3-order Fourier series approximation. Ignaccolo et al. (2008) fit the functional data to pollutant concentrations time series using B-spline system of basis, with a fixed number of knots. Then, they produce a zonal index of pollutant's concentration in Northern Italy based on a clustering of estimated coefficients. In Laguardia (2011), a Fourier basis expansion is adopted to model a very large amount of precipitation data (2,043 rain gauges). The clustering is performed using a *k*-means clustering algorithm. Our approach differs from his, first of all, for the choice of the clustering algorithm and, secondly, as in our setting, penalized B-splines are preferred to Fourier basis. Our choices are discussed below in details. We also note that in our work, a smaller amount of data than in Laguardia is considered, nevertheless returning very coherent results.

Temperature and precipitation time series can be considered as realizations of continuous processes recorded in discrete time. Thus, they are converted into functional data through the estimation of spline coefficients and the latter used for the final classification as each time series is representative of location climate variability. Here, a penalized B-spline basis system is adopted to map observations gathered at discrete time into functional data. Our proposal is named *Bsplines30 model* and reproduces data intra-annual variability by means of B-spline basis system over a 30-year period (1971–2000). A fixed number of knots guarantees a comparability of responses from the 95 and 94 time series, which constitute the dataset for the analysis of temperature and precipitation, respectively. On the contrary, a system with a free number of knots would lead each series to be smoothed according to different scale of variability and, de facto, the delineation of homogeneous zones would not be done. Finally, the estimated coefficients are partitioned by PAM classification technique and average silhouette width method is used to determine the number of climate zones (Rousseeuw 1987).

The main advantage of a functional approach to this type of data is dimensional reduction, as the information on monthly temporal pattern given by a large number of observations (time series) is summarized by a small number of coefficients that describe the basis spanning the functions (Ramsay and Silverman 1997). Furthermore, the proposed approach overcomes the problem of connecting the reduced space to the physical one. Indeed, the fitting of B-splines allows to define in a clear way which type of variability is considered.

The paper is organized as follows. The next section introduces the available data. In Section 3, we illustrate functional clustering in general terms and then we move to illustrate our proposal. Section 3.1 is devoted to the presentation of penalized B-spline; in Section 3.2, a description of the *k*-medoids clustering procedure is reported and Section 3.3 details our proposal. In Section 4, we prove and discuss the validity of the method and we compare the final grouping with the same results obtained by means of PCA method in T-mode. Relations of our proposal to Fourier analysis approach are discussed in the same section. Finally, in the last section, some concluding remarks are presented.

2 Data

In this paper, we define "Regime" as the signal obtained by averaging monthly values over the years in each station. The dataset is composed of daily precipitation and daily minimum and maximum temperature data collected by the CRA-CMA Research Unit for Climatology and Meteorology Applied to Agriculture for the period 1971–2000 from 98 Italian stations (Fig. 1). The total number of stations is 98, but only 92 are in common with temperature and precipitation; then, we have 96 stations for temperature and 94 for precipitation. This dataset is composed of climatic time series with a relatively small amount of missing values over the considered time window. This fact in addition to the continuity of these time series is a considerable advantage over other more numerous, in terms of monitoring stations, Italian dataset, such as SIMN (SIMN is affected by a severe missing data problem and, moreover, its collection ends around 1989 when it was dismissed). Among the 98 stations, seven are located above 1,500 meters and it is natural to expect that, due to correlation between temperature and altitude, they may form a cluster. For those stations, especially during winter season, observed values of precipitation might be due to snow events which amount is usually transformed into equivalent precipitation quantity. Nevertheless, this fact does not affect the analysis provided that all the mountains' stations were grouped in a unique cluster. One station time series (Pian Rosà) has been removed as the station is located

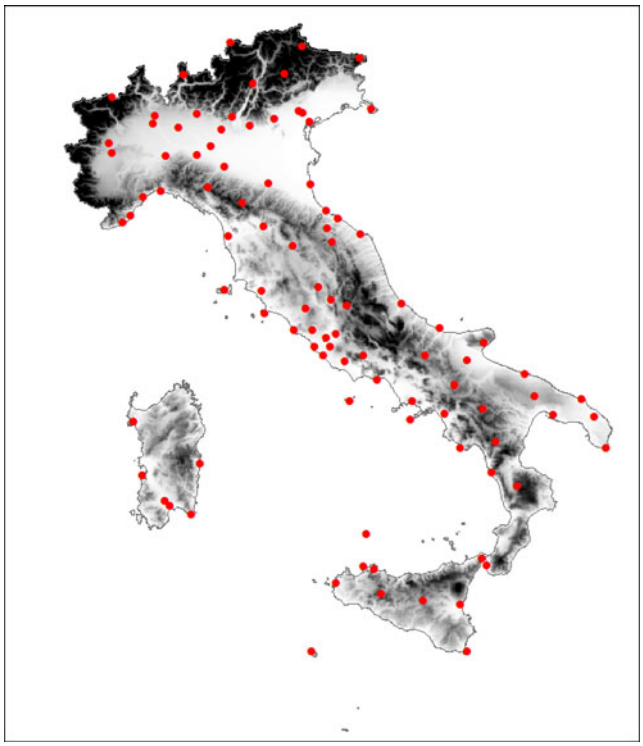


Fig. 1 Location of weather stations

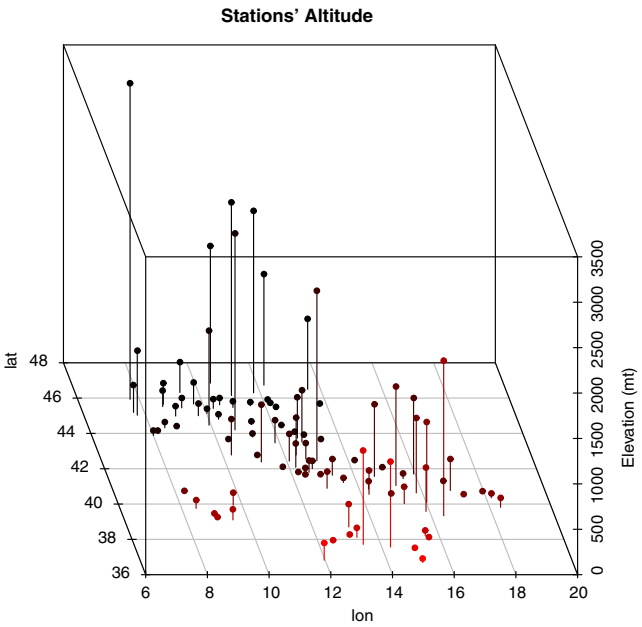


Fig. 2 Altitude of weather stations

at 3,480 m and becomes an outlier with respect to the other stations (see Fig. 2).

Minimum and maximum temperatures were averaged to obtain a rough estimate of daily medium temperatures. Then monthly mean of medium temperature (Tmed-MM) and monthly cumulated precipitation (Prec-MC) were calculated provided that at least 21 daily data in a month were registered. If not, the corresponding monthly value is set to Not Available (NA), i.e., missing value. Besides, a non-parametric test for outlier detection of these monthly values is performed. This test is based on median absolute deviation (MAD) and is suggested in Sprent (1998) as "simple and reasonably robust test." In fact, MAD is itself a robust estimator of the spread of a univariate data series. More specifically, let x_i be the element of a data series with $i = 1, \ldots, n$ and x_{Med} the median of the series; then, MAD is the median of the absolute deviation from the median:

$$\text{MAD} = \text{Median}(|x_i - x_{\text{Med}}|) \tag{1}$$

and x_i is detected as outlier if

$$\frac{|x_i - x_{\text{Med}}|}{\text{MAD}} > M \tag{2}$$

where $M = 5$ following Sprent and Smeeton (2001). They suggest this rule of thumb because of the approximate relation 5MAD = 3Sd, with Sd denoting standard deviation. The cross stations' outliers detected and successively removed by the MAD-based test applied over the period 1971–2000 are zero for Tmed-MM and 150 for Prec-MC. The latter is not significant with respect to the 33,840 overall number of data (360 monthly values × 94 stations). Summary statistics for Tmed-MM and Prec-MM are reported in Table 1 together with the overall number of missing data.

