Spatial Statistics 65 (2025) 100874

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Fuzzy clustering of mixed data with spatial regularization

Pierpaolo D'Urso [a], Livia De Giovanni [b,*], Lorenzo Federico [b], Vincenzina Vitale [a]

[a] Department of Social Sciences and Economics, Sapienza - University of Rome, P.le Aldo Moro, 5-00185, Rome, Italy
[b] Department of Political Sciences and Data Lab, Luiss University - Viale Romania, 32-00197, Rome, Italy

ARTICLE INFO

Keywords:
Mixed data
Fuzzy C-medoids clustering
Distance measure
Attribute weighting system
Contiguity matrix

ABSTRACT

A fuzzy clustering model for data with mixed features and spatial constraints is proposed. The clustering model allows different types of variables, or attributes, to be taken into account. This result is achieved by combining the dissimilarity measures for each attribute employing a weighting scheme, to obtain a distance measure for multiple attributes. The weights are objectively computed during the optimization process. The weights reflect the relevance of each attribute type in the clustering results. A spatial term is taken into account, considering a wide definition of contiguity, either physical contiguity or the adjacency matrix in a network. Simulation studies and two empirical applications, including both physical and abstract definitions of contiguity are presented that show the effectiveness of the proposed clustering model.

1. Introduction

Datasets may contain information not embedded into numeric variables or attributes. For instance, socio-economic data often come in a variety of variables, some quantitative (education, wage, labour experience, etc.), some qualitative (gender, marital status, employment status, etc.). In the case of longitudinal socio-economic datasets, among quantitative information, some are time-invariant, at least for a given period (*e.g.*, years of education, household size), and others vary over time (wage, labour experience); also qualitative information could vary over time, especially if units are observed for a long period (*e.g.*, marital status, employment status), yielding ordered sequences of items. Recently, the importance of processing spatial data with mixed type attributes has become more prominent with the wide availability of geographical remote sensing data, for example to predict landslide susceptibility (Ado et al., 2022) or detect different types of crops (Abdali et al., 2023). Often these datasets include visual data, quantitative topographic data and qualitative data on the composition of the soil. Hence, the necessity of applying clustering algorithm to data with mixed attributes, or mixed data. When more than one attribute type is collected, ignoring one or more of them in the clustering process could hamper final results. Most clustering algorithms deal with one of these data types. A first approach to deal with mixed variables consists of a pre-processing to render all variables of the same type either all numeric or all categorical (Guha et al., 1999). A second approach consists of using a dissimilarity measure that can handle mixed data, possibly by assigning a weighting system to address the relevance of each attribute type (Gower, 1971). In this paper the second approach is considered in a fuzzy framework (see, *e.g.* Antoni et al. (2014)). Tables 1 and 2 in D'Urso and Massari (2019) report clustering methods and an admittedly non-exhaustive list of papers that cope with the presence of mixed data. Mixed data in fuzzy clustering models have been considered also in D'Urso et al. (2023b).

* Corresponding author.
E-mail addresses: pierpaolo.durso@uniroma1.it (P. D'Urso), ldegiovanni@luiss.it (L. De Giovanni), lfederico@luiss.it (L. Federico), vincenzina.vitale@uniroma1.it (V. Vitale).

https://doi.org/10.1016/j.spasta.2024.100874
Received 20 August 2024; Received in revised form 5 October 2024; Accepted 13 November 2024
Available online 23 November 2024
2211-6753/

Several clustering techniques for spatial units have been proposed in the literature. The approach followed in this work belongs to the broad group of spatially constrained clustering techniques (Hu and Sung, 2006; Ambroise and Dang, 2009; Viroli, 2011; Torabi, 2016). The models include a spatial penalization term in the objective function. The role of this term and of the related tuning parameter is to smooth the membership degrees of all units contiguous to the generic ith unit in all clusters to which ith unit does not belong. Spatial constraints in fuzzy clustering models have been considered, either (D'Urso et al., 2019, 2022, 2023a).

The main purpose of the present paper is to fill, to our knowledge, a gap by presenting a clustering model for mixed data with spatial regularization. The characteristics of the proposed model are:

mixed data: the proposed clustering model is capable of handling mixed data by combining the dissimilarity measures for each attribute by means of a weighting scheme, so as to obtain a distance measure for multiple attributes;

clustering procedure: adopting the PAM (Partitioning Around Medoids) approach, the cluster prototypes (*i.e.*, medoids) are units actually observed and not "virtual" units like the "centroids" derived with a fuzzy c-means (Bezdek, 1981). Overall, having non-fictitious representative units available makes interpreting the obtained clusters easier (Kaufman and Rousseeuw, 2005). In addition, PAM procedure provides a "timid robustification" of the c-means clustering (García-Escudero and Gordaliza, 1999; García-Escudero et al., 2010);

fuzziness: fuzzy clustering appears more attractive than the traditional clustering methods when it is difficult to identify a clear boundary among clusters (McBratney and Moore, 1985; Wedel and Kamakura, 2000). In addition, the memberships indicate whether there is a second-best cluster almost as good as the best cluster, a scenario which standard clustering methods cannot uncover (Everitt et al., 2011). Furthermore, fuzzy clustering is attractive because it is easily compatible with distribution free methods (Hwang et al., 2007) and it is computationally efficient (McBratney and Moore, 1985; Heiser and Groenen, 1997). For more details, see D'Urso (2015);

spatial information: the proposed clustering model is capable of taking into account the spatial information through a spatial penalty term defined based on the following assumption: "...when a spatial unit belongs to a cluster with a high membership degree, then the penalty term forces the neighbouring spatial units to have high membership degrees in the cluster, as much as possible. In other words, it is expected that a spatial unit with high (low) membership degree in a cluster, will have neighbouring areas with low (high) membership degrees in the remaining clusters. It follows that the spatial penalty term attempts to determine a spatially smoothed membership degrees under the empirical evidence that neighbouring spatial units are often characterized by approximately similar features. Nonetheless, it may also occur that neighbouring spatial units are described by pretty diverse profiles. In this respect, there is a parameter which plays the role of increasing or decreasing the emphasis of the spatial penalty term in the clustering process" (Coppi et al., 2010).

The paper is structured as follows. In Section 2 the model FCMd-MD-SP is presented. In Section 3 a simulation study is described. In Section 4 two applications, one to environmental data of Italian municipalities and the other to Italian accounts of political coalitions in the European elections 2024 are considered.

2. Fuzzy C-medoids clustering for mixed data model with spatial constraints (FCMd-MD-SP model)

Let $\mathcal{X} = \{X_1, \dots, X_P\}$ be a set of P variables, or attributes, observed on n units, in which the P variables are of different types (mixed data), *e.g.*, quantitative, nominal, time series, sequences of qualitative data, imprecisely observed data, textual data.

More precisely, the set $\mathcal{X}$ contains S types of variables, with p_s variables for each attribute type, with

$$s = 1, \dots, S; \quad 1 < S \le P; \quad 1 \le p_s < P; \quad \sum_{s=1}^{S} p_s = P.$$

Without loss of generality, the variables are arranged so that the first p_1 variables are of the same type (for instance, quantitative), the second p_2 variables are also of the same type, different from that of the first p_1 variables (for instance, qualitative), and so on, so that

$$\mathcal{X} \equiv \{\mathcal{X}_1, \dots, \mathcal{X}_s, \dots, \mathcal{X}_S\}$$

where $\mathcal{X}_s \equiv \{X_{p_1+\cdots+p_{s-1}+1,\dots,} X_{p_1+\cdots+p_s}\}$ is the set of variables of the sth type. Finally, $\mathcal{X}_{is}$ is the set of values observed for the ith unit on the p_s variables of the sth type.

As an example, suppose that $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2\}$ where $\mathcal{X}_1$ is a set of two quantitative variables, while $\mathcal{X}_2$ is a set of two qualitative variables. Then, $S = 2$, $p_1 = p_2 = 2, P = 4$, and $\mathcal{X}_1 = \{X_1, X_2\}$, $\mathcal{X}_2 = \{X_3, X_4\}$.

Depending on the nature of the attribute, $\mathcal{X}_{is}$ could be a vector, a matrix, or could have a more complicated structure. For instance, in the case of quantitative variables, $\mathcal{X}_{is} \equiv \mathbf{x}_{is}$ is the vector of p_s values observed on the ith unit. In the case of time series of length T, $\mathcal{X}_{is} \equiv \mathbf{X}_{is}$ is a $T \times p_s$ matrix whose columns are represented by the p_s time series observed on the ith unit, and the rows are the values observed at time t ($t = 1, \dots, T$). In the case of ordered sequences of qualitative items $\mathcal{X}_{is}$ is a set of p_s sequences (see D'Urso and Massari (2013)). Similarly for a set of p_s time series of different lengths.

Continuing with the example, $\mathcal{X}_{i1} = \mathbf{x}_{i1} \equiv \{(x_{i1}, x_{i2}) : i = 1, \dots, n\}$, $\mathcal{X}_{i2} = \mathbf{x}_{i2} \equiv \{(x_{i3}, x_{i4}) : i = 1, \dots, n\}$, where (x_{i1}, x_{i2}) are numeric values, (x_{i3}, x_{i4}) are categorical values.

