

Time series clustering by a robust autoregressive metric with application to air pollution



Pierpaolo D'Urso^{a,*}, Livia De Giovanni^b, Riccardo Massari^a

^a Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, P.za Aldo Moro, 5-00185 Rome, Italy

^b Dipartimento di Scienze Politiche, LUISS Guido Carli, Viale Romania, 32-00197 Rome, Italy

ARTICLE INFO

Article history:

Received 17 July 2014

Received in revised form 8 November 2014

Accepted 10 November 2014

Available online 18 November 2014

Keywords:

Autoregressive model

Robust fuzzy C-medoids clustering

Outliers

Environmental chemistry

Pollutant concentration

Nitrogen monoxide (NO) emissions

ABSTRACT

In this paper, following a fuzzy approach and adopting an autoregressive parameterization, we propose a robust clustering model for classifying time series. In particular, by adopting a fuzzy partitioning around medoids approach, the suggested clustering model is able to define the so-called medoid time series, which is a representative time series of each cluster, and the membership degrees of each time series to the different clusters. The robustness of the proposed clustering model is guaranteed by the adoption of a suitable robust metric for time series, i.e. the so-called exponential distance measure. In this way, the clustering model is able to tolerate the presence of outlier time series in the clustering process. In particular, it is capable of neutralizing and smoothing the disruptive effect of outlier time series, preserving the original clustering structure of the dataset, by assigning to outlier time series approximately the same membership degrees across clusters. To illustrate the usefulness and effectiveness of the suggested time series clustering model, a simulation study and an application to air pollution time series are carried out. Comparison with some existing clustering procedures suggested in the literature shows several advantages of the proposed model.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The literature on time series clustering methods has increased considerably over the last two decades, with a large range of applications in many different fields, including environmental sciences, pharmaceuticals, genetics, neurosciences, computational biology, biomedical sciences, finance, econophysics, and neuromarketing. In particular, in the experimental studies the usefulness and effectiveness of the time series clustering proves to be of particular interest. For instance, in *genetics* time series clustering has been used to group genes considering profiles of time expression from cDNA microarrays experiments; in *biomedicine*, to classify signals caused by particular illnesses connected to those of healthy people (i.e., EEG and EMG time series); in *neuroscience*, to classify fMRI (functional magnetic resonance imaging) time series; in *pharmaceutics*, to cluster drug effects attending to patients' time-response after drug intake; in chemometrics, to classify natural products according to chemical profiles recorded at different times (e.g. chemical composition of wines in different years); in environmetrics, for checking the performances of an environmental monitoring network based on a set of air pollutant emissions recorded in different times (air pollution time series) by a set of monitoring stations, it is useful to classify the stations in homogeneous clusters.

For some references on the application of time series clustering/classification on the above-mentioned and other experimental areas, see [1].

As we can see below, we focus our empirical attention on the clustering of environmental time series. For an overview of the literature on the theoretical aspects of time series clustering/classification and their applications in environmental sciences and other experimental fields, see Section 2.

In this paper, we propose a robust clustering model for classifying time series based on a suitable parametric representation of univariate time series, i.e. an autoregressive representation. In particular, following a fuzzy approach, the proposed robust clustering model is based on the partitioning around medoids procedure. The robustness of the clustering model is guaranteed by the adoption of a proper robust metric, i.e. the so-called exponential distance measure. In this way, our clustering model is able to tolerate the presence of outlier time series in the clustering process; by neutralizing and smoothing the disruptive effect of outlier time series, preserving as almost invariant the clustering structure of the dataset, by assigning to outlier time series almost the same membership degrees across clusters. The proposed model is suitable for clustering time series exhibiting persistence over time. Notice that the fuzzy approach also allow us to identify peculiar patterns of time series, like switching time series, i.e. time series showing a pattern typical of a given cluster during a certain time period, and a completely different pattern (representative of another cluster) in another time period [2,3].

* Corresponding author.

E-mail addresses: pierpaolo.durso@uniroma1.it (P. D'Urso), ldegiovanni@luiss.it (L. De Giovanni), riccardo.massari@uniroma1.it (R. Massari).

For detailed discussions on the benefits connected to autoregressive model-based clustering approach, fuzzy methodology, partitioning around medoids procedure see D'Urso et al. [2,3]. We observe that the fuzzy approach has been employed for clustering/classification of sequences in other research areas, for instance in computational biology [4–15].

Following the guidelines proposed in [16–23], we will document the proposed algorithm for clustering time series according to the following procedures in order to make its development clearer and more useful: (i) select the information to be extracted by time series; (ii) select a proper distance measure for comparing time series; (iii) introduce or develop a powerful algorithm to operate the clustering; (iv) properly evaluate the accuracy of the clustering model; (v) establish resources for the development of the algorithm that are publicly available.

The paper is organized as follows. In Section 2, we present a review of the literature on time series clustering/classification methods, with particular attention to their application to environmental sciences. In Section 3, we introduce the fuzzy robust clustering model for time series. To illustrate the performances of the clustering model, and to assess its classification accuracy, we present and discuss the results of a simulation study in Section 4. As we focus our attention in particular on the usefulness and effectiveness of the time series clustering in environmental sciences, in Section 5 we utilize our clustering model for classifying air pollution time series. Conclusions are provided in Section 6.

2. Literature on clustering/classification of time series

In the literature several time series clustering methods have been suggested. For a survey on possible theoretical approaches, see, e.g., [1,24].

We remark that, following [1,24], time series clustering methods can be classified into three classes: 1) observation-based clustering: the time series clustering methods belonging to this approach are based on the actual time observations, i.e. observed time series or their transformations (see, e.g. [25,26]); 2) feature-based clustering: in this case, the methods are based on features derived for the time series (see, e.g., [27–31]); 3) model-based clustering: these methods are based on parameters estimates of model fitted to the time series (e.g. ARMA or ARIMA models) (see, e.g., [2,3]).

In this paper, we adopt the model-based approach. As we focus our attention in particular on the usefulness of time series clustering in environmental sciences in Section 2.1 we show a detailed overview on this field.

2.1. Literature of time series clustering/classification in environmental sciences

Clustering and classification methods have a crucial role in monitoring of the air quality [3]. In fact, “air quality monitoring is the main tool of local governments for the management and evaluation of air quality status. This practice follows technical regulation. An air monitoring network is usually composed of sites which measure atmospheric pollutants and weather variables. Classification of these monitoring stations is a method of network analysis and optimization. Classification highlights similarities among sites with respect to pollutant concentration levels and/or temporal profiles. Displaying groups on a map allows the identification of spatial patterns” [32]. Moreover, “in designing and maintaining a cost-effective monitoring network, it is important to recognize similarities and differences in the evolution of the variables sampled at different sites, in order to avoid or, at least, reduce redundancy. On the other hand, information collected by a redundant network, for example in the initial exploratory phase of a surveillance monitoring, could be extremely useful for partitioning a transition water body into homogeneous regions, for which different quality objectives may be established. With regard to the above issues, cluster

analysis methods (or unsupervised classification) can play a very important role” [33].

In the literature, different clustering-based techniques have been proposed for analyzing air pollution. Sanchez Gomes and Ramos Martin [34] considered the C-means clustering method for identifying sources in Valladolid (Spain). Bohm et al. [35] used cluster analysis for detecting temporal patterns of ozone. Sanchez et al. [36], Dorling et al. [37] and Ruijgrok and Romer [38] considered pollutant data and wind data in the cluster analysis. Miranda et al. [39] analyzed the concentration in Mexico City utilizing the correlation coefficient and the Ward clustering method. Romo-Groger et al. [40] considered a similar analysis in Chile using the average linkage algorithm. Ludwig et al. [41] used cluster analysis for studying the daily ozone maxima in California. Lavecchia et al. [42] employed a complete linkage-based procedure on the monitoring network in Lombardia (Northern Italy) to evaluate similarities among ozone monitoring sites in terms of concentration levels and temporal trends. The authors compared the ozone patterns by using the Euclidean distance and the correlation coefficient in the clustering procedure. Wongphatarakul et al. [43] clustered sampled sites with similar characteristics by considering PM_{2.5} chemical databases from seven sites around the world. By considering the Euclidean distance and the Ward clustering method, Ionescu et al. [44] classified estimated pollutant concentration fields obtained by utilizing the so-called *thin plate spline* functions, analyzing nitrogen dioxide data during peak episodes in Paris. Hierarchical clustering has been used to identify distinct sources of volatile organic compounds based on the grouping of the measured concentrations [45]. Moreover, hierarchical clustering could provide a description of regional chemical and transport processes associated with particular regimes and could provide information about the most relevant sources in the development of pollution episodes. Saksena et al. [46] clustered monitoring sites in Delhi on the basis of nine years of monthly average concentration data for three pollutants, i.e. nitrogen dioxide, sulfur dioxide and suspended particulate matter. They considered four agglomerative hierarchical clustering methods -i.e. average linkage, single linkage, complete linkage and centroid method- and Euclidean and Squared Euclidean distance. They observed that the most consistent results are obtained using the Euclidean distance and the average linkage method. A study of the data from Santiago's monitoring network was done by Silva and Quiroz [47] considering an index of multivariate effectiveness, based on Shannon information index. They found that air pollution data (CO, PM₁₀, O₃ and SO₂) from one of the stations (Parque O'Higgins) could be reproduced by using information from the other stations. In order to identify the representative stations for subsequent analysis of ozone concentration Gabusi and Volta [48] considered a hierarchical clustering approach for classifying Northern Italy measurement stations. Beaver and Palazoglu [49] used an aggregate solution of *k*-means clustering to characterize classes of ozone episodes occurring in the San Francisco bay. Cluster analysis based on the Pearson correlation coefficient is used by Gramsh et al. [50] on particulate matter and ozone data collected by Santiago de Chile's network to find city sectors with similar pollution behavior. Cluster analysis has been used to cluster back trajectories, in order to identify different classes of synoptic regimes over the duration of the trajectories [51,52]. Morlini [53] classified monitoring stations of ozone, sulfur dioxide, and carbon monoxide in Emilia Romagna region (Northern Italy) using a dynamic time-warping cost function as dissimilarity measure in average and complete linkage algorithms. In this framework, cluster analysis is used to classify fields obtained from observed data to identify “prototype” of spatial patterns. By considering a functional representation, Bengtsson and Cavanaugh [54] modeled the observed time series in a state space setup and classified the sites via hierarchical clustering methods relying on disparity measures based on Kullback information. Kim et al. [55] employed *k*-means clustering for classifying sites based on the temporal fluctuation of PM_{2.5}. In order to identify city areas with similar air pollution behavior and to locate emission sources, Pires et al. [56,57]

applied Principal Components Analysis and Cluster Analysis i.e., the Euclidean-based average linkage method, to the mass concentrations of SO₂ and PM₁₀ [56] and CO, NO₂ and O₃ [57] collected in the air quality monitoring network of Oporto Metropolitan Area. Lau et al. [58] used the complete linkage algorithm and Euclidean distance to analyze NO₂ and PM₁₀ measurements. Ibarra-Berastegi et al. [59] suggested a procedure to identify redundant sensors and evaluate a network's capability to correctly follow and represent SO₂ fields in Bilbao, in the frame of a continuous network optimization process. They used Self-Organizing Maps (SOMs), hierarchical cluster analysis (i.e. the Euclidean distance-based single linkage method), and Principal Component Analysis. The procedure is developed and tested at this particular location (Bilbao), but it is general enough to be useable at other places as well, since it is not tied neither to the particular geographical characteristics of the place, nor to the phenomenology of the air quality over the area. Pakalapati et al. [60] used hierarchical clustering and sequencing to group air flow patterns associated with elevated ozone concentrations. Karaca and Camci [61] evaluated the effects of long-range transport patterns of air masses to the particulate matter (PM₁₀) concentrations observed in Istanbul using the Self Organizing Maps (SOM). Lu et al. [62] considered Principal Component Analysis and Cluster Analysis (i.e. between-group linkage method) for optimizing and managing the air quality monitoring stations in Hong Kong. Ignaccolo et al. [36] used a partition around medoid (PAM) algorithm to site classification in the air quality network of Piemonte (Northern Italy). They chose PAM for practical reasons, since "PAM provides representative objects for each cluster which gives policy makers real sites to look at in order to quickly monitor general trends in a region. Moreover, it gives a suggestion about the appropriate choice of clusters' number and provides exhaustive clustering features by a graphical display (called silhouette plot). Furthermore, the so-called silhouette width represents a belonging measure of a site to a cluster such that if a site misclassification happens we have a warning. Thus, policy makers can decide to move a site to the neighboring cluster suggested by PAM" [32]. The procedure suggested by Ignaccolo et al. [32] has the advantage of clustering different sites around prototype sites, i.e. medoid sites that are not virtual sites but real sites belonging to the set of considered sites. However, as noted by Pastres et al. [33], among other this approach is not entirely satisfactory since the uncertainty in the partition is not quantified. "This issue could be relevant when deciding to redesign a monitoring network, since the location of 'uncertain' sites may be considered when establishing the boundaries between homogeneous areas" [33]. In order to overcome both limitations, Pastres et al. [33] proposed to combine functional data analysis and probabilistic cluster analysis methods, which allow one to estimate the probability that a given object belongs to a given cluster. Clustering has been used to describe diurnal variation in gaseous and particle pollutants by Adame et al. [63] and Flemming et al. [64]. Austin et al. [65] used cluster analysis to identify distinct daily multipollutant profiles at a given site, Boston. Austin et al. [66] used cluster analysis to group sites across the United States based on their PM_{2.5} composition profiles using data collected between 2003 and 2008. The main interest is to identify long-term