An *imputation* of missing monthly data has been performed accounting for seasonal variability and 3-year climate cycle, since the completeness of the series makes the application of FDA method easier from a computational point of view and it eases the output interpretation. In particular, to estimate spline coefficients from a series, completeness is necessary, but the values of the curve—giving rise to that series—can be observed on an irregular grid. Therefore, if the amount of missing data is small, it is possible to omit NAs and estimate spline coefficients even when time series is not complete. Nevertheless, the completeness of the series is fundamental for establishing the announced connection to physical variability. Briefly, the missing data $\tilde{y}_{ij}$ in year i and month j, are imputed as

$$\begin{aligned}\tilde{y}_{ij} \;=\; & 1/3(\overline{y}_j + [1/2 \cdot (y_{i-1,j} + y_{i+1,j})] \\ & + [1/2 \cdot (y_{i,j-1} + y_{i,j+1})])\end{aligned} \tag{3}$$

where $\overline{y}_j$ is the 30-year average corresponding to the jth month value. Whenever contiguous missing data are found, they are directly imputed with the 30-year average. With the number of the monthly missing values being small, we decided not to adopt a complex statistical model (such as ARIMA or VARMAX) for imputation. We use the above-described procedure that takes into account the general

Table 1 Tmed-MM and Prec-MC summary statistics and percentage of monthly missing data calculated across the overall locations

Variables	Min	1st Quartile	Median	Mean	3rd Quartile	Max	% of missing data
Tmed-MM	−14.2	8.5	13.5	13.7	20.0	30.2	4.36
Prec-MC	0	19.6	46.7	60.1	86.3	393.6	5.36

features of monthly regimes, and it is conservative in terms of variability since we use climatological levels. We experimented with other techniques, such as spline imputation as implemented in the `na.spline()` function in R (package "zoo") and other versions of our approach, observing that the proposed functional clustering approach is robust to imputation of missing values, i.e., stations are classified in the same way regardless of the chosen imputation technique. This result is most likely due to the reduced number of missing values present in the data. Besides, because of extremely high variability of precipitation, a Cox–Box transformation with coefficient $\lambda = 0.5$ has been performed on monthly precipitation data (Box and Cox 1964). This transformation corresponds to a square root of the initial data and determines tighter high-scale data and looser low-scale data. Finally, our dataset is composed by 95 and 94 time series of 360 monthly values of Tmed-MM and square root Prec-MC, respectively, since the removed station of Pian Rosà was originally included only in the temperature dataset. In the following, we mention Prec-MC which always refers to the square root of Prec-MC whereas the levels expressed in millimeters are back-transformed to the original scale.

At the end of this section, we want to draw the attention on a major result of the exploratory data analysis, that is the strong influence of local factors in determining both temperature and precipitation spatial patterns. In fact, the complex topography of the Italian Peninsula along with the strong influence of sea over air masses flow generates a large amount of small-scale atmospheric variability which is able to modulate not only the temperature field but also the precipitation one (Trigo and Coauthors 2006). These local variabilities make the building of climatic homogeneous classification particularly challenging.

3 Methodology: functional clustering

Functional clustering combines the functional representation through a given basis expansion of a time series with a cluster algorithm with the aim of finding observed units homogeneous groups. The choice of a basis implies the type of features of the series that are to be enhanced or hidden in the representation (Ramsay and Silverman 1997) and then become relevant in the classification building. The two most commonly chosen bases are Fourier and B-splines. The first one is mostly adopted when data are assumed to have an important periodic component; the second one is particularly suitable when no periodicity is anticipated in the data or periodicity is affected by some type of changing component. A B-spline smoothing is able to incorporate the shifts in the mean level of the time series caused by a breakpoint into the estimates of the coefficients. This may constitute an advantage especially in the study of climate variables and obviously depending on the scope of the analysis. For instance, an ad hoc analysis can be conducted combining two B-spline systems of basis: the first smoothing involves placing a knot every year in order to model the *trend* component and, in case, the breakpoints and the second smoothing involves placing a knot every 2, 3, or 4 months for modeling the *seasonal* component (see Chapter 7 of Ramsay and Silverman (2002)). In particular, the penalized version of B-splines, which we adopt here, becomes useful when the interest is in representing smooth functions without completely removing local behavior in time, such as changes in the time series level that persist for a limited time (Ramsay and Silverman 1997). Furthermore, this basis allows to capture specific variability patterns with an appropriate choice of knot localization.

The second element of functional clustering is the clustering algorithm. In the literature, k-means algorithm has already been used in application to precipitation data (Laguardia 2011). Here, we prefer a partitioning around medoids algorithm (Kaufman and Rousseeuw 1990). The k-medoids algorithm is a clustering algorithm related to the k-means algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, the k-medoids algorithm chooses data points as centers (medoids or exemplars), making easier to identify group features. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances (Kaufman and Rousseeuw 1990).

In what follows, we report a brief description of our main tools: penalized B-spline basis and partitioning around medoids algorithm.

3.1 Functional data smoothing

FDA transforms discrete data y_j in a functional form using a system of basis. B-spline basis is piecewise polynomials of degree d joined at $k + 1$ fixed points named *knots*. Two

adjacent polynomials are required to have matching $d-1$ (continuous) derivatives. The order of the polynomial B-splines is $d+1$ and the free parameters are $k+d+1$. The degree of smoothing is determined both by the location and the number of knots. Thus, a function $x(t)$ can be represented as a linear combination of k known basis functions ϕ_j:

$$x(t) = \sum_{j=1}^{k+d+1} c_j \phi_j(t). \tag{4}$$

The coefficients of the expansion c_j are determined by minimizing a least squares criterion. In the penalized B-splines basis, a penalization term is added to ensure control over local variability and to reduce outliers' influence on the least squares estimates. The penalization term involves a smoothing parameter λ and a linear differential operator $\mathrm{PEN}(x)$ which is a measure of the function roughness (it is the value of an approximate integral over the x range of the square of the $d-1$ derivative of the curve, which quantifies the total curvature of the function). The penalized least squares criterion adopted for coefficients estimation is

$$\mathrm{PENSSE}_\lambda(x \mid \mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]' \mathbf{W} [\mathbf{y} - x(\mathbf{t})] + \lambda \mathrm{PEN}(x) \tag{5}$$

where $\mathbf{W}$ is a symmetric, positive definite weight matrix. The smoothing parameter λ is chosen by *generalized cross validation* (GCV) criterion:

$$\mathrm{GCV}(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{\mathrm{PENSSE}}{n - df(\lambda)}\right) \tag{6}$$

where df are the degrees of freedom in the smoothing curve and its value depends on the number of knots and the spline degree that will be specified in Section 4. The best choice of λ is associated to the minimum value of $\mathrm{GCV}(\lambda)$. For large values of λ, the curve approaches the standard linear regression. A penalized B-spline smoothing with a small number of coefficients is able to capture the shape of the curve and accommodate for local features. Indeed, by using penalized B-spline, we found that outliers in the data do not affect coefficient estimates. We run our method with and without outliers in the data and the interesting feature is that conclusions are not affected by the presence of outliers. However, in our data, we have only few *anomalies*; then, as a good practice, we suggest to remove outliers identified by MAD. Notice that this identification method finds very extreme values (approximately larger than three times the standard deviation) and no outliers are found for Tmed-MM while 150 values are identified for Prec-MC. Thus, the estimate of the coefficients we use in the clustering method is robust. Simple polynomial regression does not have this kind of robustness, and small changes in the data can dramatically affect the coefficients estimates (Abraham et al. 2003).

In practice, the construction of the "best" penalized B-spline representation proceeds by iterating two steps: (i) fix the number of parameters (knots and polynomial degree) and choose λ by GCV and (ii) compute root-mean-square error (RMSE), then change the number of parameters and go back to (i); repeat this two steps until no more sensible reduction in RMSE is obtained. Finally, the combination of λ and parameters' number that returns the smallest RMSE is chosen. In general, this sequence of steps can be carried on automatically or a data-driven choice of parameters can be performed. In our case study we choose the latter as we want the final clustering to have a physical meaning and, at the same time, we want to minimize the number of estimated parameters (details are given in Section 3.3).

3.2 Partitioning around medoids classification method

k-medoids algorithm is based on the object called *medoid* (most centrally located point in the cluster) instead of the centroid of k-means algorithm (average of objects coordinates in the cluster). This has two advantages: firstly, the medoid is a real object and it is representative of group features; secondly, there is no need to calculate distances at each iteration since the reference is the distance matrix between objects. The steps of k-medoids algorithm can be summarized as follows:

1. Choose randomly k objects of the n data points to be the initial cluster medoids;
2. Assign objects to the cluster with the closest medoid;
3. Recalculate the k medoids of clusters formed at step 2;
4. Repeat steps 2 and 3 until the medoids do not change.

Step 3 is performed by finding the object i which minimizes

$$\sum_{j \in C_i} d(i; j) \tag{7}$$

where C_i is the *cluster* including i and $d(i; j)$ is any measure of dissimilarity (common choices are Euclidean and Manhattan norms) between observations i and j.

Among k-medoids algorithm, the most used and powerful is PAM algorithm proposed by Kaufman and Rousseeuw (1990). This algorithm is characterized by an efficient procedure for determining the set of medoids, which can be described in two phases: the "build" and the "swap." The gain in the algorithm efficiency introduced with PAM is described in Reynolds et al. (1992). In the build phase, the algorithm looks for a good initial set of medoids. Then, in the swap phase, it calculates the loss in the objective function determined by changing medoid. More specifically, consider the effect of removing object i from the set of

medoids and replacing it with object h. The total cost of the change is given by the sum of the cost associated to each object j that moves from other clusters to the new cluster h determined by the change. In particular, there are three cases:

(a) The cost is zero since object j does not move;
(b) Object j is closer to the initial medoid i than any other medoid before the swap; then, the cost associated to the moving is $d(j,h) - d(j,i)$;
(c) j is further from i than from some other medoid; then, the cost of the moving is $d(j,h) - D_j$, where D_j is the distance of object j from the closest medoid (if the closest is h, then the cost is zero).