The distance between units i and i' computed according to the nature of the sth variable type — on this, see Remark 2 below — can be formalized as:

$$_s d_{ii'} = d(\mathcal{X}_{is}, \mathcal{X}_{i's}). \tag{1}$$

Then

$$d^2_{ii'} = \sum_{s=1}^{S} (w_s \cdot {}_s d_{ii'})^2 = \sum_{s=1}^{S} \left[w_s \cdot d(\mathcal{X}_{is}, \mathcal{X}_{i's}) \right]^2 \tag{2}$$

is the overall weighted squared distance considering the S attribute types. As observed by Everitt (1988), the weights of the squared distance are in a quadratic form. As explained in Deza and Deza (2009, Section 4.2), as long as every ${}_s d, s = 1, \dots, S$ is a valid distance over the s−th attribute space, d is a valid distance over the product of all the attribute spaces. The role of the weights will be discussed at large in Remark 3.

In our example, ${}_1 d_{ii'} = d(\mathcal{X}_{i1}, \mathcal{X}_{i'1})$, ${}_2 d_{ii'} = d(\mathcal{X}_{i2}, \mathcal{X}_{i'2})$ are the matrices of the pairwise distances—say, Euclidean distance for $\mathcal{X}_1$ and overlapping distance for $\mathcal{X}_2$, respectively. Then

$$d^2_{ii'} = (w_1 \cdot {}_1 d_{ii'})^2 + (w_2 \cdot {}_2 d_{ii'})^2.$$

When dealing with spatial data, the within-group dispersion has to be minimized and the spatial autocorrelation between contiguous spatial units has to be factored in. In the literature, there are different ways of defining neighbourhood and consequently there are different ways of constructing proximity matrices among spatial units (Páez and Scott, 2005). Two of the most common definitions are based on connectivity, *i.e.* travel time or distance between pairs of units, and physical contiguity. A wide definition of contiguity may also be adopted, as represented by the adjacency matrix in a network. Contiguity can be specified in several ways, for instance, two spatial units can be contiguous: if they are adjacent (neighbours); if they belong to the same macro-area, even if they are not adjacent. In both cases, a binary index $0 - 1$ can be created where 1 is assigned to contiguous spatial units, 0 otherwise. Different definitions of connectivity and contiguity can be embedded in the clustering procedure.

In this paper, the fuzzy Partitioning-Around-Medoids (PAM) algorithm, also known as Fuzzy C-Medoids (FCMd), is adopted thanks to its great advantage of obtaining non-fictitious representative medoids as the final result. This allows for more appealing and easy to interpret results of the final partition (Kaufman and Rousseeuw, 2005). From a computational perspective, fuzzy clustering algorithms are generally more efficient and they are less affected by both local optima and convergence problems (Everitt et al., 2011; Hwang et al., 2007).

Once the formal notation and the overall distance have been described, in the following the clustering algorithm can be illustrated. Following the PAM approach in a fuzzy framework, let $\widetilde{\mathcal{X}}_s \equiv \{\widetilde{\mathcal{X}}_{1s}, \dots, \widetilde{\mathcal{X}}_{cs}, \dots, \widetilde{\mathcal{X}}_{Cs}\}$ be a subset of $\mathcal{X}_s$ with cardinality C, and $\widetilde{\mathcal{X}}_{cs} \in \widetilde{\mathcal{X}}_s$ the values observed for the cth elements of $\widetilde{\mathcal{X}}_s$. Then, $\widetilde{\mathcal{X}}_s \equiv \{\widetilde{\mathcal{X}}_{1s}, \dots, \widetilde{\mathcal{X}}_{cs}, \dots, \widetilde{\mathcal{X}}_{Cs}\}$ is a subset of $\mathcal{X}$ with cardinality C. Let $\mathbf{A}$ be the $(n \times n)$ contiguity (adjacency) matrix.

Formally, the proposed clustering model, called Fuzzy C-Medoids Clustering of Mixed Data model and spatial constraints (FCMd-MD-SP model) is characterized in the following way:

$$\begin{cases} \min: & \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m d_{ic}^2 + \frac{\gamma}{2} \displaystyle\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m \sum_{i'=1}^{n} \sum_{c' \in C_c} a_{ii'} u_{i'c'}^m \\ & = \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{S} (w_s \cdot {}_s d_{ic})^2 + \frac{\gamma}{2} \displaystyle\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m \sum_{i'=1}^{n} \sum_{c' \in C_c} a_{ii'} u_{i'c'}^m \\ & = \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{S} \left[w_s \cdot d(\mathcal{X}_{is}, \widetilde{\mathcal{X}}_{cs}) \right]^2 + \frac{\gamma}{2} \displaystyle\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic} \sum_{i'=1}^{n} \sum_{c' \in C_c} a_{ii'} u_{i'c'}^m \\ \text{(s.t.)} & \sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0 \\ & \sum_{s=1}^{S} w_s = 1,\ w_s \geq 0 \end{cases} \tag{3}$$

where:

- u_{ic} indicates the membership degree of the ith objects to the cth cluster;
- $m > 1$ is a weighting exponent that controls the fuzziness of the obtained partition;
- $\widetilde{\mathcal{X}}_{cs}$ is the sth component of th cth medoid, related to the sth variable type;
- ${}_s d_{ic} = d(\mathcal{X}_{is}, \widetilde{\mathcal{X}_{cs}})$ denotes the distance between the ith observation and the cth medoid, according to the sth variable type; for comparison's sake across attribute types, the S distances ${}_s d_{ic}$ are normalized to vary in the range [0, 1];
- $d_{ic}^2 = \sum_{s=1}^{S} [w_s \cdot d(\mathcal{X}_{is}, \widetilde{\mathcal{X}}_{cs})]^2$ is the overall weighted squared distance between unit i and the medoid c based on all variable types;
- w_s is the weight associated to the sth attribute type, and, hence, to the sth distance ($s = 1, \dots, S$).
- $\frac{\gamma}{2} \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} \sum_{i'=1}^{I} \sum_{c' \in C_c} a_{ii'} u_{i'c'}^m$ is the spatial penalty term;
- $\gamma \geq 0$ is the tuning parameter of the spatial information (spatial coefficient);
- $a_{ii'}$ is the generic element of the $(n \times n)$ "contiguity" matrix $\mathbf{A}$; C_c is the set of the C clusters, with the exclusion of cluster c.

For each spatial unit i and each cluster c, the higher the membership of i to c, the more the sum of the membership degrees of the contiguous/neighbouring spatial units (as indicated in matrix $\mathbf{A}$) in all the clusters except cluster c (summarized C_c) is optimized to be as small as possible. We can observe that the spatial coefficient γ tunes the trade-off between internal cohesion based on the feature vectors and the spatial homogeneity of the clusters. For $\gamma = 0$ the spatial regularization is not taken into account.

The weights w_s constitute specific parameters to be estimated within the clustering procedure.

Proposition 1. *Beginning new equations*

The solutions of (3) are:

$$u_{ic} = \frac{\left[\sum_{s=1}^{S}(w_s \cdot {}_s d_{ic})^2 + \gamma \sum_{i'=1}^{n} \sum_{c' \in C_c} a_{ii'} u_{i'c'}^m\right]^{-\frac{1}{m-1}}}{\sum_{c'=1}^{C}\left[\sum_{s=1}^{S}(w_s \cdot {}_s d_{ic'})^2 + \gamma \sum_{i'=1}^{n} \sum_{c'' \in C_{c'}} a_{ii'} u_{i'c''}^m\right]^{-\frac{1}{m-1}}} \tag{4}$$

$$w_s = \frac{1}{\sum_{s'=1}^{S}\left[\frac{\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \cdot {}_s d_{ic}^2}{\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \cdot {}_{s'} d_{ic}^2}\right]}. \tag{5}$$

Proof. In the following, we derive the iterative solutions (4)–(5).

First, fixing w_s, we determine the membership degrees u_{ic}. We consider the Lagrangian function:

$$L_m(\mathbf{u}_i, \lambda) = \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{S}(w_s \cdot {}_s d_{ic})^2 + \frac{\gamma}{2}\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \sum_{i'=1}^{n}\sum_{c' \in C_c} a_{ii'} u_{i'c'}^m - \lambda\left(\sum_{c=1}^{C} u_{ic} - 1\right) \tag{6}$$

where $\mathbf{u}_i = (u_{i1}, \ldots, u_{ic}, \ldots, u_{iC})'$ and λ is the Lagrange multiplier. Therefore, we set the first derivatives of (6) with respect to u_{ic} and λ equal to zero, yielding:

$$\frac{\partial L_m(\mathbf{u}_i, \lambda)}{\partial u_{ic}} = 0 \Leftrightarrow m u_{ic}^{m-1}\left[\sum_{s=1}^{S}(w_s \cdot {}_s d_{ic})^2 + \gamma \sum_{i'=1}^{n}\sum_{c' \in C_c} a_{ii'} u_{i'c'}^m\right] - \lambda = 0 \tag{7}$$

$$\frac{\partial L_m(\mathbf{u}_i, \lambda)}{\partial \lambda} = 0 \Leftrightarrow \sum_{c=1}^{C} u_{ic} - 1 = 0 \tag{8}$$

We define

$$\theta_{ic} = \sum_{s=1}^{S}(w_s \cdot {}_s d_{ic})^2 + \gamma \sum_{i'=1}^{n}\sum_{c' \in C_c} a_{ii'} u_{i'c'}^m. \tag{9}$$

From (7) we obtain:

$$u_{ic} = \left(\frac{\lambda}{m}\frac{1}{\theta_{ic}}\right)^{\frac{1}{m-1}} \tag{10}$$

and, by considering (8):

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \left(\frac{1}{\sum_{c=1}^{C}\theta_{ic}^{-\frac{1}{m-1}}}\right). \tag{11}$$

Finally, substituting (9) and (11) in (10) we obtain u_{ic} as in (4).