differences in the composition of PM_{2.5} across the different sites. These clusters of cities will then be characterized and validated based on a profilation of physico-chemical characteristics, geographic locations, emission profiles, population density and position with respect to major emitter sources. The work of Chaparro et al. [67] focused on the study of lichen species and their relationship with pollutants released by urban and industrial activities, using Principal Coordinate Analysis and *k*-means clustering. D'Urso et al. [3] suggested a redundancy analysis for air pollution monitoring systems by using a procedure based on the so-called Autoregressive model-based Fuzzy C-Medoid Clustering (AR-FCM_dC) algorithm. In particular, they analyzed the daily time series of CO and NO emissions in Rome. Ignaccolo et al. [68] proposed to partition a land in zones characterized by different criticality levels of atmospheric pollution considering pollutant time series as functional data (Functional Zoning). Specifically, they considered air pollutant time series of Piemonte (Northern Italy) provided by a deterministic air quality model on a regular grid, and preprocessed by assimilating observations, as functional data. Thus, they classified them by using functional clustering, where the Partitioning Around Medoids (PAM) algorithm is embedded [32] in place of the *k*-means one, as proposed by Abraham et al. [69]. Elangasinghe et al. [70] analyzed PM₁₀ and PM_{2.5} time series for a coastal site using artificial neural network modeling and *k*-means clustering. Malley et al. [71] applied hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification. Ensor et al. [72] introduced a strategy to identify point source impact on air pollution time series observed at each monitor by modeling observed hourly counts of exceedances above a pollutant threshold. They focused their study on benzene levels that exceed 0.4 parts per billion volume (ppbv) in the state of Texas. First of all, an observation-driven negative binomial regression model is used to capture autocorrelation in daily counts over time. Because there are many days in which more zero counts (representing no exceedances) are observed than would be expected for a negative binomial distribution, they included a zero-inflation component to account for this effect. Furthermore, they adopted the Gaussian plume model (GPM) for atmospheric dispersion to create covariates designed to measure the impact of emissions based on the locations of the leading point source contributors. These covariates represent the effect of an emissions source registered at a monitor. They also incorporated atmospheric conditions, such as wind speed, wind direction, and solar radiation in covariate construction. Finally, they developed a model-based approach to clustering the zero inflated count series obtained from each monitor. In particular, they used an empirical Kullback-Leibler divergence measure to quantify the similarity or dissimilarity between the modeled time series and a hierarchical clustering algorithm. In this way, the authors believe that understanding the common patterns in counts of observed threshold exceedances allows to identify similarities in the influence of the point sources on sets of monitors, as well as enables to identify monitoring sites, often spatially contiguous, representing similar (and dissimilar) behavior in pollution patterns.

3. Fuzzy clustering of time series by a robust metric

3.1. A robust distance measure for time series

Let consider a set of zero mean invertible ARIMA(*p,d,q*)(*P,D,Q*)_s processes \mathbf{Z} , $\mathbf{Z} \equiv \{Z_{it} : t = 1, \dots, T; i = 1, \dots, I\}$, where *p* is the order of the autoregressive (AR) component, *q* is the order of the moving average (MA) component, *d* is the differencing order needed to eliminate a stochastic trend, *P*, *Q* and *D* are the orders of the seasonal part of the model, and *s* is the period of the seasonal pattern.

Using the standard Box and Jenkins notation, the generic process is defined as follows:

$$\varphi_i(B)\nabla^d\nabla_s^D Z_{it} = \vartheta_i(B)\varepsilon_{it}, \quad (1)$$

where ε_{ti} is a univariate white noise (WN) process with mean 0 and constant variance σ^2 , B is the backshift operator such that $B^k Z_{ti} = Z_{(t-k)i}$, $\forall k = 0, \pm 1, \dots$, ∇ is the differencing operator such that $\nabla^d Z_{ti} = Z_{ti} - Z_{(t-d)i} = (1 - B)^d Z_{ti}$, the polynomials

$$\begin{aligned} \varphi_i(B) &= \phi_i(B)\Phi_i(B^S) = (1 - \phi_{1i}B - \dots - \phi_{pi}B^p)(1 - \Phi_{1i}B^S - \dots - \Phi_{pi}B^{Sp}) \\ \vartheta_i(B) &= \theta_i(B)\Theta_i(B^S) = (1 - \theta_{1i}B - \dots - \theta_{qi}B^q)(1 - \Theta_{1i}B^S - \dots - \Theta_{qi}B^{Sq}) \end{aligned}$$

for any $s \geq 0$, have no common factors, and all the roots of $\varphi_i(B) \cdot \vartheta_i(B) = 0$ lie outside the unit circle. Finally, let us assume that possible anomalous observations or deterministic components (i.e. mean level, calendar effects, trading days) have been suitably removed from the time series [73].

Since the processes are assumed to be invertible, Z_{ti} can be represented in terms of its past values according to the infinite autoregressive, $AR(\infty)$, formulation, i.e.:

$$\pi_i(B)Z_{ti} = \varepsilon_{ti}, \tag{2}$$

where $\pi_i(B) = (1 - B)^d(1 - B^S)^D \varphi_i(B)\vartheta_i^{-1}(B) = 1 - \sum_{j=1}^{\infty} \pi_{ji}B^j$ and $\sum_{j=1}^{\infty} |\pi_{ji}| < \infty$. The coefficients π_{ji} are denoted as π -weights.

Then, the AR distance between two processes, Z_{ti} and $Z_{t'i'}$, is defined as follows [6]:

$$d_{ii'} = d(Z_{ti}, Z_{t'i'}) = \left[\sum_{j=1}^{\infty} (\pi_{ji} - \pi_{j'i'})^2 \right]^{\frac{1}{2}}, \tag{3}$$

i.e., is the Euclidean distance between the vectors of the π -weights of the two $AR(\infty)$ formulations.

The $AR(\infty)$ formulation of the processes Z_{ti} can be approximated with a process $AR(J_i)$, so that the contributions of the π -weights π_{ji} for $j = J_i + 1, J_i + 2, \dots$, is negligible.

In the following we denote the observed time series with z_{ti} , ($i = 1, \dots, I; t = 1, \dots, T$), which represent the finite realizations of the zero mean invertible $ARIMA(p, d, q)(P, D, Q)_S$ processes. Let $\mathbf{Z} \equiv \{z_{ti} : t = 1, \dots, T; i = 1, \dots, I\}$ be the set of the observed time series, which are represented by means of the truncated AR representations $AR(J_i)$, $i = 1, \dots, I$.

Following Piccolo [74], the AR distance between two time series z_{ti} and $z_{t'i'}$ is defined as follows:

$$d_{ii'} = d(z_{ti}, z_{t'i'}) = \left[\sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{j'i'})^2 \right]^{\frac{1}{2}}, \tag{4}$$

where $J = \max(J_i, J_{i'})$, and the “hat” symbol denote the estimates of the π -weights.

Note that, given two generic processes Z_{ti} and $Z_{t'i'}$, generally $J_i \neq J_{i'}$. When the orders of the truncated AR representations of two time series differ, say $J_i > J_{i'}$, by adopting the so-called “zero-padding” approach (see, e.g. [3]) we could add to the shortest vector of the estimated π -weights $J_i - J_{i'}$ zeros to equalize the lengths of the two vectors.

As observed above, the AR distance is the Euclidean distance between two vectors of π -weights representing two time series. The Euclidean metric is widely used in real-world applications for its properties. However, Euclidean metric may not be robust in a noisy environment, as a collection of time series usually is. As a consequence, results from an objective function based on the Euclidean metric could be biased if data are contaminated by one or more outliers.

Wu and Yang [75] observed that the solution of an objective function based on the Euclidean metric can be written as a weighted sum of the observed data point with weights all equal to 1, irrespective of the fact that a data point lies close to the bulk of the data set or it is an outlier. Based on this statement, the authors proposed to adopt a more robust distance, i.e., the so-called exponential distance. The proposed exponential distance gives different weights to each data point, according to whether a data point is noisy or not. In particular, the exponential distance assigns small weights to outliers and larger weights to those data points that are more compact in the data set.

In order to neutralize the disruptive effects of possible outliers in the computation of pairwise distance measure between two processes, adopting the idea suggested by Wu and Yang [75] for not time-varying data, we suggest to compare each pair of processes Z_{ti} and $Z_{t'i'}$, provided that both processes are stationary and invertible, by means of:

$$d_{ii'} = d(z_{ti}, z_{t'i'}) = \left\{ 1 - \exp \left\{ -\beta \sum_{j=1}^{\infty} (\pi_{ji} - \pi_{j'i'})^2 \right\} \right\}^{\frac{1}{2}}. \tag{6}$$

Thus, for comparing each pair of time series z_{ti} and $z_{t'i'}$, we have:

$$d_{ii'} = d(z_{ti}, z_{t'i'}) = \left\{ 1 - \exp \left\{ -\beta \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{j'i'})^2 \right\} \right\}^{\frac{1}{2}}, \tag{7}$$

where, as above, $J = \max(J_i, J_{i'})$.

Note that, following Zhang and Chen [76], it is easy to prove that the AR-based robust distances (6) and (7) are metrics.

Wu and Yang [75] suggest that the value of β should be determined as the inverse of the variability in the data (the more the variability in the data the less the value of β). In this way, the value of β appropriately affects the distances (6), and hence (7), in terms of robustness to outliers.

Fig. 1 shows that in the presence of low variability of the data (high value of β) increasing distances receive a lower weight than in the case of high variability. See Remark 4 for further insights on how β is selected.

Remark 1. A prototypical case for the AR-based robust distance.

A prototypical case for the AR-based robust distance (6) is obtained when the time series represented by a process ARMA(1,1). Let Z_{ti} and $Z_{t'i'}$ be two stationary and invertible processes ARMA(1,1) processes. The $AR(\infty)$ coefficients corresponding to the processes are:

$$\pi_{jk} = (\phi_k - \theta_k)\theta_k^{j-1}; k = i, i'; j = 1, 2, \dots \tag{8}$$

Then, by substituting Eq. (8) in Eq. (6) we obtain the following expression:

$$\begin{aligned} d_{ii'} &= \left\{ 1 - \exp \left\{ -\beta \sum_{j=1}^{\infty} (\pi_{ji} - \pi_{j'i'})^2 \right\} \right\}^{\frac{1}{2}} = \left\{ 1 - \exp \left\{ -\beta \sum_{j=1}^{\infty} [(\phi_i - \theta_i)\theta_i^{j-1} - (\phi_{i'} - \theta_{i'})\theta_{i'}^{j-1}]^2 \right\} \right\}^{\frac{1}{2}} \\ &= \left\{ 1 - \exp \left\{ -\beta \left[(\phi_i - \theta_i)^2 \sum_{j=1}^{\infty} \theta_i^{2(j-1)} + (\phi_{i'} - \theta_{i'})^2 \sum_{j=1}^{\infty} \theta_{i'}^{2(j-1)} - 2(\phi_i - \theta_i)(\phi_{i'} - \theta_{i'}) \sum_{j=1}^{\infty} (\theta_i \theta_{i'})^{j-1} \right] \right\} \right\}^{\frac{1}{2}}. \end{aligned} \tag{9}$$

By exploiting the geometric series in Eq. (9) we have

$$\begin{aligned} \sum_{j=1}^{\infty} \theta_k^{2(j-1)} &= \frac{1}{1 - \theta_k^2}; k = i, i' \\ \sum_{j=1}^{\infty} (\theta_i \theta_{i'})^{j-1} &= \frac{1}{1 - \theta_i \theta_{i'}}. \end{aligned}$$

Then, by making use of the properties of the exponential function, from Eq. (9) we obtain:

$$d_{ii'} = \left\{ 1 - \frac{\exp \left\{ -\beta \frac{\phi_i - \theta_i}{1 - \theta_i^2} \right\} \exp \left\{ -\beta \frac{\phi_{i'} - \theta_{i'}}{1 - \theta_{i'}^2} \right\}}{\exp \left\{ -2\beta \frac{(\phi_i - \theta_i)(\phi_{i'} - \theta_{i'})}{1 - \theta_i \theta_{i'}} \right\}} \right\}^{\frac{1}{2}}. \tag{10}$$

This result could be further generalized for computing pairwise distances between processes belonging to the sub-classes AR(1), MA(1), ARIMA(1,1,0) and ARIMA(0,1,1) simply letting some parameters equals to 0 or 1 [74,77].

Remark 2. A weighted version of the robust AR-based distance

Let us now consider a more general version of the robust AR-based distance (7).