If the total cost is negative, then the move gives an improvement in the clustering. The whole neighborhood is evaluated in each iteration of the algorithm. Here, Kaufman and Rousseeuw (1990) suggest that calculating the change in cost rather than the total cost at each iteration is less operationally demanding.

Once the medoids have been fixed, clustering quality indexes can be calculated. Let $a(i)$ be the average dissimilarity between i and all objects in cluster C_i and let $d(i;C)$ be the average dissimilarity of i to all objects in C, with $C \neq C_i$. Denote with $b(i)$ the smallest distance $d(i;C)$ found among all clusters $C \neq C_i$; then, C is the neighbor cluster of i. An evaluation of how well the object i is classified in C_i or in the *neighbor cluster* is given by the *silhouette width index*:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}. \tag{8}$$

Observations with a $s(i)$ value close to 1 are very well clustered, a small value of $s(i)$ means that the observation can be assigned to two clusters, and observations with a negative $s(i)$ are misplaced. The number of clusters can be determined by the *average silhouette width*, which is the mean of $s(i)$ over all objects of any possible clustering (Rousseeuw 1987).

3.3 Proposed functional clustering

In Section 3.1, we describe a general functional smoothing for one time series. We now consider the multiple time series framework that is proper to climatological studies. The first point that requires attention is how to apply the protocol of Section 3.1 to all series in order to obtain comparable results. We propose to use the same penalized B-splines for all time series, i.e., we modify, following Ramsay and Silverman (1997), steps (i) and (ii) as follows: (i.a) we fix the same number of knots (or their position) and polynomial degree for all time series and we choose a unique smoothing parameter λ by GCV. First, for each time series, the GCV corresponding to a given value of λ is computed and then the average of these GCV values is associated to the specific λ; (ii.a) we compute RMSE for each time series and then the average RMSE. We repeat (i.a) and (ii.a) until no sensible changes are obtained in the average RMSE. Finally, we choose the combination of λ and parameters' number that returns a meaningful clustering and, simultaneously, a small average RMSE. In fact, a key point in our procedure is the choice of the number of knots and their positions. Sometimes it is necessary to compromise between a small average RMSE value and a set of knots that return a meaningful representation of the time series. For instance, in our study, by placing a knot every 4 months, we capture intra-annual variability with considerable accuracy. With a knot placed every 3 months, we obtain a smaller average RMSE but the series representation becomes more sensitive to outliers and, if no outliers are present, it is identical to the 4 months one with a considerable increase in the number of parameters to be estimated from the data. Remark that the introduction of a large number of parameters not necessarily helps the understanding of climatic features as not only information is thus added but also noise (variability).

Once the representation of the time series is obtained, the coefficients of the functional smoothing become input of the clustering algorithm with the aim of obtaining climate zones delineation. Here, we use the k-medoids algorithm PAM as implemented in the R `cluster` library of the R Development Core Team (2011) and illustrated above (Section 3.2) The number of clusters is chosen by average silhouette and climatological considerations. In other words, if the largest silhouette value is given by a very small number of clusters, say 2, that does not have a climatological meaningful interpretation, we look for the second best or the third best and so on. Besides, the choice of the proper number of clusters is done taking into account also the information associated to the PAM algorithm, as isolation, diameter of clusters, separation, and silhouette width of each group.

In the present study, we are going to call our procedure *Bspline30* as here we eventually adopt a penalized B-spline basis with a knot placed every 4 months over a 30-year period (1971–2000), which is a period commonly used as climatic normals (WMO 1989). We fix the two knots corresponding to the edges of the smoothing interval respectively on January 1971 and on December 2000 whereas the position of the interior knots, the degree of the polynomial, and the smoothing parameter are determined as illustrated in Section 3.1 and above, using `fda` library (Ramsay et al. 2011), implemented in R Development Core Team (2011). The clustering is performed using PAM implemented in the R library `cluster`. Details of the results are given in the next section.

4 Results

This paper is motivated by the need of finding a segmentation procedure of the available time series leading to homogeneous classes. Most of the analyses regarding the determination of homogeneous climatic regions are based on the monthly time scale. Then, we adopt the monthly scale using monthly averages. This time space truncation is commonly adopted in order not to include the synoptic and subsynoptic variability signals in the atmosphere variability. The climate framework of the Italian Peninsula is made complex by both the sea mitigation effect on temperature and the presence of Alps in the north as well as Appennini along the latitude extension which affect precipitation distribution. In fact, some studies based on standard clustering techniques classify Italian climate in seven-eight homogeneous subregions (Laguardia 2011; Toreti et al. 2009; Brunetti et al. 2006). On the other hand, the Mennella's basic work in 1972 describes at least 20 climate micro-regions using both observations and physical features (Mennella 1972). From a phenomenological point of view, the main advantage of functional clustering is a clear identification of variability mechanisms whereas standard methods need to find a relation between selected principal components (Pcs) and climate patterns. Recall that, with the S-mode of PCA, we look for the most significant Pcs of the information matrix over the stations; then, we map the elements of the corresponding eigenvectors (*loadings*) which are associated to each station (Ehrendorfer 1987). On the other hand, with the T-mode, we look for the most significant Pcs of information matrix over time, then mapping the *scores* (Richman 1986). In this study, we focus on *intra-annual variability* by placing penalized B-spline knots every 3, 4, and 6 months, which let us to capture intra-annual variation with scale of variability larger or equal than 3 months. The functional smoothing performed in this way preserves the bell-shaped temperature monthly distribution typical of the Italian Peninsula and the largest intra-annual precipitation pick. As an abbreviation, we use the term *4-monthly* (or 3-monthly or 6-monthly) to recall the variability scale and the placement of penalized B-spline knots. Following the approach proposed in Section 3.3 for the functional smoothing, the most interesting models among all those investigated are reported in Table 2 where the average RMSE is reported together with the penalization coefficient (λ), the number of total knots of the B-splines, and the degree of the piecewise polynomials used. Notice that with a knot placed every 3 months, the average RMSE is a little smaller than the one obtained with a knot placed every 4 months, but the total number of parameters to be estimated from the data considerably increases, without any advantage in the subsequent classification (details to support this statement are given in Sections 4.1 and 4.2).

The assessment for determining the proper number of clusters and the corresponding index to evaluate the quality of the chosen clustering are visualized in panels a and panel b of Figs. 3 and 7, for the temperature and precipitation, respectively. Medoids' locations are representative of the climate features of all stations belonging to the corresponding cluster and are enhanced in the classification maps of Figs. 4 and 8. Besides, the functional smoothing of medoids' time series over the period 1971–2000 is represented in Figs. 5 and 9. The maps of classification obtained by Bsplines30 model are reported in Figs. 4a (Tmed-MM) and 8a (Prec-MC). In the comparison procedure, we adopted different ways of summarizing time

Table 2 Tmed-MM and Sqrt Prec-MC model selection for functional data transformation with penalized B-splines piecewise polynomials degree, number of knots, penalty coefficient (lambda), and averaged across stations RMSE

Tmed-MM functional model	Degree	Knots	Lambda	RMSE (°C)
Bsplines30 6-monthly	3	60	0.06	1.97
Bsplines30 6-monthly	5	60	1	1.89
Bsplines30 4-monthly	3	90	0.16	1.58
Bsplines30 4-monthly	5	90	3.98	1.44
Bsplines30 3-monthly	3	120	0.25	1.35
Bsplines30 3-monthly	5	120	0.63	1.35
Sqrt Prec-MC functional model	**Degree**	**Knots**	**Lambda**	**RMSE (mm)**
Bsplines30 6-monthly	3	60	3.98	8.78
Bsplines30 6-monthly	5	60	15.85	8.70
Bsplines30 4-monthly	3	90	6.31	8.44
Bsplines30 4-monthly	5	90	63.1	8.39
Bsplines30 3-monthly	3	120	3.98	7.79
Bsplines30 3-monthly	5	120	63.1	8.29