End new equations

Then, fixing u_{ic} we derive w_s. The Lagrangian function is:

$$L_m(\mathbf{w}, \xi) = \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{S}(w_s \cdot {}_s d_{ic})^2 - \xi\left(\sum_{s=1}^{S}(w_s - 1)\right) \tag{12}$$

where $\mathbf{w} = (w_1, \ldots, w_s, \ldots, w_S)'$ and ξ is the Lagrange multiplier. By setting the first derivatives of (12) with respect to w_s and ξ equal to zero, we obtain respectively:

$$\frac{\partial L_m(\mathbf{w}, \xi)}{\partial w_s} = 0 \Leftrightarrow 2 w_s \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \cdot {}_s d_{ic}^2 - \xi = 0 \tag{13}$$

$$\frac{\partial L_m(\mathbf{w}, \xi)}{\partial \xi} = 0 \Leftrightarrow \sum_{s=1}^{S} w_s - 1 = 0. \tag{14}$$

From (13) we have:

$$w_s = \frac{\xi}{2\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \cdot {}_s d_{ic}^2} \tag{15}$$

and using (14):

$$\frac{\xi}{2} = \frac{1}{\sum_{s=1}^{S}\left(\frac{1}{\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \cdot {}_s d_{ic}^2}\right)}. \tag{16}$$

Then, replacing (16) in (15), we obtain w_s, as in (5). □

2.1. Some general remarks

Remark 1 (*Algorithm and Computational Issues*).

1. The fuzzy clustering algorithm that minimizes (3) is built by adopting an estimation strategy based on Fu's heuristic algorithm (Fu and Albus, 1977). Indeed, the alternating optimization estimation procedure cannot be adopted because the necessary conditions cannot be derived by differentiating the objective function in (3) with respect to the medoids. The fuzzy clustering procedure is illustrated in Algorithm 1.

Algorithm 1 Fuzzy C-Medoids Clustering for Mixed Data and SPatial constraints (FCMd-MDSP) algorithm

Fix C and $max.iter$;
Set $iter = 0$;
Pick initial medoids: $\tilde{\mathcal{X}}_s \equiv \{\tilde{\mathcal{X}}_{1s}, \dots, \tilde{\mathcal{X}}_{Cs}\}$, $s = 1, \dots, S$;
repeat
  Store the current medoids $\tilde{\mathcal{X}}_{OLD,s} = \tilde{\mathcal{X}}_s$, $s = 1, \dots, S$;
  Compute $\mathbf{u}_i$ $(i = 1, \dots, n)$ by using (4);
  Compute $\mathbf{w}$ by using (5);
  Select the new medoids: $\tilde{\mathcal{X}}_{cs}$, $c = 1, \dots, C, s = 1, \dots, S$:
  for $c = 1$ to C **do**
    $q = \arg\min_{1 \le i' \le n} \sum_{i''=1}^{n} u_{i''c}^{m} \sum_{s=1}^{S} (w_s \cdot {}_s d_{i',i''})^2 + \frac{\gamma}{2} \sum\limits_{i=1}^{n} \sum\limits_{c=1}^{C} u_{ic}^{m} \sum\limits_{i'=1}^{n} \sum\limits_{c' \in C_c} a_{ii'} u_{i'c'}^{m}$
    return $\Rightarrow \tilde{\mathcal{X}}_{cs} = \mathcal{X}_{qs}$
  end for
  $iter \leftarrow iter + 1$;
until $\tilde{\mathcal{X}}_{OLD,s} = \tilde{\mathcal{X}}_s$, $s = 1, \dots, S$ or $iter = max.iter$

2. The computational complexity of the algorithm is due to four components: (i) the computation of the S dissimilarity matrices for each attribute type; (ii) the exhaustive search for the medoids; (iii) the computation of the penalty term, (iv) the computation of the attribute weights. While it is difficult to deal with the latter issue, it is possible to cope with the former three. First, the PAM approach requires that the distance matrix is computed only once at the beginning of the clustering process, and not at each iteration, thus decreasing the computing time required. Secondly, the search for the optimal medoids can be accelerated by "linearizing" the clustering process, as in Krishnapuram et al. (2001). In the medoids selection phase, for each cluster the search is restricted to the n' $(n' < n)$ objects with the highest membership degrees with that cluster, where n' is selected to be smaller than the average number of units in each cluster. $n' \le n/c$. In this way, the overall complexity is linear in the number of units.
3. The degree of fuzziness of the resulting clusters is determined by m. The parameter can be pre-estimated by considering the usual fuzzy cluster-validity indices (see D'Urso and Maharaj (2009)). However, since the medoid always has a membership of one in the cluster, raising its membership to the power of m has no effect on the medoid, while all other memberships decrease to 0. Thus, when m is high, the mobility of the medoids from iteration to iteration may be lost. For this reason, a value of m between 1 and 1.5 is recommended (Krishnapuram et al., 2001).

Remark 2 (*Distances and Dissimilarities*).

One crucial decision in the clustering process for mixed data is the choice of a suitable distance, or dissimilarity, measure for each attribute type. The choice is mainly heuristic, based on the data at hand and on the peculiar properties of each distance measure.

An admittedly non-exhaustive list of possible distance measures for several attribute types is reported in Table 2 in D'Urso et al. (2023b).

It should be highlighted that the proposed model is adaptable to any kind of dissimilarity measure, leaving to the user the choice of the measures that are better suited for the data at hand.

Remark 3 (*Weighting System*). By means of the weighting system (5) we take into account the relevance of different attribute types towards the clustering process. An attribute type which displays a good separation into different groups should play a more significant role in clustering of data objects, against all other attribute types (Yeung and Wang, 2002; Ahmad and Dey, 2007). Indeed, the weight w_s measures the total intra-cluster deviance, *i.e.*, the within clusters similarity, for variables of the sth type; it increases as long as the intra-cluster deviance for the sth variable type decreases—compared with the remaining variable types. Thus, the optimization procedure gives more relevance to the variable types capable of increasing the within-cluster similarity among the units. In this sense, the proposed weighting scheme is able to provide an objective solution to the balance between different attributes, without requiring user-specified weights.

If one or more attributes have negligible weights, then it is likely that these attributes can be excluded from the analysis causing little, if any, differences in the final results.

Remark 4 (*Determining the Optimal Number of Clusters*). A widely used cluster validity criterion for selecting C is the Xie–Beni criterion (Xie and Beni, 1991), the ratio between compactness and separation among clusters, which can be suitably adapted for FCMd-MD-SP as follows:

$$\min_{C \in \Omega_C} : I_{XB} = \frac{\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m d_{ic}^2}{n \cdot \min_{c,c'} d_{cc'}^2} \tag{17}$$

where Ω_C represents the set of possible values of C $(C < n)$, and $d(.)$ is the overall weighted distance (2). The smaller I_{XB}, the more compact and separate the clusters.

The numerator of I_{XB} represents the total within-cluster distance. The ratio J/n measures the compactness of the fuzzy partition. The smaller this ratio, the more compact a partition with a given number of clusters. Therefore, letting the number of clusters vary over the set Ω_C, the optimal number of clusters is identified in correspondence with the lowest value of I_{XB}.

Remark 5 (*Comparison of Partitions*). Since the fuzzy nature of the partition obtained, the Fuzzy Rand Index FRI (Hüllermeier et al., 2012) is adopted to compare different partitions and/or to compare a given partition with a reference one. FRI is a fuzzy extension of the Rand index based on agreements and disagreements in the two partitions, and it ranges from 0 (total disagreement) to 1 (complete agreement).

3. Simulation study

The simulation study aims to highlight three main features of the FCMd-MD-SP algorithm, the capability of correctly clustering objects; the capability to find a suitable weighting of the attribute types according to their contribution to the optimal clustering results, the capability to take into account a contiguity matrix.

A dataset of $n = 90$ objects, with two numeric variables, X_1, X_2 and three categorical variables, X_3, X_4, X_5 $(S = 2)$ were generated. In particular, X_1 and X_2 were both generated from the Uniform distribution. X_3 is a binary variable, X_4 and X_5 are polytomous variables, with three and four categories respectively. Then, the set of variables is:

$$\mathcal{X} = \{X_1, X_2, X_3, X_4, X_5\} = \{\mathcal{X}_1, \mathcal{X}_2\}$$

where

$$\mathcal{X}_1 = \{X_1, X_2\}, \qquad \mathcal{X}_2 = \{X_3, X_4, X_5\}.$$

Three simulation scenarios were considered:

1. according to the numeric variables there is not a clear clustering structure (Fig. 1(a)).
   On the contrary, objects are grouped into three well-separated and equal-sized clusters according to the categorical variables. By looking at the distribution of the categories in Fig. 1(b), it can be seen that almost always in each cluster the same category is selected for each categorical variable;
2. objects are grouped into three well-separated and equal-sized clusters according to both numeric variables (Fig. 1(c)) and categorical (Fig. 1(d));
3. objects are grouped into three well-separated and equal-sized clusters according to the numeric variables (Fig. 1(e)). For the categorical variables, objects are grouped into three overlapping clusters, as it can be seen from the distribution of the categories in Fig. 1(f), for each variable and for each cluster.