Let $\hat{\pi}_i = \{\hat{\pi}_{i1}, \dots, \hat{\pi}_{iJ_i}\}$ and $\hat{\pi}_{i'} = \{\hat{\pi}_{i'1}, \dots, \hat{\pi}_{i'J_{i'}}\}$ be the vectors of the estimates of the parameter of the truncated processes AR(J_i) and AR($J_{i'}$), with $J = \max(J_i, J_{i'})$, respectively, which represent the observed time series z_{ti} and $z_{t'i'}$. We can define a weighted version of the distance $d_{ii'}$ as follows:

$$\Omega d_{ii'} = \left\{ 1 - \exp \left\{ -\beta (\hat{\pi}_i - \hat{\pi}_{i'})' \Omega (\hat{\pi}_i - \hat{\pi}_{i'}) \right\} \right\}^{\frac{1}{2}} \tag{11}$$

where Ω is a matrix of weights. Possible choices of Ω have been shown in D'Urso et al. [3].

Remark 3. A correlation-based version of the robust AR-based distance

Notice that, the robust AR-based squared Euclidean distance $d_{ii'}^2$ can be formalized by the correlation coefficient; i.e. -since $\sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{j'i'})^2 = 2J(1 - r_{ii'})$, where $r_{ii'}$ denoted the correlation coefficient between z_{ti} and $z_{t'i'}$ - we have $d_{ii'} = d(z_{ti}, z_{t'i'}) = 1 - \exp\{-2\beta J(1 - r_{ii'})\}$.

3.2. A robust fuzzy clustering model for time series

Let $\mathbf{z}_t \equiv \{z_{t1}, \dots, z_{ti}, \dots, z_{tJ}\}, \forall t = 1, \dots, T$ be a set of I observed time series, and $\tilde{\mathbf{z}}_t \equiv \{\tilde{z}_{t1}, \dots, \tilde{z}_{tc}, \dots, \tilde{z}_{tC}\}, \forall t = 1, \dots, T$ be a subset of \mathbf{z}_t with cardinality C . Also, let $\hat{\pi}_j \equiv \{\hat{\pi}_{j1}, \dots, \hat{\pi}_{ji}, \dots, \hat{\pi}_{jJ}\}, \forall j = 1, \dots, J$ be the corresponding autoregressive coefficients of the truncated AR representation of the I time series, and $\hat{\pi}_j \equiv \{\hat{\pi}_{j1}, \dots, \hat{\pi}_{jc}, \dots, \hat{\pi}_{jC}\}, \forall j = 1, \dots, J$ be a subset of $\hat{\pi}_j$ with cardinality C . Note that we are assuming that all the truncated AR representations are of the same order J . This is a viable assumption, since the zero padding approach explained above.

By considering the (squared) distance (Eq. (7)), we obtain the following AR-based Fuzzy C-Medoid Clustering with Exponential Distance (AR-FCMdC-Exp) model:

$$\left\{ \begin{aligned} \min : & \sum_{i=1}^I \sum_{c=1}^C u_{ic}^m \left[1 - \exp \left\{ -\beta \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{jc})^2 \right\} \right] \\ & \sum_{c=1}^C u_{ic} = 1, \quad u_{ic} \geq 0. \end{aligned} \right. \tag{12}$$

where $m > 1$ is a weighting exponent that controls the fuzziness of the obtained partition; $\hat{\pi}_{jc}$ ($j = 1, \dots, J$) is the medoid for cluster c ; u_{ic} indicates the membership degree of the i -th unit in the c -th cluster.

Following Wu and Yang [75], the local optimal solution for the objective function in Eq. (12) is:

$$u_{ic} = \left(\frac{\sum_{c'=1}^C \left[\frac{1 - \exp\left\{-\beta \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{jc'})^2\right\}}{1 - \exp\left\{-\beta \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{jc})^2\right\}} \right]^{\frac{1}{m-1}}}{\sum_{c'=1}^C \left[\frac{1 - \exp\left\{-\beta \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{jc'})^2\right\}}{1 - \exp\left\{-\beta \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{jc})^2\right\}} \right]^{\frac{1}{m-1}}} \right)^{-1} \quad (13)$$

In a C-means framework, Wu and Yang [7] state that, from a theoretical standpoint, the C-means clustering model based on the exponential distance is more robust than the model based on the Euclidean norm.

The value of β , determined as the inverse of the variability in the data (the more the variability in the data the less the value of β), appropriately affects the membership degree (Eq. (13)) in terms of robustness to outliers.

Fig. 2 shows different membership curves, for different values of β , obtained with the AR-FCMdc-Exp model applied to a simulated dataset with two clusters with centers in 0.5 and 0.6. The curve with circle points represents the AR-FCMdc membership. If β is very small (high variability in the data) the AR-FCMdc-Exp membership curve is very close to the AR-FCMdc membership curve which well represents fuzzy boundaries; if β is very large (low variability in the data) the AR-FCMdc-Exp membership curve (shaped as a step function) is very different from to the AR-FCMdc membership curve as it assigns membership 0.5 to data that are only slightly distant from the centers, well representing the characteristic of separation between the clusters. See Remark 4 for further insights on how β is selected.

Remark 4. Computational characteristics of AR-FCMdc-Exp

The computational characteristics of the AR-FCMdc-Exp model are the following:

Selection of β : we select β in the following way:

$$\beta = \left(\frac{\sum_{i=1}^I \sum_{j=1}^J (\hat{\pi}_{ji} - \hat{\pi}_{jk})^2}{I} \right)^{-1} \quad (14)$$

where

$$\hat{\pi}_{jk} : k = \operatorname{argmin}_{1 \leq j' \leq I} \sum_{i'=1}^J (\hat{\pi}_{ji'} - \hat{\pi}_{j'k})^2$$

i.e., $\hat{\pi}_{jk}$, $j = 1, \dots, J$ are the AR coefficients of the time series which is the closest to all other time series.

Algorithm The steps of the algorithm are shown below:

Algorithm 1: AR-FCMdc-Exp.

Fix C , RS and *max. iter*;

Compute β by using Eq. (14);

Set $rs = 0$;

Repeat

Set *iter* = 0

Repeat

Pick initial medoids: $\hat{\Pi} \equiv \{\hat{\pi}_{j1}, \dots, \hat{\pi}_{jc}, \hat{\pi}_{jc}; j = 1, \dots, J\}$;

Store the current medoids: $\hat{\Pi}_{\text{OLD}} = \hat{\Pi}$;

Compute u_{ic} by using Eq. (13);

Select the new medoids $\hat{\Pi} \equiv \{\hat{\pi}_{j1}, \dots, \hat{\pi}_{jc}, \hat{\pi}_{jc}; j = 1, \dots, J\}$;

$q = \operatorname{argmin}_{1 \leq i' \leq I} \sum_{i''=1}^I u_{i''c}^m [1 - \exp\{-\beta \sum_{j=1}^J (\hat{\pi}_{ji''} - \hat{\pi}_{ji'})^2\}]$;

return $\hat{\pi}_{jc} = \hat{\pi}_{jq}$

iter = *iter* + 1;

Until $\hat{\Pi}_{\text{OLD}} = \hat{\Pi}$ or *iter* = *max.iter*;

$rs = rs + 1$

Until $rs = RS$.

Local optima: the algorithm 1 falls in the category of Alternating Cluster Estimation paradigm [78]; as for other recursive algorithm, it is not guaranteed that the global minimum is reached. Thus, more than one random start (RS) is suggested to obtain a stable solution.

Detection of C : the number of clusters C can be pre-determined by considering fuzzy cluster-validity indices (see [28]).

The validity criterion considered in this paper is the Fuzzy Silhouette (FS) [79]. The individual silhouette s_i ($i = 1, \dots, I$) is a measure of the closeness of i to the objects in the highest membership cluster, with respect to the distance to objects in other clusters, and is defined as [80]:

$$s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}}$$

where a_i is the average distance of time series i to all other objects belonging to its highest membership cluster and b_i is the average of the minimum distances of time series i to all the time series belonging to another cluster.

FS is a weighted average of s_i , with weights that take into account the membership degrees of each unit:

$$FS = \frac{\sum_{i=1}^I (u_{ir} - u_{iq})^\alpha s_i}{\sum_{i=1}^I (u_{ir} - u_{iq})^\alpha} \tag{15}$$

where u_{ir} and u_{iq} are the first and second largest elements of the i -th row of the fuzzy partition matrix $\mathbf{U} = \{u_{ic} : i = 1, \dots, I; c = 1, \dots, C\}$; α is an optional user defined weighting coefficient. In our case, we set $\alpha = 1$. The harder the partition matrix, the smaller the impact of changes in α . The Crisp Silhouette (CS), i.e. the average silhouette width, is obtained as a particular case of FS by setting $\alpha = 0$. The higher the value of FS, the better the assignment of the objects to the clusters.

Detection of m : following Kamdar and Joshi [81], in the application we set $m = 1.5$. However, in the simulation study, different values of m are selected, in order to check the influence of the fuzziness parameter on the results (see Section 3).

Computational complexity: the proposed AR-FCM \hat{C} -Exp algorithm could be very computationally intensive with large sample, since it is based on an exhaustive search for the medoids. Following Krishnaputan et al. [82], a “linearized” version of the AR-FCM \hat{C} -Exp algorithm, Lin-AR-FCM \hat{C} -Exp, can be introduced to cope with this issue. For each cluster, medoid update is not based on the examination of all I units, but only on a subset of p ($p \ll I$) units with the highest membership degrees in the considered cluster. In Fig. 3 results of a simulation study is reported. The computation time required by the AR-FCM \hat{C} -Exp algorithm is represented by a solid line. The computation time required by its linearized version is represented by a dashed line when $p = I/C$, and by a dotted line when $p = 0.5 \cdot I/C$. For this simulation we have set $C = 2$, $m = 2$ and $RS = 1$. As it can be seen, the computation time of the AR-FCM \hat{C} -Exp rapidly increases as the sample size gets larger. On the contrary, Lin-AR-FCM \hat{C} -Exp is less affected by the sample size, especially with $p = 0.5 \cdot I/C$.

Remark 5. Based on the distance (Eq. (10)) illustrated in Remark 1, the AR-FCM \hat{C} -Exp model, for ARMA(1, 1) processes, can be written as:

$$\left\{ \begin{array}{l} \min : \sum_{i=1}^I \sum_{c=1}^C u_{ic}^m \left[1 - \frac{\exp\left\{-\beta \frac{\hat{\phi}_i - \hat{\theta}_i}{1 - \hat{\theta}_i^2}\right\} \exp\left\{-\beta \frac{\hat{\phi}_c - \hat{\theta}_c}{1 - \hat{\theta}_c^2}\right\}}{\exp\left\{-2\beta \frac{(\hat{\phi}_i - \hat{\theta}_i)(\hat{\phi}_c - \hat{\theta}_c)}{1 - \hat{\theta}_i \hat{\phi}_c}\right\}} \right] \\ \sum_{c=1}^C u_{ic} = 1, \quad u_{ic} \geq 0 \end{array} \right. \tag{16}$$

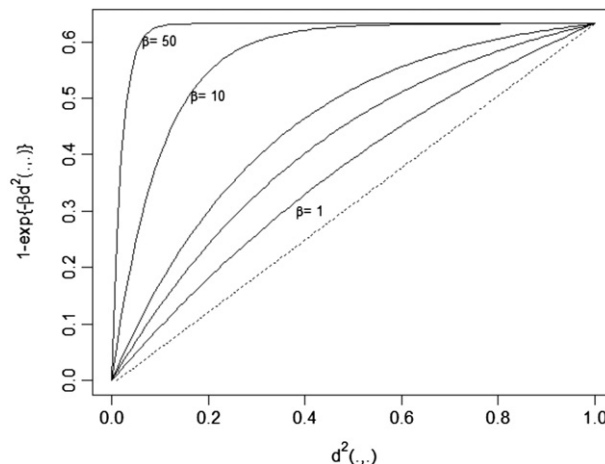


Fig. 1. Effect of β on the distance (6).