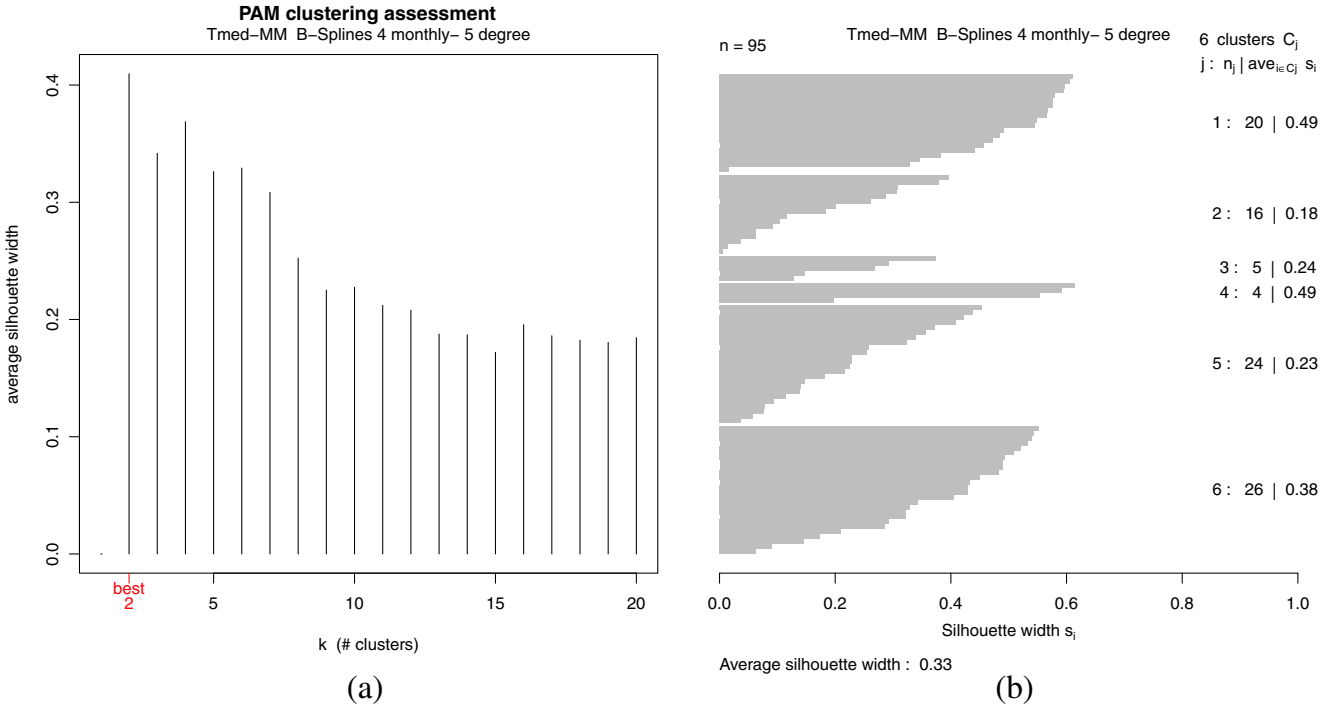


Fig. 3 Cluster algorithm assessment of Tmed-MM for 4-monthly variability functional data: **a** Silhouette average width for determining the number of clusters and **b** silhouette width index for each group and for each unit included in the correspondent six-group clustering

series features: PCA in T-mode and Fourier basis functional smoothing. The latter includes, following Laguardia (2011), 12- and 6-monthly harmonics that should be enough to capture monthly regimes. The final classifications have always been obtained using PAM as in Section 3.3. Classification maps of temperature and precipitation clusters obtained trough PCA standard method are reported in Figs. 4b and 8b; finally, the Fourier functional smoothing clusters are mapped in panel c of the same figures to facilitate comparison.

4.1 Results for Tmed-MM

The chosen model for Tmed-MM is *Bsplines30 4-monthly 5-degree* with 90 fixed knots (a knot placed every 4 months) and five-degree piecewise polynomials which corresponds to functional smoothing of order 6 (see Table 2). This choice produces a good smoothing with an average RMSE value of 1.44 °C although the 1.35 °C minimum value of RMSE is achieved with Bsplines30 3-monthly 3-degree. Nevertheless, as mentioned above, the gain in the smoothing is not enough to justify the increase in the number of parameters to be estimated (from 94 to 122) as the bell shape of monthly temperature distribution typical of the Italian Peninsula is well reproduced by the 4-monthly scale of variability and, moreover, it does not add any useful information for the final classification. In fact, the best number of groups obtained from Bsplines30 3-monthly 3-degree model is 5. This choice is done taking into account all clustering indexes and climate patterns. The maximum value of average silhouette width index corresponds to three-group clustering that has no climatic meaning. The five-group clustering has an average silhouette width of 0.36 with one misplaced unit and returns equivalent results to our 4-monthly 5-degree model except for the northern mountain region. There is a single cluster found by the 3-monthly model, while two clusters (clusters 1 and 2 in our classification mapped in Fig. 4a) are given by the 4-monthly model, the latter being more meaningful from a climatic point of view.

The average silhouette width index reported in panel a of Fig. 3 is our tools to choose the number of clusters, and we report its value from 2 to 20 groups derived from our chosen model. The best value is obtained with two groups, which is not very meaningful from a climatological point of view. As mentioned in Section 3.3, we take into account climate features in the choice of the optimal number of clusters and we select the six-group partition as a good compromise between average silhouette width value and description of climate features. Moreover, the silhouette width values of single groups shown in panel b of Fig. 3 suggest an appropriate classification with no misclassified units (recall that with misclassified units, a negative value of silhouette width index is obtained).

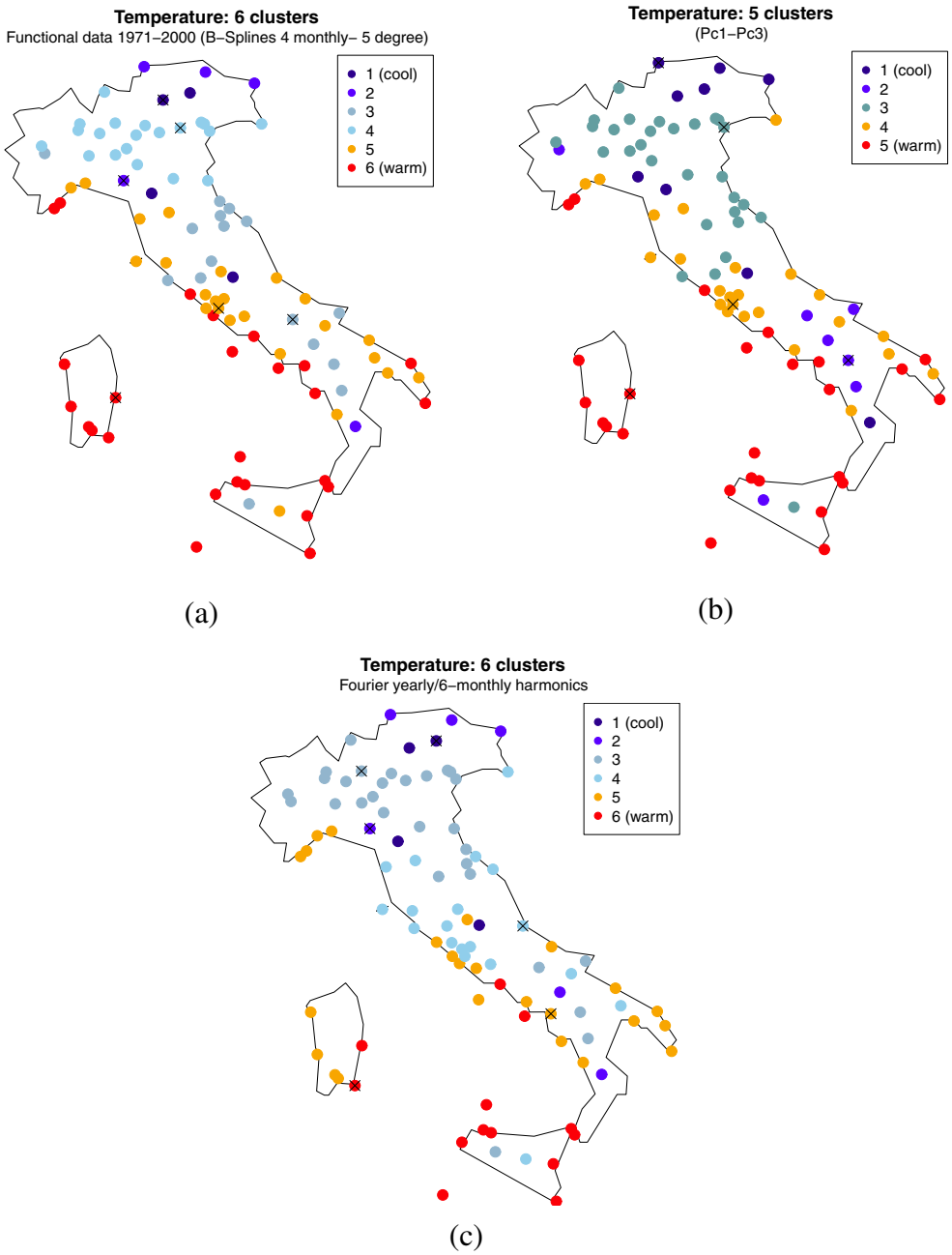


Fig. 4 Cluster maps of Tmed-MM for 4-monthly variability functional data (**a**), PCA T-mode method using three principal components (**b**), and Fourier with five bases of 12- and 6-monthly harmonics. *Crosses* in the maps indicate the location of cluster's medoids