Three adjacency matrices were generated, $\mathbf{P}_1$ concordant with the "separated" variables, either numerical or categorical, $\mathbf{P}_2$ as a stochastic block model with three blocks of size 30 each and edge probabilities equal to 0.4 within the blocks and 0.1 between the blocks (`pm<-cbind(c(.4,0.1,0.1), c(0.1,0.4,0.1), c(0.1,0.1,0.4)) sample_sbm(90, pref.matrix = pm, block.sizes = c(30,30,30)))` and $\mathbf{P}_3$ generated according to the Erdos-Renyi model in which all edges are present independently with equal probability 0.1 (Gilbert, 1959) (erdos.renyi.game(90,0.1,type="gnp")) (Fig. 2). A value of γ ranging from 0.01 to 0.1 was used.

The clustering algorithm should weigh more the categorical variables in the first scenario, and the numeric variables in the third scenario, while it should give approximately the same weight to the two attributes in the second scenario. Given the weighting structure, FCMd-MD-SP should be able to correctly group the objects, even though one attribute does not present a clear clustering structure.

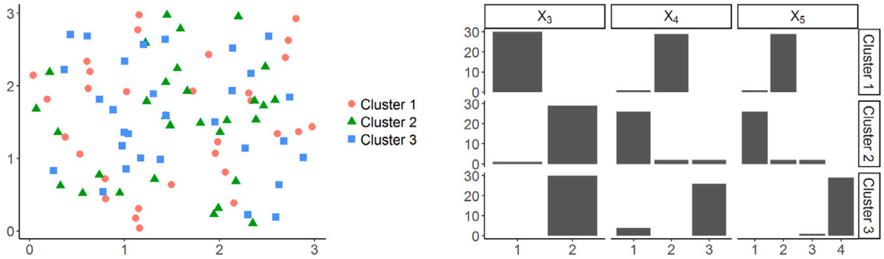
The clustering algorithm should take into account the adjacency matrix.

The correctness of the clustering is evaluated by employing the Fuzzy Rand Index to compare the obtained fuzzy partition with the reference crisp partition (30 objects in each cluster).

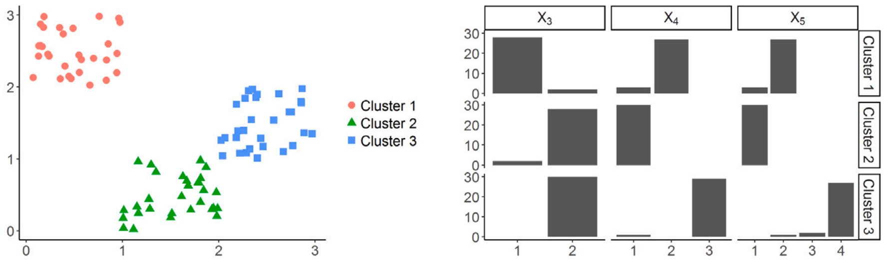
The FCMd-MD-SP model features the expected performances in the presence of adjacency matrices increasingly inconsistent with the values of the attributes, even with a small value of γ. Increasing the value of γ the performances with adjacency matrices $\mathbf{P}_2$ and $\mathbf{P}_3$ decrease (Table 1).

The model FCMd-MD-SP model weighs more the categorical variables in the first scenario, the numeric variables in the third scenario, gives approximately the same weight to the two attributes in the second scenario (Table 2).

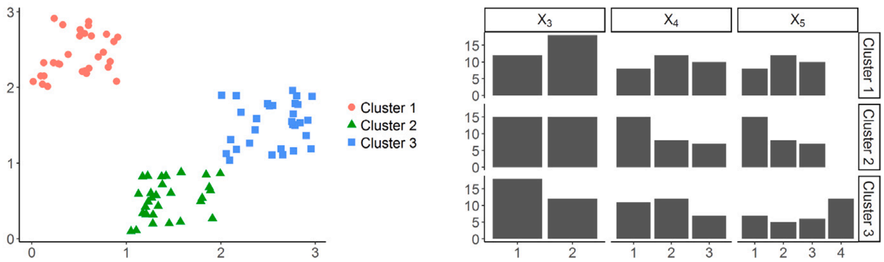
The simulation study has shown the capability of the FCMd-MD- SP algorithm to cluster correctly objects, to find endogenous weights of the attribute types according to their contribution to the optimal clustering results, to make the clustering depend on the contiguity matrix.



(a) Scenario 1: numeric variables (b) Scenario 1: categorical variables



(c) Scenario 2: numeric variables (d) Scenario 2: categorical variables



(e) Scenario 3: numeric variables (f) Scenario 3: categorical variables

Fig. 1. Simulated data - Simulation study 1.

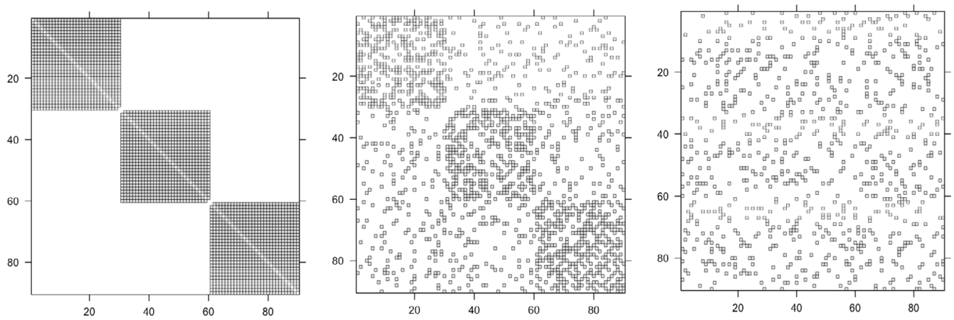


Fig. 2. Three adjacency matrices from left $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$.

4. Empirical applications

The aim of the applications is to show the performances of the FCMd-MD-SP model on environmental data with physical contiguity and on social network data where contiguity is represented by the adjacency matrix of the network.

To identify the fuzzy units, a membership degree in the interval (0.3, 0.7) in the case with two clusters and in the interval (0.3, 0.6) in that with three clusters is set, so as to obtain fuzzy membership degrees across clusters (D'Urso et al., 2014 and references

Table 1
Fuzzy Rand Index.

Scenario		FCMd-MD	FCMd-MD-SP $\gamma = 0.01$			FCMd-MD-SP $\gamma = 0.1$		
			P$_1$	**P**$_2$	**P**$_3$	**P**$_1$	**P**$_2$	**P**$_3$
1	FRI	0.79	0.91	0.81	0.00	1.00	0.00	0.00
2	FRI	1.00	1.00	0.87	0.80	1.00	0.03	0.02
3	FRI	0.94	0.99	0.88	0.80	1.00	0.00	0.00

Table 2
Weights of the continuous variables (w_N) and of the categorical variables (w_C) - $\gamma = 0.01$.

Scenario	FCMd-MD		FCMd-MD-SP					
			P$_1$		**P**$_2$		**P**$_3$	
	w_N	w_C	w_N	w_C	w_N	w_C	w_N	w_C
1	0.10	0.90	0.13	0.87	0.11	0.89	0.83	0.17
2	0.59	0.41	0.58	0.42	0.57	0.43	0.63	0.37
3	0.96	0.04	0.97	0.03	0.96	0.04	0.95	0.05

therein). Detection of cluster membership in more than one cluster is possible by fuzzy clustering, hence giving it a distinct advantage over hard clustering because of this additional information that is gained.

4.1. Environmental data in municipalities

The Survey of Environmental Data in Cities, carried out annually by Istat (National Institute of Statistics) since 2000, is a census survey covering eight themes: Water, Air, Eco-management, Energy, Urban mobility, Urban waste, Noise and Urban green. The universe of respondents consists of the 109 municipalities that are provincial capitals or metropolitan cities, to which the Municipality of Cesena has been added, on a voluntary basis, since the 2020 edition. The data are collected at the municipal level and make it possible to analyse, in their different components, both the quality of the environment and environmental services in urban areas (following their evolution over time) and the environmental policies of the local administrations. The survey is included in the National Statistical Programme (code IST-00907) and envisages the obligation to respond.

The theme of Urban waste was considered, for the year 2022.

For data on the quantity of municipal waste produced and collected separately (by product fraction) the data source is the Ispra Waste Register. Data on prevention, reduction and recycling policies, the collection service and initiatives to facilitate and incentivize correct disposal (*e.g*: good practices at schools/offices/etc.; reduction of food waste, repair and reuse centres, awareness campaigns, composting, characteristics of the collection service and types of waste collected) come from direct surveys and are derived from the thematic archives of the administrations.