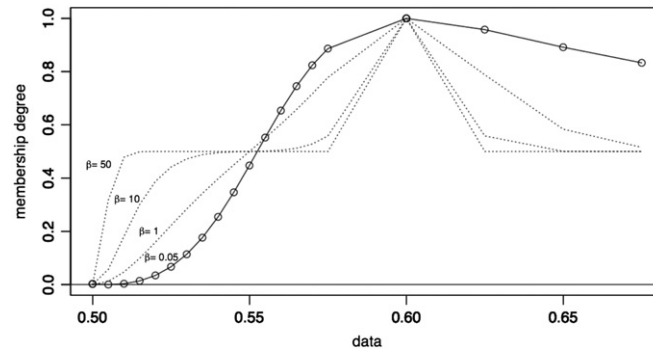


Fig. 2. Effect of the parameter β on the membership degrees (13).

and thus:

$$u_{ic} = \left(\sum_{c'=1}^C \left[\frac{\exp\left\{-\beta \frac{\hat{\phi}_i - \hat{\theta}_i}{1 - \hat{\theta}_i^2}\right\} \exp\left\{-\beta \frac{\hat{\phi}_c - \hat{\theta}_c}{1 - \hat{\theta}_c^2}\right\}}{1 - \frac{\exp\left\{-2\beta \frac{(\hat{\phi}_i - \hat{\theta}_i)(\hat{\phi}_c - \hat{\theta}_c)}{1 - \hat{\theta}_i \hat{\theta}_c}\right\}}{1 - \hat{\theta}_i^2} \exp\left\{-\beta \frac{\hat{\phi}_c - \hat{\theta}_c}{1 - \hat{\theta}_c^2}\right\}} \right]^{\frac{1}{m-1}} \right)^{-1} \quad (17)$$

Remark 6. Based on the weighted AR exponential distance (Eq. (11)) illustrated in remark 2, we can obtain a more general formalization of the clustering model (12):

$$\left\{ \begin{array}{l} \min : \sum_{i=1}^I \sum_{c=1}^C u_{ic}^m \left[1 - \exp\left\{-\beta(\hat{\pi}_i - \hat{\pi}_c)' \Omega(\hat{\pi}_i - \hat{\pi}_c)\right\} \right] \\ \sum_{c=1}^C u_{ic} = 1, \quad u_{ic} \geq 0. \end{array} \right. \quad (18)$$

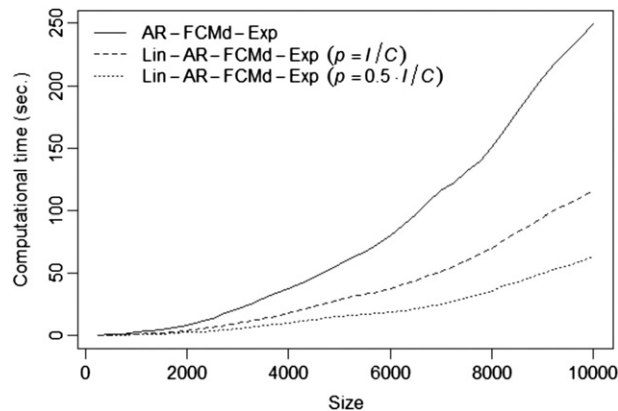


Fig. 3. Computation time for AR-FCMd-Exp algorithm and the linearized Lin-AR-FCMd-Exp algorithm.

and the corresponding solutions:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[\frac{1 - \exp\{-\beta(\hat{\pi}_i - \hat{\pi}_{c'})' \Omega (\hat{\pi}_i - \hat{\pi}_{c'})\}}{1 - \exp\{-\beta(\hat{\pi}_i - \hat{\pi}_{c'c})' \Omega (\hat{\pi}_i - \hat{\pi}_{c'c})\}} \right]^{\frac{1}{m-1}}} \tag{19}$$

Note that with $\Omega = \mathbf{I}$, where \mathbf{I} denotes the identity matrix, we obtain the clustering model (12) and the membership degrees (Eq. (13)). In the following we set $\Omega = \mathbf{I}$.

4. Simulation study

To assess the clustering performance and accuracy of the proposed robust clustering model AR-FCM_DC-Exp, in this section we show the results of a simulation study. The simulation study consisted of the generation of several simulated datasets according to fifteen different scenarios, which mimic real world situations.

For comparison purposes, we have considered the non-robust AR-FCM_DC model [3] and the robust AR-based Fuzzy C-Medoid Clustering with Noise Cluster model (AR-FCM_DC-NC) [2].

We have also drawn a comparison between the fuzzy approach and the crisp approach, by considering the crisp versions of AR-FCM_DC and AR-FCM_DC-NC, i.e. the AR Crisp C-Medoid (AR-CCM_DC) model and the AR Crisp C-Medoids with Noise Cluster (AR-CCM_DC-NC) model.

Finally, we have considered some hierarchical clustering models, viz. Single, Complete and Average linkage and Ward's model with autoregressive coefficients.

For each scenario, we have generated time series of length $T = 256$ from ARMA processes, in particular AR(1) and MA(1) processes. Each generation process has produced well separated clusters of four time series. In addition, in some scenarios data were contaminated with switching and/or outlier time series. Each simulated time series has been fitted with an AR(k) model, where the order k was determined by Akaike's Information Criterion (AIC).

Some suggestive scenarios are illustrated in Table 1, i.e. the scenario 1 with two well separated clusters, and scenarios 2, 5 and 6, contaminated with one outlier time series and/or one switching time series.

For a complete overview of the fifteen scenarios, see Table I in the supplementary material to this paper.

In brief, in scenarios 1 and 3 we have generated two and three well separated clusters, respectively. In scenarios 2 and 4 we have considered the same simulated dataset as in Scenarios 1 and 3, respectively, adding an outlier time series.

In scenarios 5 and 7 there are two well separated clusters and a switching time series. In scenarios 6 and 8 we have added to the former scenarios an outlier time series.

Similarly, in scenarios 9 and 11 we have considered three well separated clusters and a switching time series, adding an outlier time series in scenarios 10 and 12.

In scenario 13 we have generated three well separated clusters with two switching series, while in the scenarios 14 and 15 we have considered the same simulated dataset as in the scenario 13, adding one and two outliers time series, respectively.

The clustering performance and accuracy of each clustering model were evaluated according to whether time series generated from the same process were grouped in the same cluster, with membership degrees equal or close to one in that cluster.

Since switching time series should belong simultaneously to more than one cluster, the performance of each clustering model was also assessed according to the capability of the model to individuate switching time series.

Finally, when the dataset is contaminated with one or more outlier time series, the robustness of each model was evaluated by considering the effect of the presence of anomalous data in the clustering process.

For each scenario, 10 sets of 100 simulations were carried out. For each set of simulation, the percentage of times the objects (time series) were correctly identified as belonging to one of the cluster, or as switching time series, or as outlier time series, is computed. Then, we have computed the average percentage of correct classification over the 10 sets of the 100 simulations for each scenario. The average percentage of correct classification is a measure of clustering accuracy of the model. The higher this value, the better the classification performance of the model considered. We have chosen this strategy to be consistent with that used by D'Urso et al. [2,3].

To assign each time series to a specific cluster we have set cut-off values. In the first four scenarios, with no switching series, we have assigned the i -th time series to the c -th cluster if its fuzzy membership degree was $u_{ic} > 0.7$ or $u_{ic} > 0.5$. In the remaining scenarios contaminated with switching time series, we have set the cut-off value as $u_{ic} > 0.7$ or $u_{ic} > 0.6$, depending on whether there were two or three separated clusters, respectively.

To identify the switching time series, we have set the membership degrees in the interval (0.3, 0.7) in the scenarios with two clusters, and in the interval (0.3, 0.6) in those with three clusters, so as to obtain fuzzy membership degrees across clusters.

Note that the selected cut-off values are compatible with those suggested in literature: for simulation studies, see [28,29,83], while for empirical applications see [84].

As for the identification of outlier time series, one should note that the exponential distance approach does not explicitly label the outlier time series but spreads the membership degrees of outlier series over the k clusters, as mentioned above. Therefore, for AR-FCM_DC-Exp we have also computed the percentage of times in which, providing the remaining time series were correctly identified, the membership degrees of outlier time series were split across clusters.

Finally, in all scenarios we have set the fuzziness parameter m at different values, namely 1.3, 1.5, 1.7 and 2, to check if results are influenced by the degree of fuzziness.

Overall results for clustering performance and accuracy of the model compared are summarized in Table 2.

For each scenario, values in Table 2 are obtained by averaging the percentages of correct classification for the different values of m , of membership degrees and, if any, the different partitions obtained with each model. Note that we do not report the results for the AR-CCM_DC model and for the hierarchical clustering for Scenarios 5–15, since they provided partitions which were never close to the real data structure.

In addition in Table 3 are reported the membership degrees obtained on random generated datasets from the scenarios illustrated in Table 1.

More detailed results of the simulation for each scenario carried out are reported in the supplementary material to this paper (Tables II-XXXI).

In particular, in the even numbered tables in the supplementary material the clustering results for each scenario are reported. Results for crisp models AR-CCM_DC and AR-CCM_DC-NC are reported in the row corresponding to $m = 0$.

In the odd numbered tables in the supplementary material, the membership degrees obtained with AR-FCM_DC, AR-FCM_DC-NC, and with their crisp counterpart, and AR-FCM_DC-Exp for each scenario are reported.

Scenarios 1 and 3 are our baseline scenarios, with no outliers and/or switching time series. As expected, in both cases for all considered

clustering models the average percentage of success in the classification of the generated dataset is always near or equal to 100% (Table 2).

The crisp medoid based-models and the hierarchical procedures perform slightly worse than the fuzzy models in the case of the scenario 3.

Scenarios 2 and 4 are similar to the previous scenarios, with the only difference given by the inclusion of a single outlier time series. As one would expect, the robust fuzzy models on average are able to correctly classify the time series and to neutralize the influence of the outlier. As mentioned above, for AR-FCMdc-Exp we have reported in italics the average percentage of times in which the outlier is split into the clusters. For scenario 2, see also Table 3, where the membership degrees of the outlier time series are highlighted in gray. As it can be seen from the membership degrees reported in Table 3, in scenario 2 AR-FCMdc considered outlier time series as a switching time series. This inconvenience occurs also for scenario 4 (see Table IX in the supplementary material). This evidence could help to explain the very low average percentage of success reported in Table 2 for this method in correspondence with scenarios 2 and 4.

Scenarios 5 and 6 are similar to the first two scenarios, with the addition of a switching time series. In scenario 5 all the fuzzy models deal rather well with the presence of the switching time series. All the fuzzy models are capable of classifying the switching time series in a vague manner, as it can be seen from the membership degrees reported in Table 3. Adding an outlier (scenario 6) the robust fuzzy models still perform well, while AR-FCMdc is effective in classifying the generated time series only when the outlier is considered as switching between clusters (see Table 3).

Similar patterns are observed for scenarios 7 and 8, where the switching time series is more erratic than in the previous two scenarios.

As for the scenario 9, in which we have three separated clusters and a switching time series, all the fuzzy models are able to identify the switching series rather well (see also Table XIX in the supplementary material).

When an outlier is added to the data (scenario 10), the disruptive effect of the anomalous time series heavily influences the average performance of success of the fuzzy model AR-FCMdc.

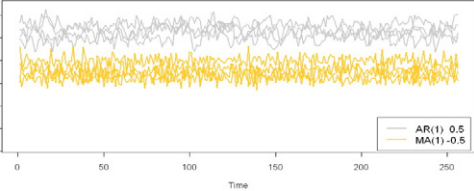
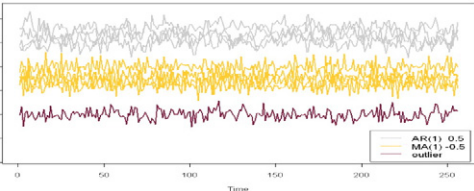
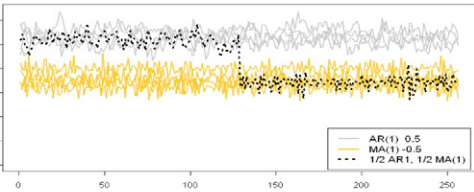
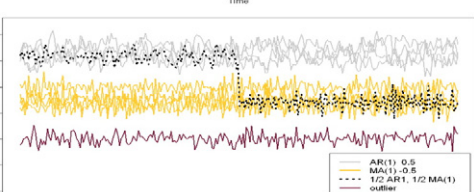
Scenarios 11 and 12 are similar to scenarios 9 and 10, with a switching time series that shifts between clusters more frequently than in the previous scenarios. As one would expect, findings are similar to that observed for scenarios 9 and 10.

Scenarios 13 to 15 consist of three well separated clusters and two switching time series, the first switching between the first and the second cluster, the second between the second and third cluster. In scenarios 14 and 15 one and two outlier time series are added, respectively. The proposed AR-FCMdc-Exp model outperforms the remaining models both when the dataset is not contaminated by outliers (see also Table XXVI in the supplementary material, for scenario 13) and in presence of one or two outliers (see also Tables XXVIII and XXX, for scenarios 14 and 15, respectively).

Overall, as for the capability of a correct identifying of outlier time series, the performances of AR-FCMdc-NC and AR-FCMdc-Exp are comparable in most of the situations taken in consideration in this simulation study. Both models are preferable to AR-FCMdc in presence of outliers.

As a final remark, it has to be noted AR-FCMdc-Exp outperforms AR-FCMdc-NC, especially when the presence of outlier time series are combined with the presence of one or more switching time series and in the scenarios with three well separated clusters.

Table 1
Scenario 1 with two clusters and scenarios 2, 5, and 6 contaminated with one outlier and/or one switching time series.