Renaming clusters from colder (1) to warmer (6), we obtain the map in Fig. 4a where the medoids are also indicated. The mapping of Bsplines30 functional clustering highlights the following: there are two coldest clusters of mountain stations in the north (clusters 1 and 2); cluster 3 covers a part of the central area mainly close to the Adriatic Sea and some mountain stations in the south which are not included in the northern mountain stations' clusters because of latitude's mitigation effect; and cluster 4 represents cold stations of the northern area, clusters 5 and 6 correspond to the warm southern region near the Adriatic Sea, and nearly the whole of stations located along the Tyrrhenian Coast and both the islands of Sicily and Sardinia. By visualizing the smoothed time series of medoids in Fig. 5, it is also worth noticing the peculiarity of "hot winters" occurred in mountain regions from 1988 to 1992 (clusters 1 and 2); how cluster 4 differs from cluster 3 for hotter summer temperatures while cluster 5 differs from cluster 6

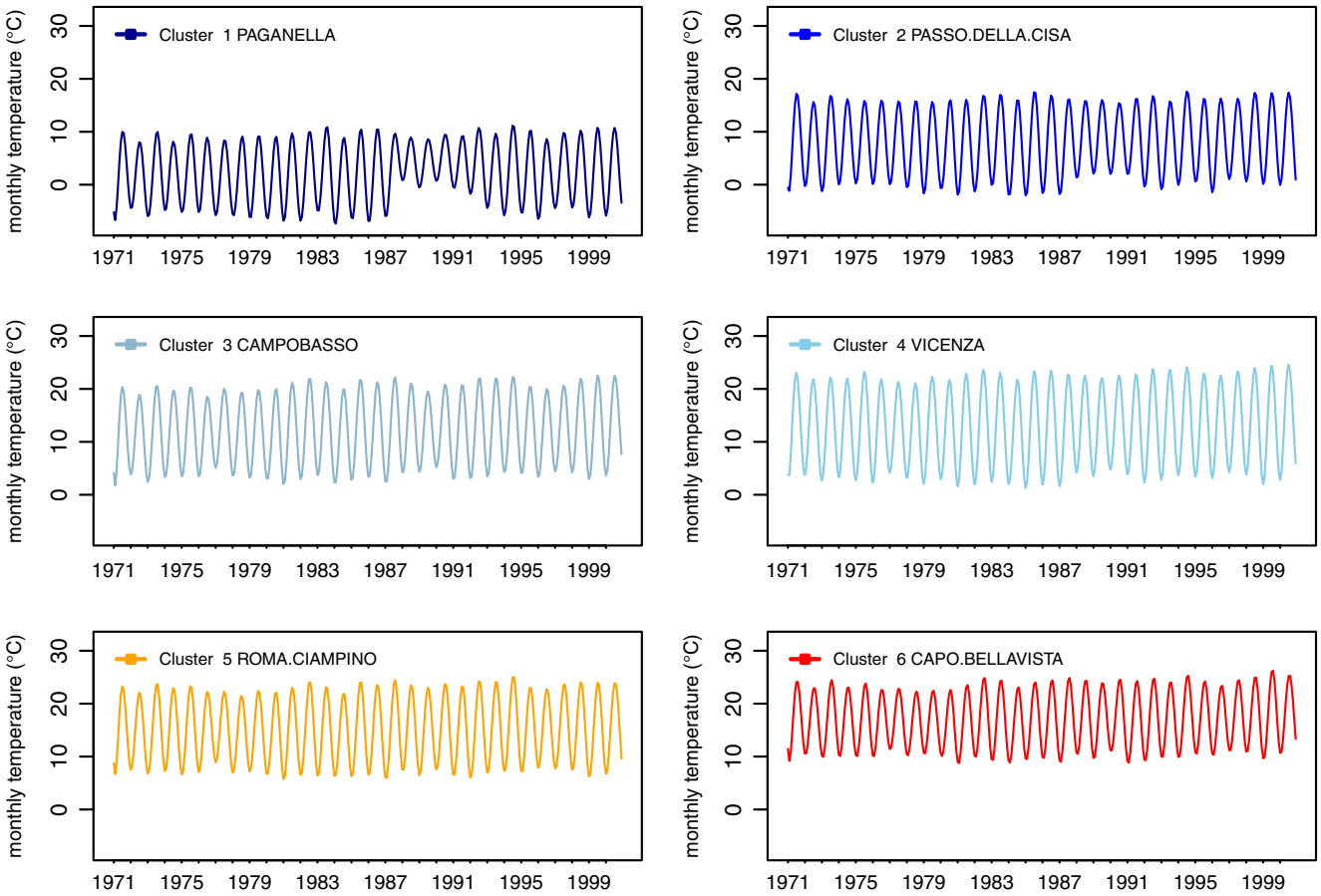
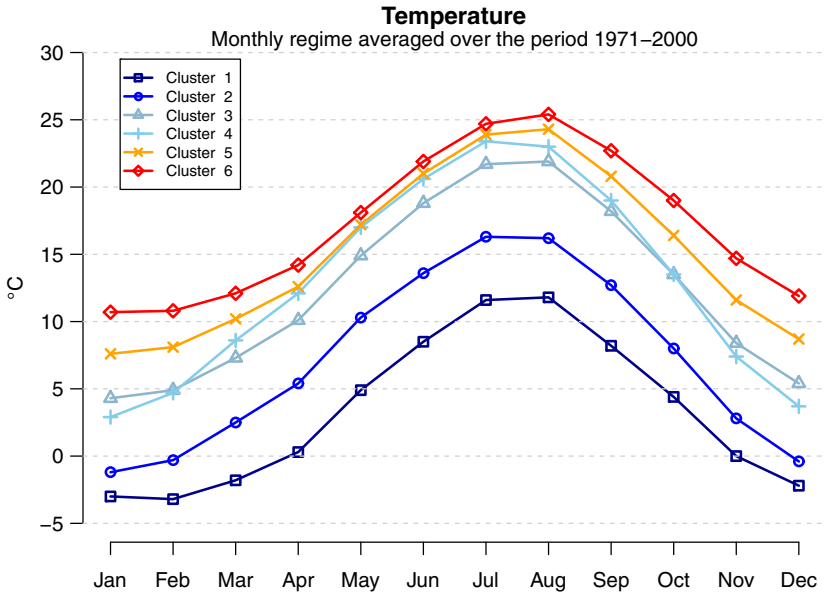


Fig. 5 Functional smoothing of the six medoids of temperature time series 1971–2000 (B-splines 4-monthly 5-degree)

for colder winter temperatures. This cluster analysis can be immediately related to the exposure to the main atmospheric circulations of the different regions. Warm clusters' (clusters 5 and 6) location, in the south and along the Tyrrhenian Coast, is linked to the southwestern flows forced by both cyclonic and anticyclonic circulation over Western Mediterranean Basin. In this area, only mountain stations such as those over the Appennini ridge and Mount Argentario belong to other clusters. The locations of cold clusters are linked to northeastern flows driven by cyclonic circulation over Eastern Europe and blocking condition over Central Europe that bring cold air masses into the Mediterranean Basin. A detailed summary of monthly and seasonal Tmed-MM 30-year averaged values of each group is given in Fig. 6. The Bsplines30 approach leads to results similar to the benchmark PCA in T-mode using three Pcs with respect to highlighted climatological features. The PCA based classification returns a unique coldest cluster of mountain stations in the north, whereas in Bsplines30, this cluster is more correctly divided into two separated groups. Examining Fig. 6, where the general features of cluster are depicted, it appears that the monthly levels of cluster 1 and cluster 2 are clearly different. In panel c, the Fourier-based map is reported. There, we choose six groups. As for the Bsplines30, clusters are very similar; however, there are some relevant differences: Sardinia is divided into two clusters and several stations around Rome are in a colder cluster with respect to the Bspline30; furthermore, the two Rome stations are classified into two different clusters. The general classification has several unclear aspects from a phenomenological point of view. In terms of the best clustering quality, Bspline30 with six groups reports an average silhouette width of 0.33 with no misclassified stations, PCA with five groups has an average silhouette width equal to 0.42 and two misplaced units, and Fourier with six groups reports an average silhouette width of 0.41 and one misplaced unit. Say k is the number of groups of each cluster. In the case of PCA, the best value of silhouette average width corresponds to $k = 2$, $k = 3$ is the second best, and our choice $k = 5$ is the third best. In the Fourier case, the best value of silhouette average width is found for $k = 3$, $k = 2$ is the second best, $k = 4$ is the third, and our choice $k = 6$ is the fourth best.