The considered municipalities are 109, Latina was omitted due to missing data. Two municipalities are contiguous in the contiguity matrix if their distance is smaller or equal to 80 km (Fig. 3).

Municipal waste accounts for a small fraction of the total waste produced (17.9% in 2021), but its management is particularly complex due to the heterogeneity of its composition and origin. High quality and quantity standards of separate collection facilitate the achievement of the targets for preparation for reuse and recycling set by the Circular Economy Package (Directive 2008/851/EU) and the National Plan for Recovery and Resilience NRRP (Mission 2 Component 1). In 2022, at the national level, separate collection is 63.7% of the municipal waste produced, but only 63 (57.8%) of the municipalities have reached the 65.0% target set by Legislative Decree 152/2006 for 2012. In the capital municipalities, the share of separate collection is 55.1%.

The highest quotas are found in the North-East (68.5%), the North-West (60.8%) and the Centre (53.2%); the South (46.6%) and the Islands (38.2%) still lag behind, despite the increase compared to the previous year.

The FCMd-MD-SP model was used for different values of C and γ and the best combination with respect to the Xie–Beni index was $C = 3$, $\gamma = 0.3$ (I_{XB}=1.89). The value of the fuzziness parameter m is equal to 1.5 (D'Urso (2015)). The variables and their summary statistics are reported in Table 3; alongside the weights computed in the clustering process for the different attributes types as in (5). The complete data are presented in Table 9 (Appendix). The two considered quantitative variables were N3, and N4.

The medoids are Caserta (cluster 1), Messina (cluster 2), and Udine (cluster 3), in bold in Table 9 in Appendix. The cardinalities of the clusters are 23, 8, and 70, respectively. The partition obtained is presented in Table 5 and in Fig. 4.

The mean values of the variables in the three clusters are presented in Table 4.

Cluster 2, with medoid Messina, is made up of municipalities with all the categorical variables well under or equal to the mean (Presence of rebates or actions to encourage self-composting at households over the mean), and the lowest values of the variable Separate municipal waste collection (% over kg/inhabitant) - Palermo 15.6% the lowest.

Cluster 3, with medoid Udine, is made up of municipalities with all the categorical variables over (Reduce food waste at markets, restaurants, canteens, stores equal to) the mean, Separate municipal waste collection 2022 (% over maximum, that is, the value measured in Piacenza) slightly over the mean and Separate municipal waste collection (% over kg/inhabitant) well over the mean - Ferrara 87.6% the highest.

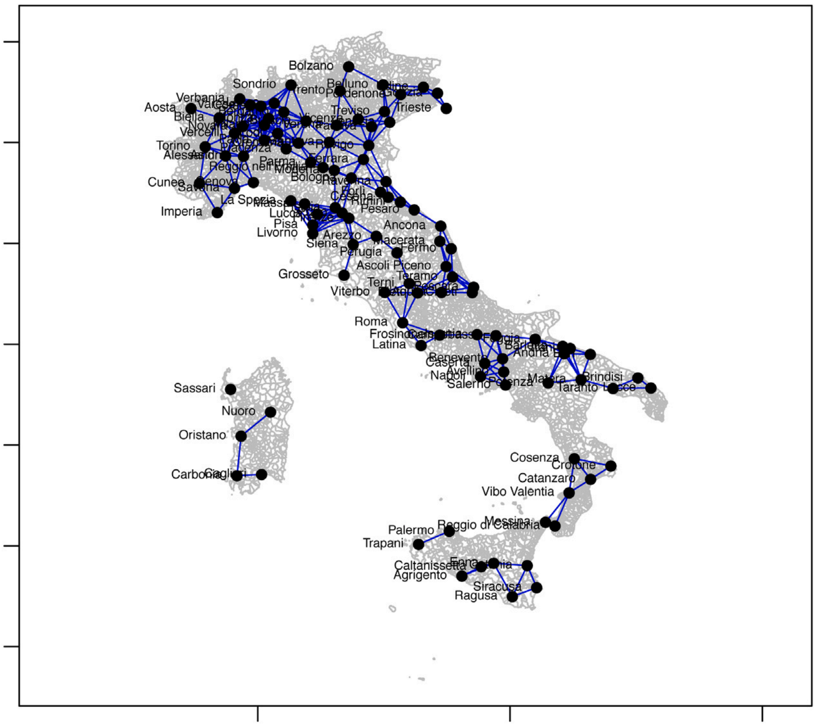


Fig. 3. Contiguity of the municipalities.

Table 3
Environmental waste variables.

Attribute	Variables	Mean	Weight (5)
Categorical	C1. Reduce food waste	0.23	0.30
	C2. Reduce packaging	0.12	
	C3. Dematerialize advertising and communications	0.11	
	C4. Reduce food waste at markets, restaurants, canteens, stores	0.29	
	C5. Awareness campaigns on prevention 2022	0.58	
	C6. Initiatives or concessions to purchase washable diapers	0.12	
	C7. Discounts to non-households that: implement policies to prevent their own municipal waste	0.28	
	C8. Discounts to non-households that: Send their municipal waste for recycling	0.45	
	C9. Presence of rebates or actions to encourage self-composting at households	0.76	
Numeric	N1. Municipal waste production in provincial capitals/metropolitan cities (kg/inhabitant)	518.80	0.70
	N2. Separate municipal waste collection (kg per inhabitant)	383.33	
	N3. Separate municipal waste collection(% over maximum)	68.7%	
	N4. Separate municipal waste collection(% over kg/inhabitant)	63.7%	

Table 4
Mean values of the variables in each cluster.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	N1	N2	N3	N4
1	0.09	0.13	0.13	0.30	0.39	0.00	0.09	0.22	0.70	491.27	262.95	65.04	55.06
2	0.00	0.00	0.13	0.00	0.50	0.00	0.00	0.50	0.88	468.42	225.72	62.01	49.36
3	0.29	0.13	0.10	0.32	0.64	0.17	0.36	0.51	0.77	532.14	360.31	70.45	67.75

Cluster 1, with medoid Caserta, is in between cluster 2 and cluster 3. Cluster 1, compared to cluster 2, has better values of the categorical variables and of the numerical variable Separate municipal waste collection (% over kg/inhabitant).

In particular, the three clusters contain 9 (39.1%), 3 (37.5%), 51 (72.9%) of the 63 municipalities that have reached the 65.0% target for the share of Separate municipal waste collection.

The partition defines two geographical areas, North-Centre and South, and a small area within the South, composed by eight municipalities (Cosenza, Reggio di Calabria, Catanzaro, Vibo Valentia, Crotone, Palermo, Trapani, Messina) which form two components disconnected from the rest of southern Italy by the sparsely populated areas in northern Calabria and central Sicily. It is worth noting that the municipalities of Sardinia belong to the cluster composed of the municipalities in North-Centre due to

Table 5
Membership degrees and highest membership cluster - $C = 3$ EFMd-SpID and FMd-SpID. In bold the EFMd-SpID medoids, in italic the EFMd-SpID fuzzy municipalities.

	Municipality	*cluster* 1	*cluster* 2	*cluster* 3	*cluster*
1	Torino	0.002	0.002	0.996	3
2	Novara	0.000	0.000	1.000	3
3	Vercelli	0.000	0.000	1.000	3
4	Cuneo	0.000	0.000	1.000	3
5	Mantova	0.000	0.000	1.000	3
6	Lodi	0.000	0.000	1.000	3
7	Verbania	0.000	0.000	1.000	3
8	Foggia	0.996	0.002	0.002	1
9	Aosta	0.000	0.000	1.000	3
10	Cremona	0.000	0.000	0.999	3
11	Ferrara	0.001	0.001	0.998	3
12	Ravenna	0.001	0.001	0.999	3
13	Pisa	0.002	0.001	0.997	3
14	Asti	0.000	0.000	1.000	3
15	Arezzo	0.001	0.001	0.999	3
16	Terni	0.000	0.000	1.000	3
17	Alessandria	0.001	0.001	0.999	3
18	Modena	0.002	0.002	0.996	3
19	Ancona	0.000	0.000	0.999	3
20	Venezia	0.001	0.001	0.998	3
21	Trento	0.004	0.003	0.993	3
22	Como	0.000	0.000	1.000	3
23	Avellino	0.999	0.001	0.001	1
24	Piacenza	0.002	0.002	0.996	3
25	Parma	0.000	0.000	0.999	3
26	Lecce	0.994	0.003	0.003	1
27	Perugia	0.000	0.000	1.000	3
28	Rieti	0.000	0.000	1.000	3
29	Varese	0.000	0.000	1.000	3
30	Biella	0.000	0.000	1.000	3
31	Sondrio	0.000	0.000	0.999	3
32	Pavia	0.000	0.000	1.000	3
33	Livorno	0.000	0.000	1.000	3
34	Prato	0.000	0.000	1.000	3
35	Lucca	0.001	0.001	0.998	3
36	Lecco	0.000	0.000	1.000	3
37	Milano	0.000	0.000	1.000	3
38	Bergamo	0.000	0.000	1.000	3
39	Brescia	0.000	0.000	1.000	3
40	Grosseto	0.005	0.005	0.990	3
41	Bolzano - Bozen	0.007	0.007	0.987	3
42	Udine	0.000	0.000	1.000	3
43	Belluno	0.000	0.000	0.999	3
44	Vicenza	0.000	0.000	0.999	3
45	Gorizia	0.000	0.000	1.000	3
46	Trieste	0.006	0.006	0.988	3
47	Monza	0.000	0.000	1.000	3
48	Padova	0.000	0.000	1.000	3
49	Verona	0.000	0.000	1.000	3
50	Rovigo	0.000	0.000	1.000	3
51	Siena	0.001	0.001	0.999	3
52	Pordenone	0.000	0.000	0.999	3
53	Treviso	0.000	0.000	0.999	3
54	Ascoli Piceno	0.000	0.000	1.000	3
55	Imperia	0.001	0.001	0.998	3
56	Pesaro	0.000	0.000	1.000	3
57	Genova	0.005	0.005	0.990	3
58	Cesena	0.001	0.001	0.998	3
59	Bologna	0.000	0.000	1.000	3
60	Forlì	0.000	0.000	1.000	3

(*continued on next page*)

the virtuous behaviour concerning the sharing of Separate municipal waste collection (Nuoro 83.8%, Oristano 80.6%). The model identifies two fuzzy provinces: Frosinone and Sassari, the provinces with the lowest membership to cluster 3.