Scenario	Description	Figure	Models
1	Two well-separated clusters; four time series were simulated from each of AR(1) with $\phi = 0.5$ and MA(1) with $\theta = -0.5$		<ul style="list-style-type: none"> –Fuzzy models: AR-FCMdc, AR-FCMdc-NC, AR-FCMdc-Exp –Crisp models: AR-CCMdc, AR-CCMdc-NC –Hierarchical procedures: Single Linkage, Complete and Average Linkage, Ward's method
2	Two well-separated clusters and outlier time series; four time series were simulated from each of AR(1) with $\phi = 0.5$ and MA(1) with $\theta = -0.5$. The outlier time series was simulated from a process ARMA(1,1) with $\phi = 0.9$ and $\theta = -0.9$		<ul style="list-style-type: none"> –Fuzzy models: AR-FCMdc, AR-FCMdc-NC, AR-FCMdc-Exp –Crisp models: AR-CCMdc, AR-CCMdc-NC –Hierarchical procedures: Single Linkage, Complete and Average Linkage, Ward's method
5	Two well-separated clusters and a switching time series; four time series were simulated from each of AR(1) with $\phi = 0.5$ and MA(1) with $\theta = -0.5$ and a switching time series 1/2 AR(1) and 1/2 MA(1).		<ul style="list-style-type: none"> –Fuzzy models: AR-FCMdc, AR-FCMdc-NC, AR-FCMdc-Exp –Crisp model: AR-CCMdc-NC
6	Two well-separated clusters, a switching time series and an outlier time series; four time series were simulated from each of AR(1) with $\phi = 0.5$ and MA(1) with $\theta = -0.5$ and a switching time series 1/2 AR(1) and 1/2 MA(1). The outlier time series was simulated from a process ARMA(1,1) with $\phi = 0.9$ and $\theta = -0.9$		<ul style="list-style-type: none"> –Fuzzy models: AR-FCMdc, AR-FCMdc-NC, AR-FCMdc-Exp –Crisp model: AR-CCMdc-NC

Up until now we have reported the results of the simulation study in a pure descriptive way. In order to test the improvement of the proposed method over the other methods that have been considered, we have conducted the one-sided Wilcoxon signed rank test. The test verifies the significance of differences between pairs of data of two dependent samples. Let δ be the median of these differences. The null and alternative hypotheses of the test are the following: $H_0 : \delta = 0$; $H_1 : \delta > 0$. The null hypothesis is that the average performances evaluated for each pair of models are not significantly dissimilar. The alternative hypothesis is that the average performance evaluated for the first model in the comparison is significantly better than the one observed for the second model.

We have computed the one-sided Wilcoxon signed rank test for each pair of models by considering the average percentages of success of each model reported in Table 2. Note that we have drawn comparisons of pairs only between Partitioning Around Medoids based models, i.e. AR-FCMdc, AR-FCMdc-NC, AR-FCMdc-Exp and AR-CCMdc-NC.

Table 4 shows, for each pair of models, the values of the test-statistic W and the accompanying p -values.

Let consider two generic models. We reject the null hypothesis that there is no difference in the performance of the two model, concluding that the average performance of the first model is significantly better than that of the other model when the p -value of the test-statistic W is less than the 0.05 significance level.

From Table 4 we can see how the proposed AR-FCMdc-Exp model outperforms the remaining clustering models.

5. Application: the use of the AR-FCMdc-Exp model for clustering air pollution time series

In this section we illustrate an air quality study based on daily nitrogen monoxide (NO) emissions detected in fourteen monitoring stations in Rome: Arenula, Bufalotta, C.so Francia, Castel di Guido, Ciampino, Cinecittà, Cipro, Fermi, L.go Magna Grecia, L.go Perestrello, Malagrotta, Tenuta del Cavaliere, Tiburtina, Villa Ada. We have considered both urban and non urban stations since, given that NO emissions are mainly related to human activities, it is interesting to assess whether there are differences between areas with different population density. Data were collected in 2012, from January 1 December 31. NO concentration derives from different sources, such as heating systems and vehicular traffic.

The data source is the database BRACE,¹ which is maintained by ISPRA (Istituto per la Protezione e la Ricerca Ambientale), the Italian public institution which deals with the environment preservation.

Raw data are transformed by considering the log-differences of the daily emissions of NO, in order to remove non-stationarity both in the levels and in variance. Data are also affected by weekly seasonality, since, as observed above, NO emissions are also related to traffic conditions that are likely to vary during the week: usually, car traffic is more congested during weekdays and less on Sunday. To remove the seasonal pattern, we have regressed data against six daily dummy variables. The aim of this case study is to detect similarities among NO monitoring stations in terms of temporal trends of daily changes of concentration levels.

We have assessed the stationarity of the transformed data by means of the augmented Dickey Fuller (ADF) test. The null hypothesis of the ADF test is that the time series has unit roots, i.e. it is not stationary. To reinforce our findings, we have employed the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. KPSS test is a nonparametric test in which the null hypothesis is the stationarity of time series. We have reported the p -values of both tests in Table 5. Both tests led us to conclude that data are stationary.

Raw data and transformed data are shown in Fig. 4, panel (a) and (b), respectively.

Once the stationarity of data has been assessed, to choose the best ARMA models for the time series of daily rates change of NO emissions in Rome, we considered the Box and Jenkins modeling procedure.

Results for this procedure are reported in Table 6. In brackets are reported the standard errors.

In the last column of Table 5 are reported the Ljung–Box statistics up to ten lags and the corresponding p -values, in brackets. The null hypothesis is that the residual series from the fitted models are white noise and hence the model fit is appropriate. If the p -value is greater than 0.05, we can accept the null hypothesis at a 5% significance level. As it can be seen, the null hypothesis is accepted for all the fitted time series.

We fit each time series with the truncated $AR(\infty)$ representation corresponding to the ARMA model estimated. Then, we apply the proposed AR-FCMdc-Exp model and, for comparison purposes, we also consider the timid robust fuzzy clustering, i.e. AR-FCMdc (Autoregressive-based Fuzzy C-Medoid Clustering) proposed by D'Urso et al. [3] and the robust fuzzy clustering AR-FCMdc-NC (Autoregressive-based Fuzzy C-Medoid Clustering with Noise Cluster) proposed by D'Urso et al. [2].

Since the model proposed is based on the partitioning around medoid approach, once the number of clusters is assessed, we are able to identify medoid stations, i.e., stations representative of the whole cluster. The fuzzy approach allows us also to identify stations whose temporal trends match with more than one cluster. Finally, the robustness of the model allows us to identify anomalous stations, whose profiles are not typical of any cluster. Then, for a quick monitoring of the network, policy makers could look at the medoid station, the fuzzy allocated stations and the outliers. All these stations carry out relevant information about the air quality in different sites.

By adopting the FS criterion, the optimal number of clusters is $C = 3$ for all models. The membership degrees of each station obtained with the three fuzzy AR-based models are reported in Table 7.

First of all, both the robust fuzzy clustering models, AR-FCMdc-NC and AR-FCMdc-Exp, detect the presence of one outlier (C.so Francia, station 3). Hence, results for AR-FCMdc model are likely to be undermined by the presence of one anomalous time series.

With AR-FCMdc the medoid stations are Arenula, Cipro and Tiburtina (respectively stations 1, 7 and 13). We can also observe that most of the stations are assigned to one cluster with very high membership degrees ($u_{ic} > 0.8$), the only exception being Tenuta del Cavaliere (station 12) which is fuzzy allocated between cluster 2 and 3.

The partition obtained with AR-FCMdc-NC is similar to that obtained with AR-FCMdc, the main difference being the fact that C.so Francia is allocated in the noise cluster and, hence, is considered an outlier. In fact, by looking at the ARMA model reported in Table 6, C.so Francia displays a temporal evolution of daily rates of change of NO emission which is at odds with the remaining station.

This evidence is confirmed also by AR-FCMdc-Exp, which splits the membership degrees of C.so Francia uniformly across the cluster. Beside this, the partition obtained is only marginally different from those observed with the previous models.

Considering the values of the $AR(\infty)$ coefficients (π -weights) of the time series (related to the values of the original ARMA coefficients in Table 6), graphically represented via box plot in Fig. 5, we observe that C.so Francia displays the lowest values of the coefficients; then from the lowest values to the highest we find a first group of stations with lower values of the coefficients (Arenula, Cinecittà, L.go Magna Grecia, L.go Perestrello, Tiburtina, Villa Ada) and then a second group of stations with higher values (Bufalotta, Castel di Guido, Ciampino, Cipro, Fermi, Malagrotta, Tenuta del Cavaliere), Bufalotta and Fermi showing the highest values.

All the partitions presented in Table 7 detect these differences, in particular by the AR-FCMdc-Exp model, which smoothes the influence of C.so Francia, allocates stations Bufalotta and Fermi that show the

¹ <http://www.brace.sinanet.apat.it/web/struttura.html>.

Table 2
Average clustering results for scenarios 1–15.

Scenarios	Characteristics of the scenarios			Non-hierarchical models					Hierarchical procedures with AR coefficients ^b			
	Number of:			Fuzzy clustering models			Crisp clustering models		Ward's method	Single linkage	Average linkage	Complete linkage
	Clusters	Outlier Time series	Switching Time Series	AR-FCMdc	AR-FCMdc-NC	AR-FCMdc-Exp ^a	AR-CCMdc	AR-CCMdc-NC				
1	2	0	0	99.84	99.18	100	100	96.20	100	100	100	100
2	2	1	1	33.75	99.18	100	94.28	48.70	96.20	100	100	100
3	3	0	0	95.20	94.91	99.90	99.90	89.30	90.20	98.60	88.40	92.50
4	3	1	0	30.35	96.08	95.41	93.20	27.70	88.50	96.30	88.50	91.50
5	2	0	1	99.53	98.98	99.88	–	–	50.00	–	–	–
6	2	1	1	32.76	96.85	99.58	99.58	–	48.00	–	–	–
7	2	0	1	99.55	99.15	99.88	–	–	50.00	–	–	–
8	2	1	1	32.60	99.70	99.58	99.58	–	50.00	–	–	–
9	3	0	1	88.10	90.30	95.28	–	–	49.90	–	–	–
10	3	1	1	31.76	92.53	97.78	97.78	–	41.30	–	–	–
11	3	0	1	88.80	89.68	93.85	–	–	49.00	–	–	–
12	3	1	1	31.34	91.90	97.58	97.58	–	42.55	–	–	–
13	3	0	2	81.00	85.95	98.03	–	–	23.95	–	–	–
14	3	1	2	20.60	79.65	84.43	84.43	–	21.20	–	–	–
15	3	2	2	23.68	81.10	84.95	65.48	–	23.50	–	–	–

(a): in italics are reported the percentage of cases in which AR-FCMdc-Exp splits the membership degrees of the outlier time series into two or three clusters (according to the scenario).
 (b): in presence of one outlier time series (scenarios 2 and 4), hierarchical methods classify the outlier time series as a singleton in a separated cluster.

highest values of the AR(∞) coefficients to a separate cluster, and assigns stations Cinecittà and Tiburtina to the cluster of the stations with the lowest values.

Adopting AR.FCMdc-Exp, it is interesting to note that cluster 1 is characterized by urban stations, mainly located in residential areas (Arenula, Cinecittà, L.go Magna Grecia, L.go Perestrello), while cluster

Table 3
Membership degrees of random generated datasets from scenarios 1, 2, 5 and 6 (two clusters).

Scenarios	Time series	AR-FCMdc		AR-FCMdc-NC		Noise cluster	AR-FCMdc-Exp	
		Cluster1	Cluster2	Cluster1	Cluster2		Cluster1	Cluster2
		u_{i1}	u_{i2}	u_{i1}	u_{i2}		u_{i1}	u_{i2}
1	1	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	2	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	3	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	4	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	5	0.02 (0)	0.98 (1)	0.00 (0)	0.98 (1)	0.02 (0)	0.00	1.00
	6	0.00 (0)	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	7	0.00 (0)	1.00 (1)	0.00 (0)	0.98 (1)	0.02 (0)	0.01	0.99
	8	0.01 (0)	0.99 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.01	0.99
2 (One outlier time series)	1	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	2	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	3	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00 (0)	0.99	0.00
	4	0.99 (1)	0.01 (0)	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	5	0.00 (0)	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	6	0.02 (0)	0.98 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	7	0.00 (0)	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	8	0.00 (1)	0.99 (1)	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
Outlier 5 (one switching time series)	9	0.61 (1)	0.39 (0)	0.03 (0)	0.01 (0)	0.96 (1)	0.58	0.42
	1	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	2	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	3	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	4	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	5	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	6	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	7	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
Switching 6 (one outlier and one switching time series)	8	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	9	0.50	0.50	0.42 (0)	0.44 (1)	0.14 (0)	0.52	0.48
	1	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	2	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	3	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	4	1.00	0.00	1.00 (1)	0.00 (0)	0.00 (0)	1.00	0.00
	5	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	6	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
Switching outlier	7	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	8	0.00	1.00	0.00 (0)	1.00 (1)	0.00 (0)	0.00	1.00
	9	0.57	0.43	0.51 (1)	0.47 (0)	0.02 (0)	0.56	0.44
	10	0.66	0.34	0.03 (0)	0.05 (0)	0.92 (1)	0.50	0.50

Note: in brackets are reported the membership degrees for the crisp models AR-CCMdc and AR-CCMdc-NC. Bold font indicates the higher membership degree of the time series or the membership degrees for switching time series.

Table 4
Wilcoxon signed rank test directional hypothesis.

	W	p-value
AR-FCMdc-Exp vs. AR-FCMdc-NC	117	0.0007
AR-FCMdc-Exp vs. AR-FCMdc	120	0.0004
AR-FCMdc-Exp vs. AR-CCMdc-NC	120	0.0004
AR-FCMdc-NC vs. AR-FCMdc	110	0.0025
AR-FCMdc-NC vs. AR-CCMdc-NC	120	0.0004
AR-FCMdc vs. AR-CCMdc-NC	63	0.4435

2 is characterized by non-urban stations (Castel di Guido, Ciampino, Malagrotta and Tenuta del Cavaliere).

Overall, AR-FCMdc-Exp and AR-FCMdc-NC provide similar partitions, but the partition obtained with AR-FCMdc-Exp seems to be more interesting from a policy maker's point of view.