	Units (#)	DJF (°C)	MAM (°C)	JJA (°C)	SON (°C)
Cluster 1	4	−2.8	1.1	10.6	4.2
Cluster 2	5	−0.6	6	15.4	7.8
Cluster 3	16	4.9	10.8	20.8	13.4
Cluster 4	20	3.8	12.6	22.3	13.3
Cluster 5	24	8.1	13.3	23.1	16.3
Cluster 6	26	11.2	14.8	24	18.8

Fig. 6 Monthly and seasonal values of Tmed-MM averaged over 1971–2000 for six areas delineated by 4-monthly variability functional data (DJF December, January, February; MAM March, April, May; JJA June, July, August; SON September, October, November)

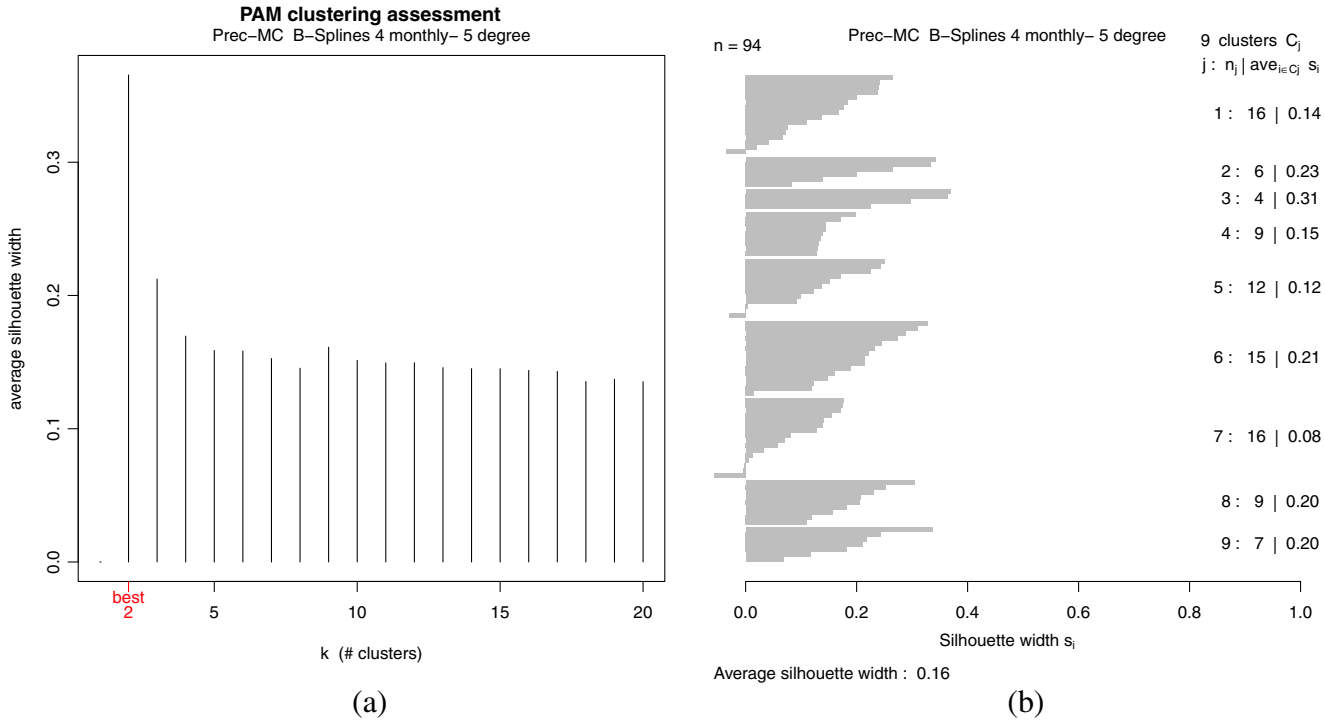


Fig. 7 Cluster algorithm assessment of square root Prec-MC for 4-monthly variability functional data: **a** Silhouette average width for determining the number of clusters and **b** silhouette width index for each group and for each unit included in the corresponding nine groups clustering

4.2 Results for Prec-MC

The more appropriate Bsplines30 model for Prec-MC is Bsplines30 4-monthly 5-degree with 90 fixed knots and five-degree piecewise polynomials. The value of average RMSE is 8.39 mm. Similar comments as in Section 4.1 on the model choice apply. Again, the smallest average RMSE is obtained with Bsplines30 3-monthly 3-degree but the increase in parameters number (from 94 to 122 parameters to be estimated) (see Table 2) and the final classification do not justify the choice of the 3-monthly model. For Prec-MC B-splines30 3-monthly 3-degree, the chosen number of groups is 7 with an average silhouette width of 0.15 which is the fourth best value and the first one with a climatic meaningful interpretation; there are 12 misplaced units according to the silhouette index and the classification is quite consistent with climate patterns. A comparison with our chosen classification reveals that several locations are wrongly

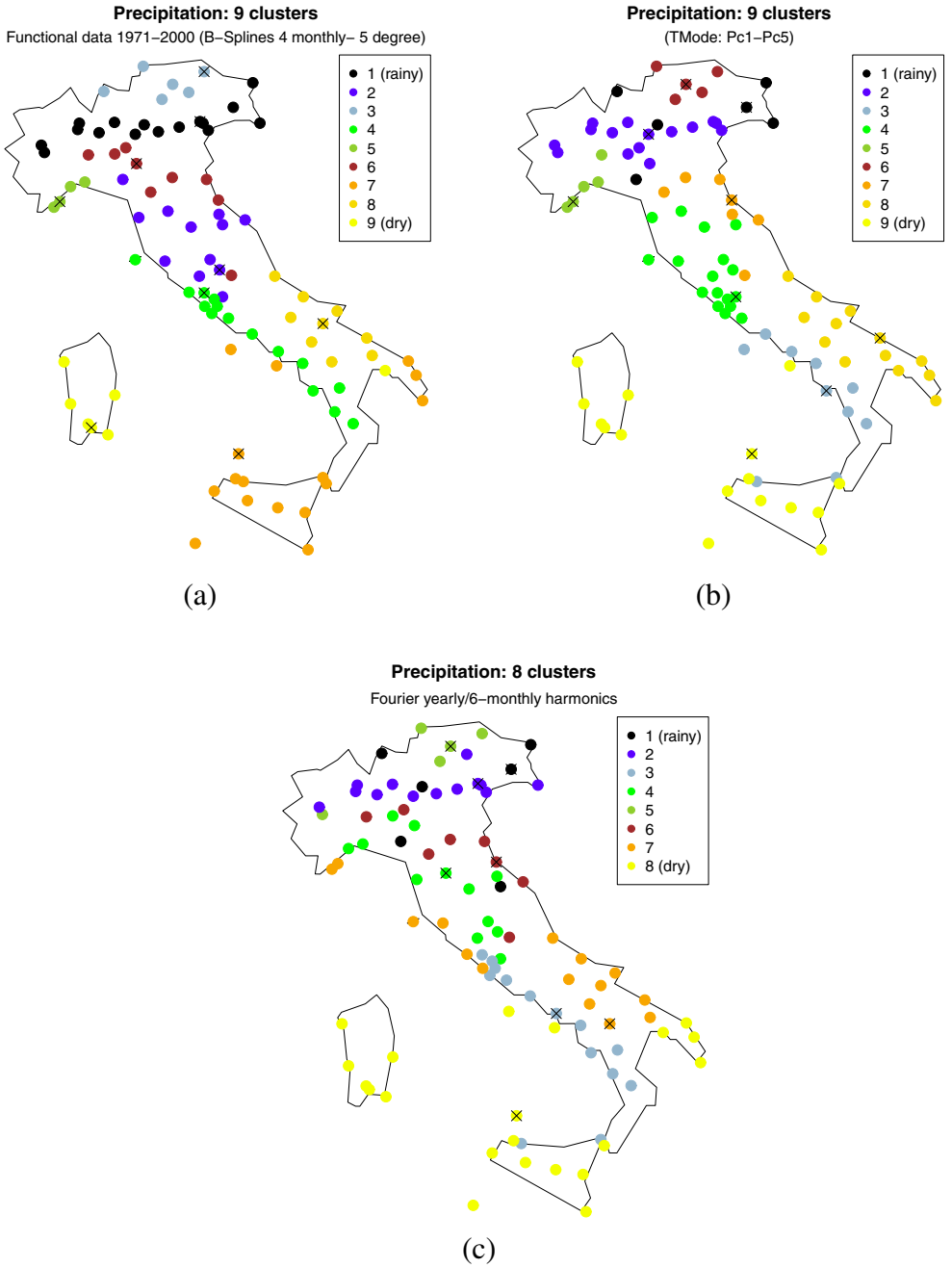


Fig. 8 Cluster maps of precipitation for 4-monthly variability functional data (**a**), PCA in T-mode method using five principal components (**b**), and Fourier with five bases of 12- and 6-monthly harmonics (**c**). *Crosses* in the maps indicate the location of cluster's medoids