Table 5 (*continued*).

	Municipality	*cluster* 1	*cluster* 2	*cluster* 3	*cluster*
61	La Spezia	0.001	0.001	0.997	3
62	Rimini	0.001	0.001	0.998	3
63	Massa	0.001	0.000	0.999	3
64	Isernia	0.976	0.012	0.013	1
65	Reggio nell'Emilia	0.001	0.001	0.997	3
66	Firenze	0.000	0.000	1.000	3
67	Pistoia	0.000	0.000	0.999	3
68	Caserta	1.000	0.000	0.000	1
69	Fermo	0.000	0.000	1.000	3
70	Benevento	1.000	0.000	0.000	1
71	Macerata	0.000	0.000	1.000	3
72	Roma	0.018	0.016	0.966	3
73	Viterbo	0.001	0.001	0.997	3
74	Savona	0.002	0.002	0.996	3
75	L'Aquila	0.001	0.001	0.998	3
76	Pescara	0.001	0.001	0.998	3
77	Frosinone	0.418	0.113	0.469	3
78	Chieti	0.000	0.000	1.000	3
79	Teramo	0.000	0.000	1.000	3
80	Campobasso	0.998	0.001	0.001	1
81	Napoli	0.999	0.000	0.000	1
82	Cosenza	0.000	0.999	0.000	2
83	Reggio di Calabria	0.001	0.999	0.001	2
84	Salerno	1.000	0.000	0.000	1
85	Catanzaro	0.001	0.999	0.001	2
86	Bari	0.999	0.001	0.001	1
87	Taranto	0.991	0.005	0.004	1
88	Brindisi	0.996	0.002	0.002	1
89	Barletta	0.999	0.000	0.000	1
90	Andria	1.000	0.000	0.000	1
91	Trani	0.999	0.001	0.001	1
92	Potenza	0.998	0.001	0.001	1
93	Matera	0.998	0.001	0.001	1
94	Vibo Valentia	0.000	0.999	0.000	2
95	Crotone	0.016	0.971	0.013	2
96	Palermo	0.083	0.861	0.056	2
97	Siracusa	1.000	0.000	0.000	1
98	Ragusa	0.999	0.001	0.001	1
99	Enna	0.999	0.001	0.001	1
100	Catania	0.963	0.019	0.018	1
101	Agrigento	0.997	0.001	0.001	1
102	Caltanissetta	1.000	0.000	0.000	1
103	Trapani	0.018	0.963	0.019	2
104	Messina	0.000	1.000	0.000	2
105	Nuoro	0.007	0.007	0.986	3
106	Sassari	0.152	0.286	0.563	3
107	Carbonia	0.000	0.000	0.999	3
108	Oristano	0.001	0.001	0.999	3
109	Cagliari	0.001	0.001	0.999	3

4.2. European elections 2024

All Italian-language tweets posted by accounts related to eight coalitions were collected in the period between May 13 and June 2, 2024. A binary network of Twitter accounts (nodes) was constructed based on retweets, replies, mentions, hashtags and account mentions (Fig. 5). The colour of the nodes indicates the coalition among the eight to which the accounts belong.

The accounts were clustered according to two quantitative variables, *account followers* and *account following*, transformed in logarithm to base 10, and two qualitative variables, *account political party* and *account is verified* (see Fig. 6). The value of the fuzziness parameter m is equal to 1.5 (D'Urso (2015)). The FCMd-MD-SP model was used for different values of C and γ and the best combination with respect to the Xie–Beni index was $C = 3$, $\gamma = 0.02$ (I_{XB}=1.33).

The medoids are Min_Casellati (Forza Italia), bendellavedova (Stati Uniti d'Europa), Azione_it (Azione) (Table 6). The weights of the continuous variables and of the categorical variables are 0.96, 0.04.

Cluster 1 is characterized by low values of *followers* and high values of *following*. Cluster 2 is characterized by medium values of *followers* and low values of *following*. Cluster 3 is characterized by very high values of *followers* and medium values of *following*. The official accounts of the parties and of their respective leaders are in cluster 3: Alleanza Verdi Sinistra - NFratoianni/AngeloBonelli1, Azione - CarloCalenda, forza-italia - Antonio_Tajani, FratellidItalia - GiorgiaMeloni, LegaSalvini- matteosalvini, Mov5Stelle - GiuseppeConteIT; pdnetwork - ellyesse, Stati Uniti d'Europa - matteorenzi.

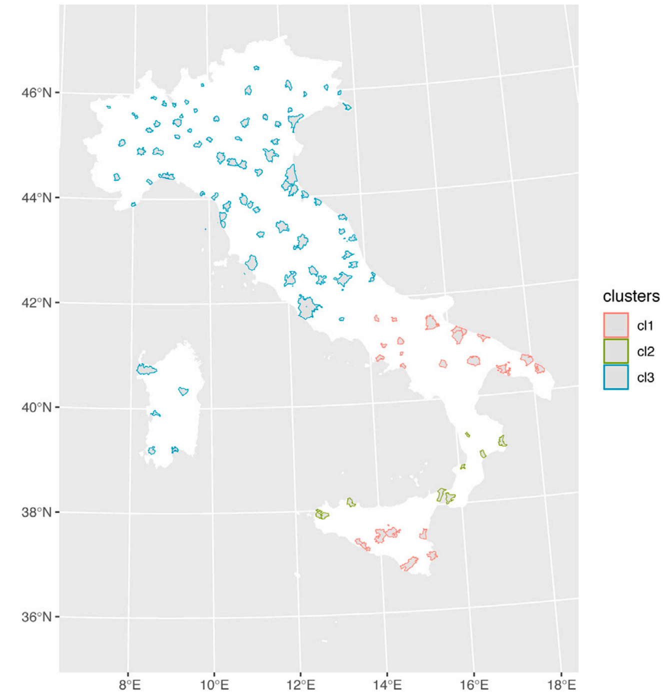


Fig. 4. Partition of the municipalities.

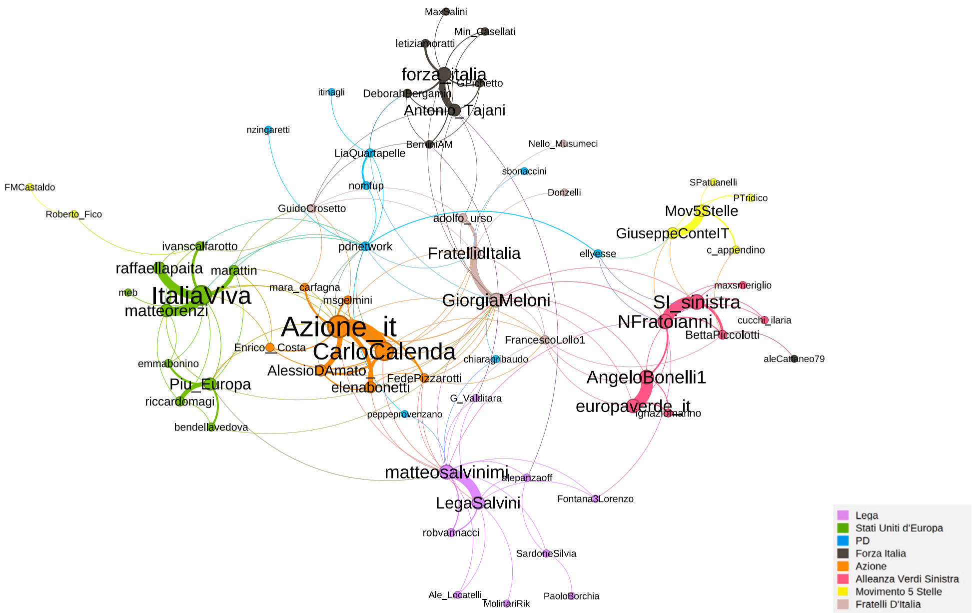


Fig. 5. Network of Italian accounts in European elections 2024.

From Tables 7, 8 it is possible to observe the accounts grouped in their political party despite the low value of γ - Alleanza Verdi Sinistra, Azione, PD, Fratelli D'Italia, Stati Uniti d'Europa, and the accounts split according to their political party, Forza Italia, Lega, Movimento 5 Stelle.