Finally, we compare AR-FCMdc-Exp model with some partitioning procedures suggested in the literature on air pollution monitoring which analyze the observed data set directly and not a proper model-based parametric representation of the data: non hierarchical clustering (*k*-means clustering) [34] and hierarchical agglomerative cluster analysis, i.e. Ward [31], single linkage [46,59], average linkage [46,56,57] and complete linkage [42,46,58].

For each method, the optimal partition is detected by means of the silhouette criterion. Results are reported in the last fifth columns of Table 7.

Hierarchical clustering method with single linkage, average linkage and complete linkage fail to identify a substantive partition. Indeed, all stations, but Castel di Guido are allocated into one cluster.

More meaningful results are obtained with hierarchical cluster analysis with Ward's method and with *k*-means. The partitions obtained with these methods are similar, with the only exception of Villa Ada which is allocated differently according to the method considered. Notice that the partitions obtained are of scarce interest from a policy makers point of view. For instance, in cluster 1 grouped stations are located in residential areas (Arenula, L.go Magna Grecia), stations in high traffic areas (C.so Francia, Fermi, Tiburtina) and non urban stations (Ciampino, Tenuta del Cavaliere). Moreover, applying these procedures it is not possible to identify representative stations for each cluster, or the presence of outliers.

Overall, the performances of robust fuzzy models are better than results obtained utilizing standard (non fuzzy) and non robust clustering procedures based on hierarchical and partitioning around centroids (e.g. *k*-means clustering) approaches.

In particular, our robust fuzzy clustering model is more appropriate than standard procedures to analyze air quality data properly, inheriting the features of the theoretical and methodological approaches adopted in the clustering process.

6. Final remarks

In this paper, we have proposed a robust fuzzy clustering model for time series. Furthermore, we have shown its performances by considering a simulation study and its empirical usefulness and effectiveness in environmetrics by applying the AR-FCMdc-Exp model to air pollution time series monitored by a set of stations.

From a theoretical point of view, the advantages of our AR-FCMdc-Exp model are listed below.

1. It takes into account the dynamic information of the time series by means of the most widely used parametric representation of the time series, i.e. their autoregressive representation.
2. It inherits all the advantages of the partitioning around medoid approach. In particular:
 - It does not depend on the order in which the time series are presented, except when equivalent solutions exist, which very rarely occurs in practice (this is not the case for many other algorithms present in literature [80].
 - As opposed to *C*-means clustering and hierarchical clustering approaches, AR-FCMdc-Exp model assigns each observed time series to the cluster represented by one of the selected representative time series. Then each cluster is represented by an observed representative time series and not by a fictitious representative time series (e.g., mean time series). The possibility of obtaining non-fictitious representative time series in the clusters is very appealing and useful in a wide range of applications. This is very important for the interpretation of the selected clusters. In fact, as affirmed by Kaufman and Rousseeuw [80] “in many clustering problems one is particularly interested in a characterization of the clusters by means of typical or representative objects [time series]. These are objects [time series] that represent the various structural aspects of the set of objects [time series] being investigated. There can be many reasons for searching for representative objects [time series]. Not only can these objects [time series] provide a characterization of the clusters, but they can often be used for further work or research, especially

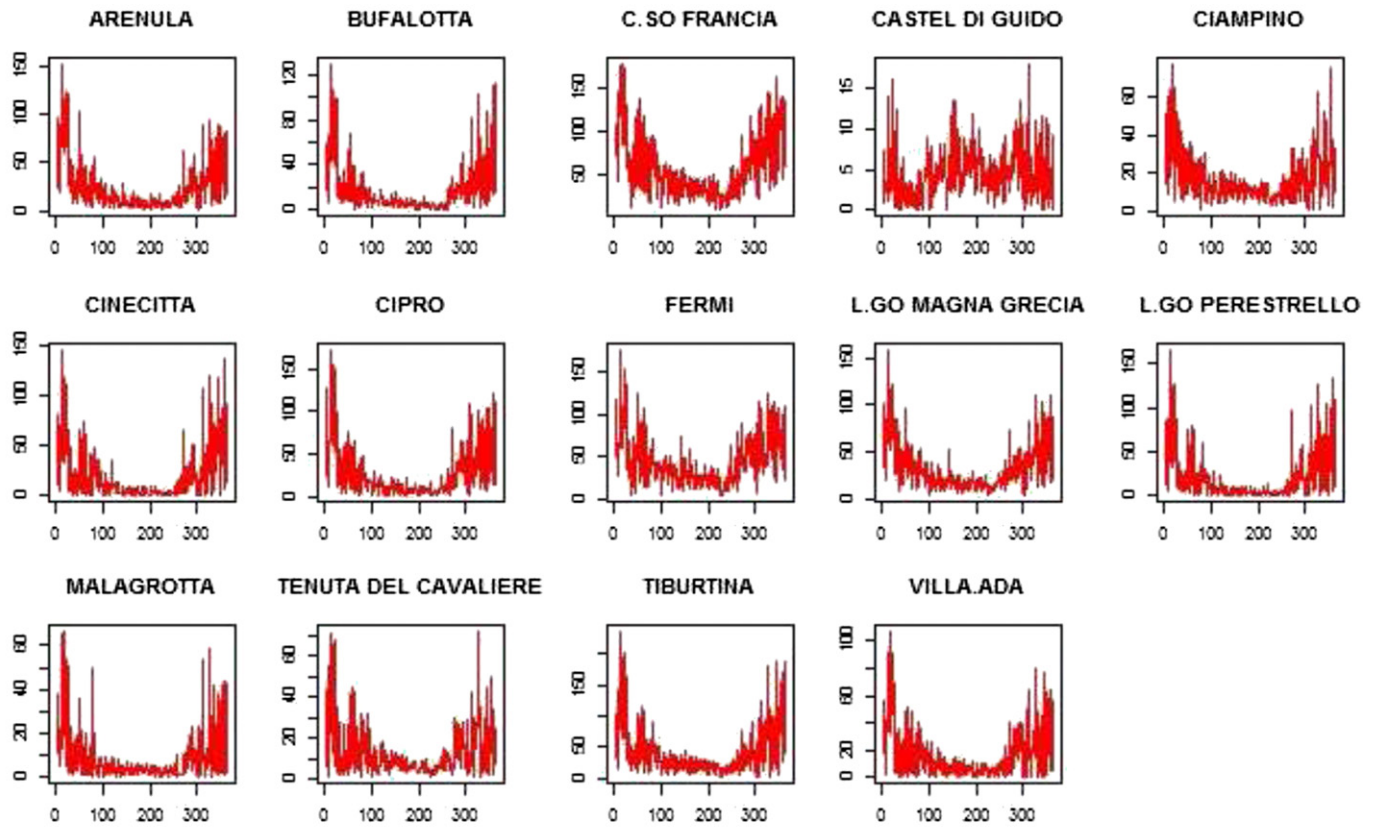
Table 5
Results of the ADF test, the KPSS test and the Ljung–Box test for daily rates change in NO emissions in Rome.

Station	ADF	KPSS	Q(10)
1 ARENULA	−13.809 (<2e−16)	0.008	14.139 (0.078)
2 BUFALOTTA	−14.07 (<2e−16)	0.009	8.482 (0.205)
3 C.SO FRANCIA	−12.505 (<2e−16)	0.009	9.536 (0.146)
4 CASTEL DI GUIDO	−12.814 (<2e−16)	0.009	4.735 (0.578)
5 CIAMPINO	−13.222 (<2e−16)	0.009	11.378 (0.181)
6 CINECITTA	−12.418 (<2e−16)	0.008	4.801 (0.441)
7 CIPRO	−13.559 (<2e−16)	0.007	10.175 (0.253)
8 FERMI	−13.223 (<2e−16)	0.009	7.651 (0.468)
9L. GO MAGNA GRECIA	−11.649 (<2e−16)	0.008	5.890 (0.436)
10L. GO PERESTRELLO	−14.075 (<2e−16)	0.008	3.901 (0.866)
11 MALAGROTTA	−14.345 (<2e−16)	0.009	8.887 (0.352)
12 TENUTA DEL.CAVALIERS	−13.61 (<2e−16)	0.008	4.093 (0.769)
13 TIBURTINA	−13.051 (<2e−16)	0.008	3.078 (0.688)
14 VILLA ADA	−13.182 (<2e−16)	0.008	5.854 (0.664)

Note: in brackets are reported the p-values for the ADF and the Ljung–Box tests.

Critical values for the KPSS test without trend are:
0.347 (10%); 0.463 (5%); 0.574 (2.5%); 0.739 (1%).

(a)



(b)

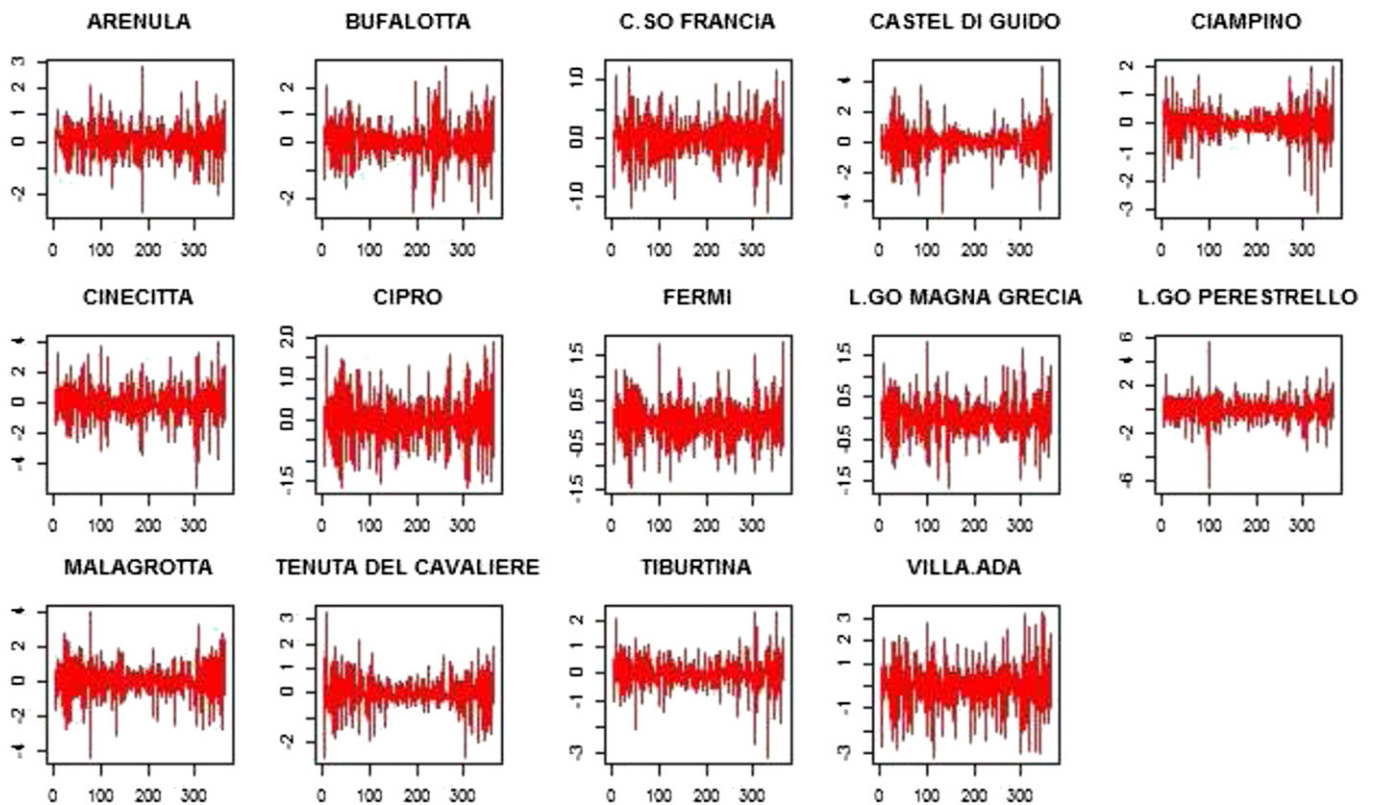


Fig. 4. Daily rates change of NO emissions in Rome: (a) raw data, (b) transformed data.

Table 6
Estimated coefficients of ARMA(*p,q*) processes for daily rates change in NO emissions in Rome.