classified into cluster 1 (stations along Po river) and two areas are not isolated in single clusters as it should be (stations near the Ligurian Sea and Sardinian stations). The Sardinian stations are correctly grouped if we consider the eight-group clustering, which is the sixth best choice in terms of average silhouette width (0.13) and counts 11 misplaced units. On the contrary, using the chosen model, we obtain an acceptable compromise between climatic interpretation of groups and statistical clustering quality indexes. This statement is corroborated by the following results. As far as the number of groups to be chosen is concerned, for precipitation data, the choice is less straightforward than with temperature data. Thus, very similar values of the average silhouette width index are obtained with 4- up to 20-cluster partitions (see Fig. 7a). Nevertheless, in spite of an average silhouette width value of 0.16 and 3 misplaced units (Fig. 7b), the nine-cluster partition is "the best" if we take into account all the information associated to the PAM algorithm, as isolation, diameter of clusters, separation, and silhouette width of each group (Fig. 8). Besides, this clustering returns a representation of climate features of precipitation which is consistent with well-known patterns of this variable for the Italian Peninsula. The smoothed time series of nine medoids represented in Fig. 9 reveals the high variability of precipitation and also highlights significant differences between groups. A detailed summary of yearly and seasonal Prec-MC 30-year averaged values of each group is given in Fig. 10, where we use line chart instead of bar chart to make a clearer graphical representation of precipitation regime. In the following, we refer to those values for the ordination of the groups from the rainiest to the driest and for a further description of the groups. As it comes out from panel a of Fig. 8, main patterns of variability are well reproduced and their identification improved with respect to the benchmark in PCA T-mode (panel b). In particular, it is worthwhile to evaluate the separation of the stations near the Ligurian Sea (cluster 5) and continental stations in the northwest (clusters 1 and 6) into different regions and the clear identification of two precipitation patterns in the northern and southern stations along the Po river (clusters 1 and 6). A central area extends from the Tyrrhenian to the Adriatic Sea (cluster 2) which is the second most rainy region (858 mm of total annual precipitation) behind the northern Po river area

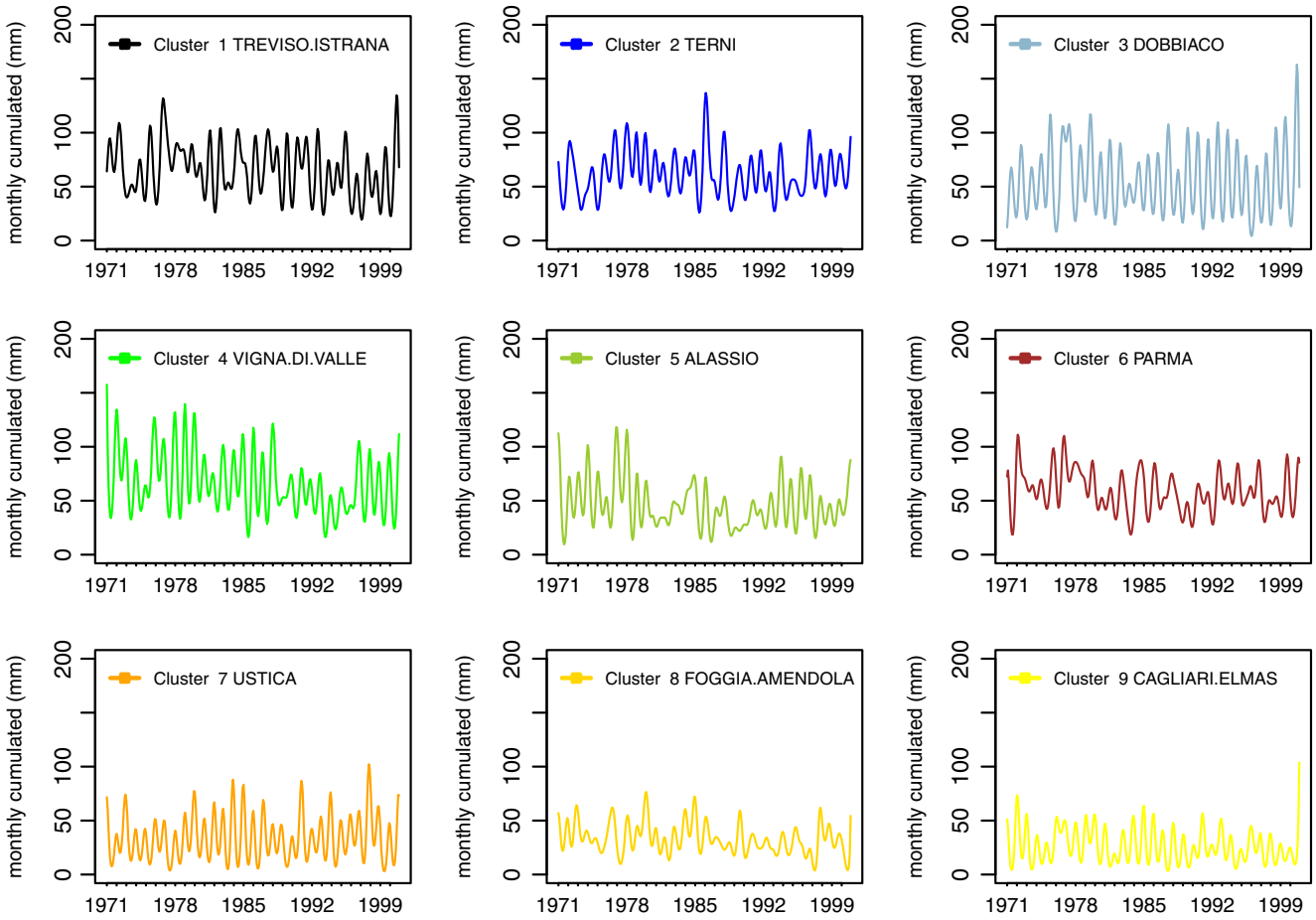
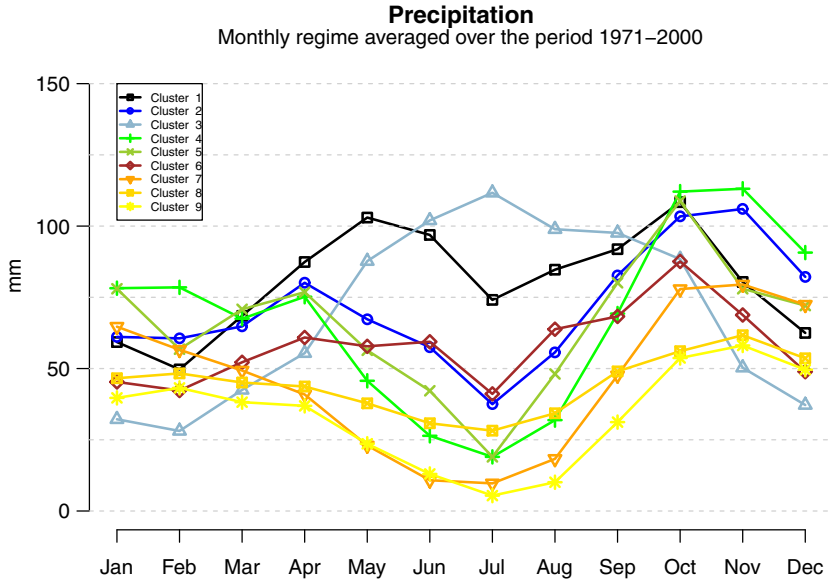


Fig. 9 Functional smoothing of the nine medoids of precipitation time series 1971–2000 (B-splines 4-monthly 5-degree)



	Units (#)	DJF (mm)	MAM (mm)	JJA (mm)	SON (mm)	YEAR (mm)
Cluster 1	16	171.5	258.9	255.6	280.7	966.7
Cluster 2	12	203.8	212.1	150.7	292	858.6
Cluster 3	6	97.6	185.6	312.6	236.3	832.1
Cluster 4	15	247.4	188.4	77.3	294.5	807.6
Cluster 5	4	206.9	203.9	109.2	266.8	786.8
Cluster 6	9	136.5	170.9	164.3	224.8	696.5
Cluster 7	16	193.8	113.3	38.8	205	550.9
Cluster 8	9	148.4	126.7	93.2	166.8	535.1
Cluster 9	7	132.5	98.8	28.5	143	402.8

Fig. 10 Monthly and seasonal values of precipitation averaged over 1971–2000 for nine areas delineated by 4-monthly variability functional data (seasonal precipitation values are obtained by cumulating monthly values; DJF December, January, February; MAM March, April, May; JJA June, July, August; SON September, October, November)