The spatial fuzzy mixed model allowed to partition the italian provinces on the basis of categorical and numerical environmental data taking into account the contiguity, for designing proper public policies. The model made it possible also to study whether,

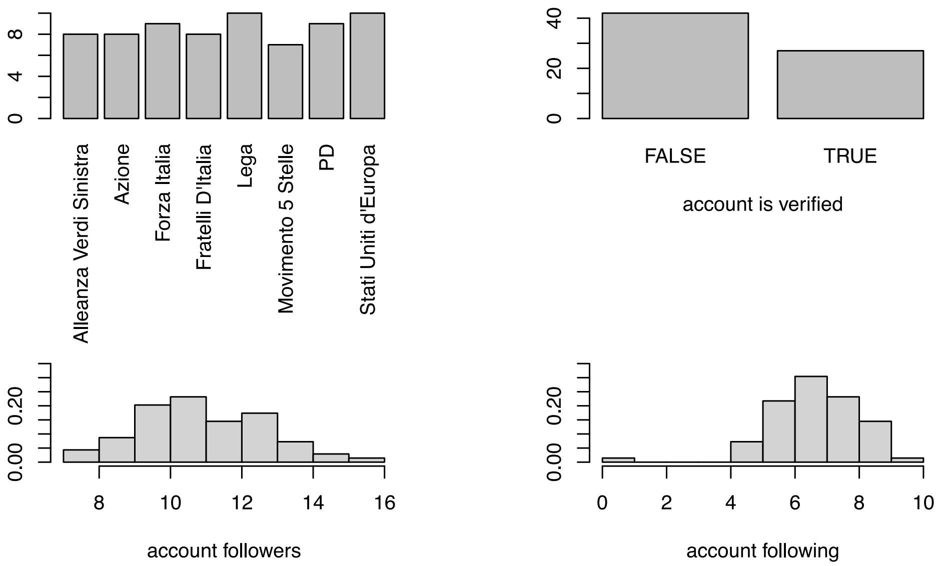


Fig. 6. Variables of the accounts (account followers and following in logarithmic to base 10 scale).

Table 6
Medoids.

	Account	*Followers* (a)	*Following* (b)	*Political party*	*is verified*	*log(a)*	*log(b)*
1	Min_Casellati	18 597.85	848.75	Forza Italia	1	4.27	2.93
2	bendellavedova	29 353.86	465	Stati Uniti d'Europa	1	4.47	2.67
3	Azione_it	82 824.94	779	Azione	1	4.92	2.89

Table 7
Account political party and clustering.

Account political party	*cluster 1*	*cluster 2*	*cluster 3*
Alleanza Verdi Sinistra	1	0	7
Azione	0	0	8
Forza Italia	4	1	4
Fratelli D'Italia	1	0	7
Lega	5	2	3
Movimento 5 Stelle	3	0	4
PD	1	0	8
Stati Uniti d'Europa	0	2	8

during the European 2024 elections, communication (the adjacency matrix) took place between candidates with similar activities and characteristics on social networks (mixed attributes) and belonging to the same political coalition.

5. Final remarks

The proposed FCMd-MD-SP model fills, to our knowledge, a gap by presenting a clustering model for mixed data with spatial constraints. The characteristics of the proposed model are: *mixed data*; *fuzziness*; *spatial information*.

A simulation study is described in which we showcase how the model can detect among the variables provided the ones that carry more relevant information. Two applications, one to environmental data of Italian municipalities and the other to Italian accounts of political coalitions in the European elections 2024 show the performances and the ability to analyse empirical data in the case of spatial mixed data of the model, respectively.

The spatial fuzzy mixed model allowed to partition the italian provinces on the basis of categorical and numerical environmental data taking into account the contiguity, for designing proper public policies. The model made it possible also to study whether, during the European 2024 elections, communication (the adjacency matrix) took place between candidates with similar activities and characteristics on social networks (mixed attributes) and belonging to the same political coalition.

The modelization of the spatial correction plays an important and delicate role, and we leave to future studies the further optimization of the spatial correction term and the parameter that governs its relevance. Entropic and robust versions of the proposed model may be considered in the future.

Table 8
Accounts.

	Account	*Followers* (a)	*Following* (b)	*Political party*	*is verified*	*log(a)*	*log(b)*	*cl*
1	maxsmeriglio	7794.75	1320	Alleanza Verdi Sinistra	1	3.89	3.12	1
2	europaverde_it	14 768.88	181	Alleanza Verdi Sinistra	0	4.17	2.26	3
3	SI_sinistra	138 815.51	3747.45	Alleanza Verdi Sinistra	0	5.14	3.57	3
4	AngeloBonelli1	13 589.35	897	Alleanza Verdi Sinistra	0	4.13	2.95	3
5	BettaPiccolotti	11 188.8	3838.47	Alleanza Verdi Sinistra	0	4.05	3.58	3
6	ignaziomarino	390 912.62	2308.2	Alleanza Verdi Sinistra	0	5.59	3.36	3
7	NFratoianni	70 006.46	2046	Alleanza Verdi Sinistra	0	4.85	3.31	3
8	cucchi_ilaria	43 421	57	Alleanza Verdi Sinistra	0	4.64	1.76	3
9	**Azione_it**	82 824.94	779	Azione	1	4.92	2.89	3
10	CarloCalenda	470 826.11	273.25	Azione	1	5.67	2.44	3
11	elenabonetti	34 207.43	755	Azione	1	4.53	2.88	3
12	AlessioDAmato_	2643.51	229.24	Azione	0	3.42	2.36	3
13	FedePizzarotti	46 720.96	116	Azione	1	4.67	2.06	3
14	msgelmini	105 746	2828.38	Azione	0	5.02	3.45	3
15	Enrico_Costa	11 804.35	92	Azione	0	4.07	1.96	3
16	mara_carfagna	239 815	1476	Azione	0	5.38	3.17	3
17	GPichetto	3468.69	629.23	Forza Italia	0	3.54	2.80	1
18	letiziamoratti	7556.33	1225.11	Forza Italia	1	3.88	3.09	1
19	**Min_Casellati**	18 597.85	848.75	Forza Italia	1	4.27	2.93	1
20	aleCattaneo79	17 916.5	1488	Forza Italia	1	4.25	3.17	1
21	MaxSalini	13 821.75	196	Forza Italia	0	4.14	2.29	2
22	Antonio_Tajani	116 087.64	3612.73	Forza Italia	0	5.06	3.56	3
23	BerniniAM	44 319.58	661	Forza Italia	1	4.65	2.82	3
24	DeborahBergamin	26 113.84	779.42	Forza Italia	0	4.42	2.89	3
25	forza_italia	195 290.07	765	Forza Italia	0	5.29	2.88	3
26	Nello_Musumeci	20 500	3395	Fratelli D'Italia	0	4.31	3.53	1
27	GuidoCrosetto	292 956	2306	Fratelli D'Italia	1	5.47	3.36	3
28	FratellidItalia	305 688.83	403	Fratelli D'Italia	0	5.49	2.61	3
29	adolfo_urso	14 491.85	1003	Fratelli D'Italia	1	4.16	3.00	3
30	DSantanche	207 636.77	1636	Fratelli D'Italia	0	5.32	3.21	3
31	FrancescoLollo1	31 972	216	Fratelli D'Italia	0	4.50	2.33	3
32	GiorgiaMeloni	2266929.71	248	Fratelli D'Italia	1	6.36	2.39	3
33	Donzelli	24 437.5	1305.5	Fratelli D'Italia	1	4.39	3.12	3
34	G_Valditara	10 411.1	1049	Lega	0	4.02	3.02	1
35	SardoneSilvia	42 174.63	3635.58	Lega	1	4.63	3.56	1
36	alepanzaoff	3871.11	1853.79	Lega	1	3.59	3.27	1
37	PaoloBorchia	2471	467	Lega	0	3.39	2.67	1
38	Ale_Locatelli_	4249	224	Lega	0	3.63	2.35	1
39	robvannacci	1378	1	Lega	0	3.14	0.00	2
40	MolinariRik	25 613	211	Lega	0	4.41	2.32	2
41	LegaSalvini	246 577.7	515	Lega	0	5.39	2.71	3
42	matteosalvinimi	1524363.04	1983.6	Lega	1	6.18	3.30	3
43	Fontana3Lorenzo	30 321.42	176	Lega	0	4.48	2.25	3
44	FMCastaldo	17 098.4	3475.6	Movimento 5 Stelle	1	4.23	3.54	1
45	PTridico	6855.5	508.36	Movimento 5 Stelle	0	3.84	2.71	1
46	SPatuanelli	18 987.5	723	Movimento 5 Stelle	0	4.28	2.86	1
47	Mov5Stelle	745 451.66	249	Movimento 5 Stelle	0	5.87	2.40	3
48	GiuseppeConteIT	1199356.36	137.44	Movimento 5 Stelle	1	6.08	2.14	3
49	c_appendino	92 964.67	943	Movimento 5 Stelle	0	4.97	2.97	3
50	Roberto_Fico	213 900	413	Movimento 5 Stelle	0	5.33	2.62	3
51	chiaragribaudo	12 591.21	1261	PD	0	4.10	3.10	1
52	peppeprovenzano	31 549.88	922.38	PD	0	4.50	2.96	3
53	nomfup	134 554.31	8093.6	PD	0	5.13	3.91	3
54	LiaQuartapelle	38 133	6246.33	PD	0	4.58	3.80	3
55	nzingaretti	574 746.57	1763	PD	1	5.76	3.25	3
56	pdnetwork	424 914.36	497	PD	0	5.63	2.70	3
57	ellyesse	192 500	7026	PD	0	5.28	3.85	3
58	sbonaccini	170 583.12	11 567.76	PD	0	5.23	4.06	3
59	itinagli	45 871.65	636.45	PD	0	4.66	2.80	3
60	**bendellavedova**	29 353.86	465	Stati Uniti d'Europa	1	4.47	2.67	2
61	riccardomagi	31 844.82	2619	Stati Uniti d'Europa	1	4.50	3.42	2
62	ItaliaViva	62 001.3	347	Stati Uniti d'Europa	0	4.79	2.54	3
63	Piu_Europa	58 834.03	116	Stati Uniti d'Europa	1	4.77	2.06	3
64	marattin	117 714.31	150	Stati Uniti d'Europa	1	5.07	2.18	3
65	raffaellapaita	17 605.45	3982.28	Stati Uniti d'Europa	1	4.25	3.60	3