Station	AR(1)	AR(2)	AR(3)	MA(1)	MA(2)	MA(3)	MA(4)	MA(5)
1 ARENULA	0.446 (0.054)	-	-	-0.966 (0.020)	-	-	-	-
2 BUFALOTTA	-0.480 (0.186)	-	-	0.033 (0.180)	-0.446 (0.101)	-0.317 (0.065)	-	-
3 C.SO FRANZIA	0.546 (0.054)	-0.067 (0.059)	0.138 (0.054)	-0.991 (0.017)	-	-	-	-
4 CASTEL DI GUIDO	-0.644 (0.059)	0.263 (0.058)	-	0.058 (0.025)	-0.915 (0.025)	-	-	-
5 CIAMPINO	-	-	-	-0.535 (0.047)	-0.369 (0.046)	-	-	-
6 CINECITTA	-	-	-	-0.525 (0.052)	-0.282 (0.058)	-0.061 (0.061)	0.081 (0.064)	-0.135 (0.054)
7 CIPRO	-	-	-	-0.546 (0.048)	-0.333 (0.050)	-	-	-
8 FERMI	0.386 (0.076)	-	-	-0.893 (0.045)	-	-	-	-
9 L.GO MAGNA GRECIA	-0.569 (0.058)	0.373 (0.058)	-	0.030 (0.040)	-0.965 (0.040)	-	-	-
10 L.GO PERESTRELLO	0.415 (0.057)	-	-	-0.961 (0.023)	-	-	-	-
11 MALAGROTTA	-	-	-	-0.62 (0.048)	-0.277 (0.047)	-	-	-
12 TENUTA DEL.CAVALIERE	0.183 (0.122)	-	-	-0.583 (0.119)	-0.336 (0.099)	-	-	-
13 TIBURTINA	-	-	-	-0.432 (0.052)	-0.314 (0.056)	-0.144 (0.060)	0.101 (0.058)	-0.135 (0.051)
14 VILLA ADA	0.426 (0.053)	-	-	-0.965 (0.018)	-	-	-	-

Note: in brackets are reported the standard errors of the estimated coefficients of the ARMA(*p,q*).

when it is more economical or convenient to use a small set of *k* objects [*C* time series in our case] instead of the large set one started off with”.

- It is more robust than *C*-means clustering method and Euclidean-based hierarchical clustering in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a centroid. Then, the objective chosen in our clustering approach is appealing because it is more robust than the error sum of squares employed in most methods [80,85,86]. Notice that, as stated by Garcia-Escudero and Gordaliza [87,88] the methods based on the partitioning around medoids procedure provide only a timid robustification of the *C*-means clustering method; then they alleviate the negative effects of the presence of outliers in the dataset but do not solve the problem. However, as we point out below (point 4), our model is robust (and not slightly robust) because in addition we consider a robust metric in the clustering process.
- 3. It inherits the computational properties and the other advantages of the fuzzy approach. For instance, it captures the switching or drifting nature of some time series in the clustering process.
- 4. It is robust, i.e. it is able to tolerate the presence of outlier time series in the clustering partition. In fact, by using a suitable robust metric in the clustering process, the AR-FCMdC-Exp model is able to neutralize and smooth the disruptive effect of outlier time series, preserving the original clustering structure of the dataset by assigning to outliers almost the same membership degree to clusters.

From an empirical point of view, in particular in air pollution monitoring studies, the benefits connected to the utilization of AR-FCMdC-Exp model are shown in the following.

1. Using the autoregressive representation of the air pollution time series, we are able to analyze a wide class of environmental time series. In fact, in the literature, this assumption is often made for air pollution time series (see, e.g., [2,3,89–91]).
2. In the literature on air quality monitoring, in many papers the uncertainty is not properly quantified in the clustering process. However, various researchers note the importance of defying a suitable measure of uncertainty. As pointed out by Pastres et al. [33] this issue could be relevant when deciding to redesign a monitoring network, since the location of ‘uncertain’ sites may be considered when establishing the boundaries between homogeneous areas. By adopting a fuzzy approach for clustering the air pollution time series monitored by a set of stations, our model is able to quantify the uncertainty -formalized in a fuzzy manner- in the assignment process of the monitoring stations to the clusters by means of the membership degrees.
3. By adopting the partitioning around medoids approach, we have a set of prototype (medoid) air pollution monitoring stations, representing for each cluster the monitoring stations with high membership degrees to the same cluster. In this way, we can obtain useful information on the possible redundancy/efficiency of the air pollution

Table 7
Membership degrees for the fuzzy clustering models and indices of cluster membership for crisp methods.

Station no.	AR-FMdC			AR-FMdC-NC				AR-FMdC-Exp			Crisp	Hierarchical clustering ^a			
	Cluster 1 with medoid	Cluster 2 with medoid	Cluster 3 with medoid	Cluster 1 with medoid	Cluster 2 with medoid	Cluster 3 with medoid	Noise cluster	Cluster 1 with medoid	Cluster 2 with medoid 5	Cluster 3 with Medoid 8	<i>k</i> -means ^a	Ward's method	Single linkage	Average linkage	Complete linkage
1	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.989	0.007	0.003	1	1	1	1	1
2	0.023	0.907	0.071	0.013	0.575	0.050	0.361	0.002	0.003	0.996	1	1	1	1	1
3	0.535	0.131	0.334	0.016	0.004	0.009	0.970	0.334	0.333	0.333	1	1	1	1	1
4	0.018	0.917	0.065	0.016	0.803	0.116	0.065	0.042	0.915	0.042	2	2	2	2	2
5	0.028	0.852	0.120	0.023	0.669	0.254	0.054	0.000	1.000	0.000	1	1	1	1	1
6	0.018	0.003	0.979	0.000	0.000	1.000	0.000	0.795	0.185	0.020	3	3	1	1	1
7	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.042	0.779	0.179	1	1	1	1	1
8	0.019	0.923	0.058	0.012	0.649	0.048	0.290	0.000	0.000	1.000	1	1	1	1	1
9	0.881	0.010	0.109	0.768	0.008	0.074	0.150	0.750	0.147	0.104	1	1	1	1	1
10	0.873	0.002	0.126	0.872	0.001	0.122	0.005	1.000	0.000	0.000	3	3	1	1	1
11	0.006	0.975	0.019	0.006	0.928	0.042	0.025	0.005	0.990	0.005	2	2	1	1	1
12	0.041	0.576	0.383	0.028	0.408	0.509	0.056	0.051	0.921	0.029	1	1	1	1	1
13	0.000	0.000	1.000	0.020	0.002	0.974	0.004	0.864	0.111	0.025	1	1	1	1	1
14	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.992	0.006	0.003	3	2	1	1	1

Note: Bold font indicates maximal membership degrees of the time series. Italic font indicates the membership degree for switching time series. (a): in the column the cluster indices are reported.

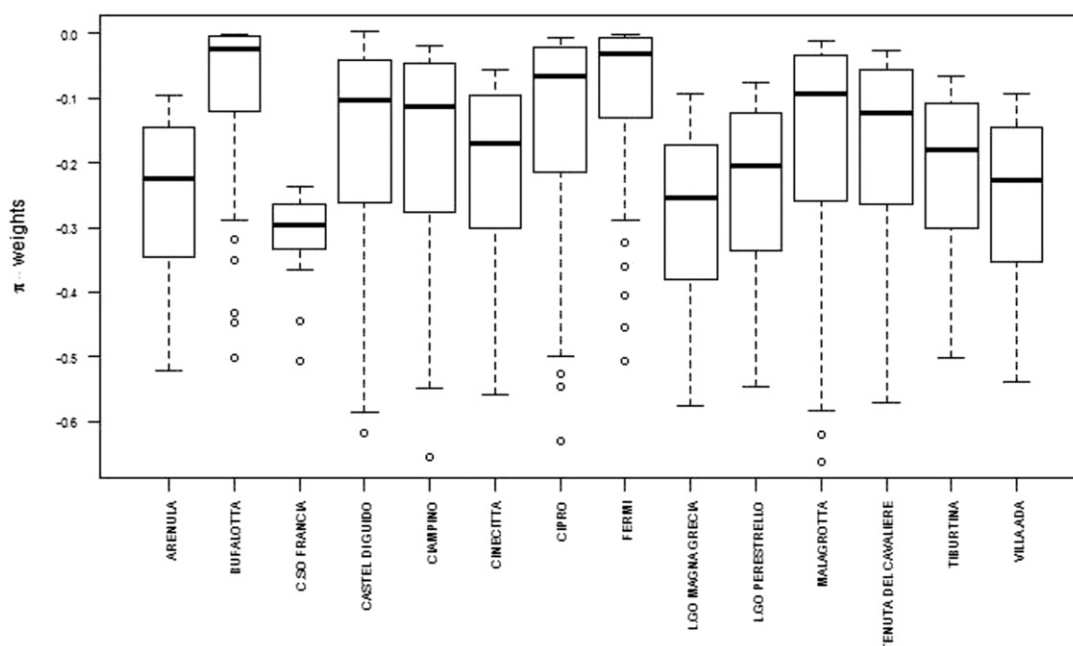


Fig. 5. Box plots of the parameter of the AR(k) processes of daily changes of NO emissions.

monitoring network. In fact, as stated by Ignaccolo et al. [32], the partitioning around medoids approach “provides representative objects for each cluster which gives policy makers real sites to look at in order to quickly monitor general trends in a region”. Furthermore, “while k -means algorithm provides centroids that show concentration levels and temporal trends for obtained clusters, PAM [Partition Around Medoids] is based on the search for k representative objects (called *medoids*) in the dataset, and groups are defined around them. Representative objects are useful for policy makers since they are an efficient support for identifying summaries in their regional reports.” [32].

- Since our clustering model is robust, it is able to tolerate the presence of outlier air pollution time series in the clustering process. In this way, we can neutralize the disruptive effects in the clustering process of monitoring stations characterized by anomalous dynamic behaviors of the air pollution time series. In this way, the optimal partition of the monitoring stations is not affected by anomalous dynamic behavior of some air pollution time series.

In future, we will investigate possible theoretical developments of our clustering model and its usefulness and effectiveness in other chemometrical areas.

As a final remark, since publicly accessible resources (like data analysis packages or user friendly web-servers [92,93]) represent the future direction for developing more useful classifiers, models, or predictors, we are currently working on an R package that will be freely downloadable.

Conflict of Interest

The authors certify that there is no conflict of interest.

Acknowledgments

The authors thank the editor and the two referees for their useful comments and suggestions which helped to improve the quality and presentation of this manuscript. The authors also thank Dario Di Lallo.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.chemolab.2014.11.003>.

References

- T.W. Liao, Clustering of time series data—a survey, *Pattern Recogn.* 38 (2005) 1857–1874.
- P. D'Urso, L. De Giovanni, R. Massari, D. Di Lallo, Noise fuzzy clustering of time series by autoregressive metric, *METRON* 71 (2013) 217–243.
- P. D'Urso, D. Di Lallo, E.A. Maharaj, Autoregressive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks, *Soft. Comput.* 17 (2013) 83–131.
- Y.S. Ding, T.L. Zhang, Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network, *Protein Pept. Lett.* 14 (2007) 811–815.
- H.B. Shen, J. Yang, K.C. Chou, Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition, *J. Theor. Biol.* 240 (2006) 9–13.
- H.B. Shen, J. Yang, X.J. Liu, Using supervised fuzzy clustering to predict protein structural classes, *Biochem. Biophys. Res. Commun.* 334 (2005) 577–581.
- P. Wang, X. Xiao, NR-2L: A two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features, *PLoS ONE* 6 (2011) e23505.
- X. Xiao, J.L. Min, P. Wang, iGPCR-drug: A web server for predicting interaction between GPCRs and drugs in cellular networking, *PLoS ONE* 8 (2013) e72234.
- X. Xiao, J.L. Min, P. Wang, iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints, *J. Theor. Biol.* 337C (2013) 71–79.
- X. Xiao, P. Wang, GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions, *Mol. BioSyst.* 7 (2011) 911–919.
- X. Xiao, P. Wang, W.Z. Lin, iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, *Anal. Biochem.* 436 (2013) 168–177.
- T.L. Zhang, Y.S. Ding, Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern, *J. Theor. Biol.* 250 (2008) 186–193.
- D.N. Georgiou, T.E. Karakasidis, A.C. Megaritis, A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory, *Open Bioinforma. J.* 7 (2013) 41–48 (open access at <http://www.benthamscience.com/open/tobioij/articles/V007/SI0025TOBIOIJ/0041TOBIOIJ.pdf>).
- D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, *J. Theor. Biol.* 257 (2009) 17–26.
- M. Hayat, A. Khan, Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC, *Protein Pept. Lett.* 19 (2012) 411–421.
- H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels, *Biomed. Res. Int.* 2014 (2014).