(966 mm). A coastal region along Tyrrhenian Sea (cluster 4) is also delineated which is fourth in the rainy ordered classification (807 mm). Regarding the south of the peninsula, Puglia is longitudinally divided into two areas (clusters 7 and 8) because of drier summer regime registered in the southern part (39 against 93 mm), which is similar to Sicily precipitation features. This is an improvement of the PCA T-mode classification. Finally, Sardinia is correctly classified as a unique cluster with the driest annual precipitation volume (402 mm) whereas the PCA-based classification proposes a unique cluster of Sicily and Sardinia. Besides, the clustering of stations reflects atmospheric patterns responsible for different precipitation regimes both at large scale and local scale. In particular, the Atlantic storm track determines the grouping of western areas (clusters 2, 4, and 5), of which clusters 2 and 4 are characterized by a prevalence of frontal precipitation and convective events, whereas cluster 5 precipitation signal is due to a more cyclogenetic and convective type of events (Harnik and Chang 2003). The continental and Alpine regions are characterized by large amount of precipitation due to an orographic enhancement mechanism driven by the presence of mountain ridges (clusters 1 and 3) and a distinct area (cluster 6) in the east side of Appennini lee ridge, which is dryer than the other northern clusters since it is not directly exposed to the moist westerly atmospheric flows. Similarly to the case of temperature, we perform functional clustering using Fourier basis as well. Following Laguardia (2011), we adopt 12- and 6-monthly harmonics. The classification map shows noticeable differences with respect to Bsplines30: the locations of the rainiest cluster are far from each other and, moreover, this spatial dispersion does not seem to have a physical motivation; cluster 2 is similar to cluster 1 of Bsplines30; locations by the Ligurian Sea do not have a clear identification as it is for our proposal; and, finally, the two major Italian islands Sicily and Sardinia are aggregated in a unique cluster (cluster 8), which is questionable as it is clear by looking at the yearly volume of precipitation of those groups when separated (Fig. 10). In terms of the best clustering quality, Bspline30 reports an average silhouette width of 0.16 with three misclassified units, PCA T-mode using five Pcs with nine groups returns a value of 0.31 average silhouette width and five misclassified units, while the Fourier-based analysis with eight groups has 0.33 average silhouette width with seven misplaced units. Say k is the number of groups of each clustering. In the case of PCA, the best value of silhouette average width is obtained for $k = 2$; $k = 11$, 12, and 13 have the same value of the index which correspond to the second best; and our choice $k = 9$ is the third best. In the Fourier case, the best value of silhouette average width

corresponds to $k = 2$, $k = 3$ is the second best, $k = 4$ is the third, and our choice $k = 8$ is the fourth best.

Finally, we replicate, as far as possible, the procedure proposed in Laguardia (2011) by adopting the same basis for the functional smoothing, i.e., the Fourier basis with 12- and 6-monthly harmonics and the *k*-means clustering method (not shown). However, in Laguardia's paper, the clustering algorithm is not entirely specified and then we choose the default Hartigan and Wong algorithm in the "stats" library of the R software with 25 random starts for the *k*-means clustering (Hartigan and Wong 1979). The method for selecting the optimal number of clusters is not specified in Laguardia; then, we choose the same number of clusters proposed by Laguardia, i.e., 6 for clustering the 94 stations of our data set. The predicted values of monthly regime obtained by Fourier functional smoothing for our data set is consistent with the climatology of the clusters' location. However, with the six clusters, several features captured by the Bsplines30 are not highlighted, and when a larger number of clusters have been tried, the same confusions seen with PAM classification are obtained.

5 Conclusion

This paper has presented a new protocol, based on functional clustering for determining homogeneous climate zones. We showed that by functional clustering, information on temporal pattern relative to the monthly time scale is summarized by a small number of coefficients and those coefficients determine a clear identification of variability mechanisms. The proposed method achieves this goal with a parametrization of function using penalized B-spline basis that returned a clear description of intra-annual variability. Description of the current distribution of local precipitation is made difficult by the high spatial and temporal variability of this parameter. Nevertheless, the regional distributions obtained not only correspond fairly well to the large, well-known physical regions of Italy but also go further, improving the classification determined by the standard methods. In fact, to identify climate regions using PCA-based methods requires a long and complex analysis of the reduced space to connect it to the physical world. In our approach, this is easily achieved by the choice of knot number and locations. Changing place and/or number of interior knots allows us to investigate different patterns of variability: long-term variability or trend (yearly variability over at least a 30-year interpolation period) and intra-annual variability (bimonthly, quarterly, 4-monthly, or 6-monthly variability across time series period). Further development of this approach is possible and has been investigated in our research to some extent (not reported in the present work). For instance, a decomposition in trend and short-term component of the time series is easily achieved by fitting a B-spline with yearly knots (trend) and a second B-spline with more knots to capture short-term features or a Fourier expansion with few harmonics to capture long-term cycles. Some caution must be used when using Fourier basis with relatively small number of station such as in our study. Indeed, the Fourier expansion reveals a tendency to over-smoothing (not shown in the paper) that influences classification results that may not be very clear, particularly with highly variable quantities such as precipitation. Thus, the Fourier smoothing seems to refer to a numerical smoothing rather than to a physical framework, and this drawback might be due to the loss of local element in the time domain. In fact, unlike Bsplines30, the smoothing of Fourier with 12- and 6-month harmonics attempts to reproduce the average features of monthly distribution of the time series smoothing out small and short-term changes. Moreover, the reproduction of Fourier predictive regime reveals that the addition of one supplementary harmonics does not let us to catch local element in time domain. However, if a very large number of monitoring stations are available as in Laguardia (2011), the strong smoothing effect of Fourier basis expansion may mitigate problems deriving from the large variability that is proper of large datasets.

In general terms our proposal, as described above, creates a very flexible framework in which analysis of climatological features can be carried out. In particular, the functional smoothing can be modified including, for example, both Fourier and penalized B-spline basis, the first to describe periodic components (regime) and the second to describe the trend; this combination of basis is especially effective when the periodicity in the data is not subjected to large changes in the considered time window. Other basis can be considered such as wavelets or combinations of B-splines with different number of parameters depending always on the aim of the study and the type of available data. In conclusion, we believe that the presented functional clustering approach is definitely much more flexible and easier to implement than the current PCA-based methods, regardless of the chosen basis of representation.

Acknowledgments This work has been developed within the context of Agroscenari project "Adaptation of Agricultural Management to climate change" funded by the Italian Ministry of Agriculture. The authors would like to thank two anonymous referees for the useful comments that helped in considerably improving the paper and Prof. M. Maugeri for his suggestions.

References

Abraham C, Cornillon PA, Matzner-Loeber E, Molinari N (2003) Unsupervised curve clustering using B-splines. Scand J Statist 30:581–595

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc, B 26:211–246

Brunetti M, Maugeri M, Monti F, Nanni T (2006) Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. Int J Climatol 26: 345–381

Ehrendorfer M (1987) A regionalization of Austria's precipitation climate using principal component analysis. Int J Climatol 7(1): 71–89

Fovell RG, Fovell MYC (1993) Climate zones of the conterminous United States defined using cluster analysis. J Climate 6: 2103–2135

Harnik N, Chang EKM (2003) Storm track variations as seen in radiosonde observations and reanalysis data. J Climate 16:480–495

Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28:100–108

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Kim BR, Zhang L, Berg A, Fan J, Wu R (2008) A computational approach to the functional clustering of periodic gene-expression profiles. Genetics 180:821–834

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

Laguardia G (2011) Representing the precipitation regime by means of Fourier series. Int J Climatol 31(9):1398–1407

Mennella C (1972) Il clima d'Italia nelle sue caratteristiche e varieta' quale fattore dinamico del paesaggi, vol II. Fratelli Conte Editore, Napoli

Preisendorfer RW, Mobley CD (1988) Principal component analysis in meteorology and oceanography. Elsevier, Amsterdam

R Development Core Team (2011) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

Ramsay JO, Silverman BW (1997) Functional data analysis. Springer, New York

Ramsay JO, Silverman BW (2002) Applied functional data analysis: methods and case studies. Springer, New York

Ramsay JO, Wickham H, Graves S, Hooker G (2011) Fda: functional data analysis. R package version 2.2.7. http://CRAN.R-project.org/package=fda

Reynolds A, Richards G, De La Iglesia B, Rayward-Smith V (1992) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5:475–504. doi:10.1007/s10852-005-9022-1

Richman MB (1986) Rotation of principal components. J Climatol 6:293–335

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20: 53–65

Sprent P (1998) Data driven statistical methods. Chapman & Hall, London

Sprent P, Smeeton NC (2001) Applied nonparametric statistical methods, 3rd edn. Chapman & Hall/CRC, London

Toreti A, Fioravanti G, Perconti W, Desiato F (2009) Annual and seasonal precipitation over Italy from 1961 to 2006. Int J Climatol 29(13):1976–1987

Trigo R et al (2006) Relations between variability in the Mediterranean region and mid-latitude variability. In: Lionello P, Malanotte-Rizzoli P, Boscolo R (eds) Mediterranean climate variability. Elsevier, Amsterdam, pp 179–226

Von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge

WMO (1989) Calculation of monthly and annual 30-year standard normals. WCDP n.10, WMO-TD/N.341, Geneva