- [17] Y.N. Fan, X. Xiao, J.L. Min, iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking, *Int. J. Mol. Sci.* 15 (2014) 4915–4937.
- [18] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, *Bioinformatics* 30 (2014) 1522–1529.
- [19] W.R. Qiu, X. Xiao, W.Z. Lin, iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, *Biomed. Res. Int.* 2014 (2014).
- [20] W. Chen, P.M. Feng, E.Z. Deng, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, *Anal. Biochem.* 462 (2014) 76–83.
- [21] Y. Xu, J. Ding, L.Y. Wu, iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, *PLoS ONE* 8 (2013) e55844.
- [22] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, iDNA-ProtDis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, *PLoS ONE* 9 (2014) e106691.
- [23] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), *J. Theor. Biol.* 273 (2011) 236–247.
- [24] J. Caiado, E.A. Maharaj, P. D'Urso, Time series clustering, in: C. Hennig, M. Meila, F. Murtagh, R. Rocci (Eds.), *Handbook of Cluster Analysis*, Chapman & Hall, 2014 (in press).
- [25] P. D'Urso, Dissimilarity measures for time trajectories, *Stat. Meth. Appl.* 1–3 (2000) 53–83.
- [26] P. D'Urso, Fuzzy clustering for data time arrays with inlier and outlier time trajectories, *IEEE Trans. Fuzzy Syst.* 13 (2005) 583–604.
- [27] J. Caiado, N. Crato, D. Peña, A periodogram-based metric for time series classification, *Comput. Stat. Data Anal.* 50 (2006) 2668–2684.
- [28] P. D'Urso, E.A. Maharaj, Autocorrelation-based fuzzy clustering of time series, *Fuzzy Sets Syst.* 160 (2009) 3565–3589.
- [29] E.A. Maharaj, P. D'Urso, Fuzzy clustering of time series in the frequency domain, *Inf. Sci.* 181 (2011) 1187–1211.
- [30] P. D'Urso, E.A. Maharaj, Wavelet-based clustering of multivariate time series, *Fuzzy Sets Syst.* 193 (2012) 33–61.
- [31] P. D'Urso, L. De Giovanni, E.A. Maharaj, R. Massari, Wavelet-based self organizing maps for classifying multivariate time series, *J. Chemom.* 28 (2014) 28–51.
- [32] R. Ignaccolo, S. Ghigo, E. Giovenali, Analysis of monitoring networks by functional clustering, *Environmetrics* 62 (2008) 672–686.
- [33] R. Pastres, A. Pastore, S.F. Tonellato, Looking for similar patterns among monitoring stations. Venice Lagoon application, *Environmetrics* 22 (2011) 712–724.
- [34] M.L. Sanchez Gomez, M.C. Ramos Martin, Application of cluster analysis to identify sources of airborne particles, *Atmos. Environ.* 21 (1987) 1521–1527.
- [35] M. Bohm, B. McCune, T. Vandetta, Diurnal curves of tropospheric ozone in the western United States, *Atmos. Environ.* 25 (1991) 1570–1590.
- [36] M.L. Sanchez, D. Pascual, C. Ramos, I. Perez, Forecasting particulate pollutant concentrations in a city from meteorological variables and regional weather patterns, *Atmos. Environ.* 26 (1990) 1509–1519.
- [37] S.R. Dorling, T.D. Davies, C.E. Pierce, Cluster analysis: a technique for estimating the synoptic meteorological controls on air and precipitation chemistry – method and applications, *Atmos. Environ.* 26 (1992) 2575–2581.
- [38] W. Ruijgrok, F.G. Romer, Aspects of wet, acidifying deposition in Arnhem: source regions, correlations, and trends, *Atmos. Environ.* 27 (1993) 637–653.
- [39] J. Miranda, T.A. Cahill, J. Morales, A.F. Roberto, M.J. Flores, R.V. Diaz, Determination of elemental concentrations in atmospheric aerosols in Mexico City using proton induced x-ray emission, proton elastic scattering, and laser absorption, *Atmos. Environ.* 28 (1994) 2299–2306.
- [40] C.M. Romo-Groger, J.R. Morales, M.I. Dinator, F. Llona, Heavy metals in the atmosphere coming from a copper smelter in Chile, *Atmos. Environ.* 28 (1994) 705–711.
- [41] F.L. Ludwig, J. Jiang, J. Chen, Classification of ozone and heather patterns associated with high ozone concentrations in the San Francisco and Monterey Bay areas, *Atmos. Environ.* 29 (1995) 2915–2928.
- [42] C. Lavecchia, E. Angelino, M. Bedogni, E. Brevetti, R. Gualdi, G. Lanzani, A. Musitelli, M. Valentini, The ozone patterns in the aerological basin of Milan (Italy), *Environ. Softw.* 11 (1996) 73–80.
- [43] V. Wongphatarakul, S.K. Friedlander, J.P. Pinto, A comparative study of PM2.5 ambient aerosol chemical databases, *Environ. Sci. Technol.* 32 (1998) 3926–3934.
- [44] A. Ionescu, Y. Candau, E. Mayer, I. Colda, Analytical determination and classification of pollutant concentration fields using air pollution monitoring network data—methodology and application in the Paris area, during episodes with peak nitrogen dioxide levels, *Environ. Model Softw.* 15 (2000) 565–573.
- [45] I.G. Kavouras, P. Koutrakis, M. Tsapakis, E. Lagoudaki, E.G. Stephanou, D. Von Baer, P. Oyola, Source apportionment of urban particulate aliphatic and polynuclear aromatic hydrocarbons (PAHs) using multivariate methods, *Environ. Sci. Technol.* 35 (2001) 2288–2294.
- [46] S. Saksena, V. Joshi, R.S. Patil, Cluster analysis of Delhi's ambient air quality data, *J. Environ. Monit.* 5 (2003) 91–499.
- [47] C. Silva, A. Quiroz, Optimization of the atmospheric pollution monitoring network at Santiago de Chile, *Atmos. Environ.* 37 (2003) 2337–2345.
- [48] V. Gabusi, M. Volta, A methodology for seasonal photochemical model simulation assessment, *J. Environ. Pollut.* 24 (2005) 11–21.
- [49] S. Beaver, A. Palazoglu, A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area, *Atmos. Environ.* 40 (2006) 713–725.
- [50] E. Gramsh, F. Cereceda-Balic, P. Oyola, D. von Baer, Examination of pollution trends in Santiago de Chile with cluster analysis of PM10 and Ozone data, *Atmos. Environ.* 40 (2006) 5464–5475.
- [51] A.C. Comrie, An all-season synoptic climatology of air pollution in the US–Mexico border region, *Prof. Geogr.* 48 (1996) 237–251.
- [52] B.F. Taubman, J.C. Hains, A.M. Thompson, L.T. Marufu, B.G. Doddridge, J.W. Stehr, C.A. Piety, R.R. Dickerson, Aircraft vertical profiles of trace gas and aerosol pollution over the mid-Atlantic United States: statistics and meteorological cluster analysis, *J. Geophys. Res.* 111 (2006) D10S07.
- [53] I. Morlini, Searching for structure in measurements of air pollutant concentration, *Environmetrics* 18 (2007) 823–840.
- [54] T. Bengtsson, J.E. Cavanaugh, State-space discrimination and clustering of atmospheric time series data based on Kullback information measures, *Environmetrics* 19 (2008) 103–121.
- [55] S.B. Kim, C. Temiyasathit, V.C.P. Chen, S.K. Park, M. Sattler, A.G. Russell, Characterization of spatially homogeneous regions based on temporal patterns of fine particulate matter in the continental United States, *J. Air Waste Manage. Assoc.* 58 (2008) 965–975.
- [56] J.C.M. Pires, S.I.V. Sousa, M.C. Pereira, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis—part I: SO2 and PM10, *Atmos. Environ.* 42 (2008) 1249–1260.
- [57] J.C.M. Pires, S.I.V. Sousa, M.C. Pereira, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis—part II: CO, NO2 and O3, *Atmos. Environ.* 42 (2008) 1261–1274.
- [58] J. Lau, W.T. Hung, C.S. Cheung, Interpretation of air quality in relation to monitoring station's surroundings, *Atmos. Environ.* 43 (2009) 769–777.
- [59] G. Ibarra-Berastegi, J. Sáenz, A. Ezcurra, U. Ganzedo, J.D. de Argandoña, I. Errasti, A. Fernandez-Ferrero, J. Polanco-Ferrero, Assessing spatial variability of SO2 field as detected by an air quality network using self-organizing maps, cluster, and principal component analysis, *Atmos. Environ.* 43 (2009) 3829–3836.
- [60] S. Pakalapati, S. Beaver, J.A. Romagnoli, A. Palazoglu, Sequencing diurnal air flow patterns for ozone exposure assessment around Houston, Texas, *Atmos. Environ.* 43 (2009) 715–723.
- [61] F. Karaca, F. Camci, Distant source contributions to PM10 profile evaluated by SOM based cluster analysis of air mass trajectory sets, *Atmos. Environ.* 44 (2010) 892–899.
- [62] W.-Z. Lu, H.-D. He, L.-Y. Dong, Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis, *Build. Environ.* 46 (2011) 577–583.
- [63] J. Adame, A. Notario, F. Villanueva, J. Albaladejo, Application of cluster analysis to surface ozone, NO2 and SO2 daily patterns in an industrial area in Central-Southern Spain measured with a DOAS system, *Sci. Total Environ.* 429 (2012) 281–291.
- [64] J. Flemming, R. Stern, R. Yamartino, A new air quality regime classification scheme for O3, NO2, SO2 and PM10 observations sites, *Atmos. Environ.* 39 (2005) 6121–6129.
- [65] E. Austin, B.A. Coull, D. Thomas, P. Koutrakis, A framework for identifying distinct multipollutant profiles in air pollution data, *Environ. Int.* 45 (2012) 112–121.
- [66] E. Austin, B.A. Coull, A. Zanolletti, P. Koutrakis, A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition, *Environ. Int.* 59 (2013) 244–254.
- [67] M.A.E. Chaparro, J.M. Lavornia, M.A.E. Chaparro, A.M. Sinito, Biomonitoring of urban air pollution: magnetic studies and SEM observations of corticolous foliose and microfoliose lichens and their suitability for magnetic monitoring, *Environ. Pollut.* 172 (2013) 61–69.
- [68] R. Ignaccolo, S. Ghigo, S. Bande, Functional zoning for air quality, *Environ. Ecol. Stat.* 20 (2013) 109–127.
- [69] C. Abraham, P.A. Cornillon, E. Matzner-Løber, N. Molinari, Unsupervised curve clustering using Bsplines, *Scand. J. Stat.* 30 (2003) 581–595.
- [70] M.A. Elangasinghe, N. Singhal, K.N. Dirks, J.A. Salmond, S. Samarasinghe, Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering, *Atmos. Environ.* 94 (2014) 106–116.
- [71] C.S. Malley, C.F. Braban, M.R. Heal, The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification, *Atmos. Res.* 138 (2014) 30–40.
- [72] K.B. Ensor, B.K. Ray, S.J. Charlton, Point source influence on observed extreme pollution levels in a monitoring network, *Atmos. Environ.* 92 (2014) 191–198.
- [73] M. Corduas, D. Piccolo, Time series clustering and classification by the autoregressive metric, *Comput. Stat. Data Anal.* 52 (2008) 1860–1872.
- [74] D. Piccolo, A distance measure for classifying ARIMA models, *J. Time Ser. Anal.* 11 (1990) 153–164.
- [75] K.-L. Wu, M.-S. Yang, Alternative c-means clustering algorithms, *Pattern Recogn.* 35 (2002) 2267–2278.
- [76] D.-Q. Zhang, S.-C. Chen, A comment on “Alternative c-means clustering algorithms”, *Pattern Recogn.* 37 (2004) 173–174.
- [77] D. Piccolo, Statistical issues on the AR metric in time series analysis, *Proceedings of the SIS intermediate conference: conference on risk and prediction*, 2007, pp. 221–232.
- [78] T.A. Runkler, J.C. Bezdek, Alternating cluster estimation: a new tool for clustering and function approximation, *IEEE Trans. Fuzzy Syst.* 7 (1999) 377–393.
- [79] R.J.G.B. Campello, E.R. Hruschka, A fuzzy extension of the silhouette width criterion for cluster analysis, *Fuzzy Sets Syst.* 157 (2006) 2858–2875.
- [80] L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, J. Wiley and Sons, New York, 1990.
- [81] T. Kamdar, A. Joshi, On Creating Adaptive Web Servers Using Weblog Mining, Technical Report TR-CS-00-05, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, 2000.
- [82] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low-complexity fuzzy relational clustering algorithms for web mining, *IEEE Trans. Fuzzy Syst.* 9 (2001) 595–607.

- [83] E.A. Maharaj, P. D'Urso, D.U.A. Galagedera, Wavelets-based fuzzy clustering of time series, *J. Classif.* 27 (2010) 231–275.
- [84] D. Dembélé, P. Kastner, Fuzzy C-means method for clustering microarray data, *Bioinformatics* 19 (2003) 973–980.
- [85] S. Mitra, An evolutionary rough partitive clustering, *Pattern Recogn. Lett.* 25 (2004) 1439–1449.
- [86] M.C.N. Barioni, H.L. Razente, A.J.M. Traina, C. Traina, Accelerating k-medoid-based algorithms through metric access methods, *J. Syst. Softw.* 8 (2008) 343–355.
- [87] L.A. García-Escudero, A. Gordaliza, Robustness properties of k-means and trimmed k-means, *J. Am. Stat. Assoc.* 94 (1999) 956–969.
- [88] L.A. García-Escudero, A. Gordaliza, A proposal for robust curve clustering, *J. Classif.* 22 (2005) 185–201.
- [89] P. Anttila, J.-P. Tuovinen, Trends of primary and secondary pollutant concentrations in Finland in 1994–2007, *Atmos. Environ.* 44 (2010) 30–41.
- [90] S. Hassanzadeh, F. Hosseinibalam, R. Alizadeh, Statistical models and time series forecasting of sulfur dioxide: a case study Tehran, *Environ. Monit. Assess.* 155 (2009) 149–155.
- [91] U. Kumar, V.K. Jain, ARIMA forecasting of ambient air pollutants (O₃, NO, NO₂ and CO), *Stoch. Env. Res. Risk A.* 24 (2010) 751–760.
- [92] H.B. Shen, Review: recent advances in developing web-servers for predicting protein attributes, *Nat. Sci.* 2 (2009) 63–92, <http://dx.doi.org/10.4236/ns.2009.12011> (open access at).
- [93] S.X. Lin, J. Lapointe, Theoretical and experimental biology in one, *J. Biomed. Sci. Eng. (JBISE)* 6 (2013) 435–442.