(*continued on next page*)

Table 8 (*continued*).

	Account	*Followers* (a)	*Following* (b)	*Political party*	*is verified*	*log(a)*	*log(b)*	*cl*
66	ivanscalfarotto	111 136.04	2030	Stati Uniti d'Europa	1	5.05	3.31	3
67	emmabonino	244 353.09	343.36	Stati Uniti d'Europa	1	5.39	2.54	3
68	matteorenzi	3325194.55	968	Stati Uniti d'Europa	1	6.52	2.99	3
69	meb	646 831.57	244	Stati Uniti d'Europa	0	5.81	2.39	3

Table 9
Membership degrees and highest membership cluster - $C = 3$.

	Municipality	C1	C2	C3	C4	C5	C6	C7	C8	C9	N1	N2	N3	N4
1	Torino	1	0	0	1	1	1	1	1	0	477.6	259.9	63.2	54.4
2	Novara	0	0	0	0	1	0	0	1	1	542.0	430.8	71.8	79.5
3	Vercelli	0	0	0	0	0	0	0	0	0	661.5	495.8	87.6	75.0
4	Cuneo	0	0	0	1	1	0	1	1	1	505.8	346.4	67.0	68.5
5	Mantova	0	0	0	1	1	0	1	1	1	521.5	432.1	69.0	82.9
6	Lodi	0	0	0	0	0	0	0	0	0	407.5	300.6	53.9	73.8
7	Verbania	0	0	0	0	0	0	1	0	1	628.6	487.9	83.2	77.6
8	Foggia	0	0	0	0	0	0	0	0	0	544.8	141.4	72.1	25.9
9	Aosta	0	0	0	0	1	0	0	0	1	490.1	339.2	64.9	69.2
10	Cremona	1	1	1	1	1	0	0	1	1	460.5	360.2	61.0	78.2
11	Ferrara	1	0	0	1	1	0	1	1	1	639.8	560.5	84.7	87.6
12	Ravenna	0	0	0	1	1	0	0	0	1	715.1	482.3	94.7	67.4
13	Pisa	0	0	0	0	0	0	1	1	1	753.4	488.1	99.7	64.8
14	Asti	1	1	1	0	1	0	0	1	1	486.1	328.2	64.4	67.5
15	Arezzo	0	0	0	0	1	0	0	1	1	577.6	313.2	76.5	54.2
16	Terni	0	0	0	0	0	0	0	0	1	455.3	335.0	60.3	73.6
17	Alessandria	0	0	1	0	1	0	1	1	1	554.3	248.8	73.4	44.9
18	Modena	1	1	0	1	1	1	1	1	1	657.0	401.1	87.0	61.0
19	Ancona	0	0	0	1	1	0	1	1	1	479.7	300.3	63.5	62.6
20	Venezia	1	1	1	1	1	0	0	1	1	628.0	393.7	83.1	62.7
21	Trento	1	1	1	1	1	1	0	0	1	443.3	365.3	58.7	82.4
22	Como	0	0	0	0	0	0	0	1	0	463.7	319.4	61.4	68.9
23	Avellino	0	0	0	0	0	0	0	0	0	418.7	277.0	55.4	66.2
24	Piacenza	1	0	1	1	0	0	0	1	1	755.4	542.1	100.0	71.8
25	Parma	1	0	0	0	0	0	1	0	1	563.7	458.0	74.6	81.2
26	Lecce	0	0	0	1	0	0	1	1	1	545.3	382.1	72.2	70.1
27	Perugia	0	0	0	1	1	0	0	0	1	556.2	397.8	73.6	71.5
28	Rieti	0	0	0	0	0	0	0	0	0	485.3	268.6	64.2	55.3
29	Varese	0	0	0	0	0	0	1	0	0	460.7	321.4	61.0	69.8
30	Biella	1	0	0	0	0	0	0	0	1	558.8	429.9	74.0	76.9
31	Sondrio	0	0	0	0	0	0	0	0	0	485.9	258.8	64.3	53.3
32	Pavia	1	1	0	0	0	1	0	0	1	497.7	300.2	65.9	60.3
33	Livorno	0	0	0	0	0	0	0	0	1	553.7	350.3	73.3	63.3
34	Prato	1	1	0	1	1	0	0	1	1	603.5	440.4	79.9	73.0
35	Lucca	1	0	0	1	0	1	1	0	1	646.3	528.4	85.6	81.8
36	Lecco	0	0	0	1	0	1	0	0	0	466.1	349.6	61.7	75.0
37	Milano	1	0	1	1	1	0	1	1	0	469.1	291.4	62.1	62.1
38	Bergamo	1	0	0	0	1	0	1	1	1	479.4	368.0	63.5	76.8
39	Brescia	0	0	0	0	1	0	0	0	0	507.0	343.8	67.1	67.8
40	Grosseto	0	0	1	0	1	0	0	1	0	583.4	348.7	77.2	59.8
41	Bolzano - Bozen	1	1	0	1	0	1	0	0	0	485.8	324.0	64.3	66.7
42	**Udine**	0	0	0	0	1	0	0	0	1	522.4	357.5	69.2	68.4
43	Belluno	0	0	0	0	1	1	0	0	1	458.3	395.5	60.7	86.3
44	Vicenza	1	0	0	0	1	1	1	1	1	613.2	464.2	81.2	75.7

(*continued on next page*)

Acknowledgement

The authors are grateful to the reviewers for the contribution to the improvement of the paper.

Appendix

See Table 9.

Table 9 (*continued*).

	Municipality	C1	C2	C3	C4	C5	C6	C7	C8	C9	N1	N2	N3	N4
45	Gorizia	0	0	0	0	1	0	0	0	1	492.5	320.6	65.2	65.1
46	Trieste	0	0	0	0	1	0	1	1	1	503.1	226.6	66.6	45.0
47	Monza	1	0	0	1	1	0	1	1	1	403.6	292.3	53.4	72.4
48	Padova	0	0	0	0	1	1	0	1	1	596.4	383.3	78.9	64.3
49	Verona	0	0	0	0	0	0	0	0	0	493.9	265.1	65.4	53.7
50	Rovigo	0	0	0	0	1	0	0	0	1	596.5	409.4	79.0	68.6
51	Siena	1	1	0	1	0	0	0	0	1	592.6	367.5	78.4	62.0
52	Pordenone	0	0	0	0	1	1	1	1	1	497.0	421.1	65.8	84.7
53	Treviso	0	0	0	1	1	1	0	0	1	453.6	395.2	60.1	87.1
54	Ascoli Piceno	0	0	0	0	0	0	0	0	0	499.2	344.3	66.1	69.0
55	Imperia	0	0	0	0	0	0	1	1	1	463.8	310.2	61.4	66.9
56	Pesaro	0	0	0	0	1	0	0	1	1	574.6	389.9	76.1	67.9
57	Genova	1	0	0	0	1	0	1	1	1	501.0	214.7	66.3	42.8
58	Cesena	0	0	0	0	0	0	0	1	1	653.2	514.3	86.5	78.7
59	Bologna	0	0	0	0	1	0	1	1	1	522.2	330.2	69.1	63.2
60	Forlì	0	0	0	0	1	0	0	0	1	445.4	363.9	59.0	81.7
61	La Spezia	1	0	0	1	0	0	1	1	1	515.9	408.6	68.3	79.2
62	Rimini	0	0	0	1	1	0	1	1	1	671.2	446.5	88.9	66.5
63	Massa	0	0	0	0	0	0	1	1	1	665.8	434.8	88.1	65.3
64	Isernia	0	0	0	1	0	0	0	0	1	442.7	211.9	58.6	47.9
65	Reggio nell'Emilia	1	0	0	1	1	0	1	1	1	646.0	535.2	85.5	82.8
66	Firenze	0	0	0	0	1	0	0	0	1	614.5	338.0	81.3	55.0
67	Pistoia	1	1	0	1	1	0	0	1	1	515.3	251.0	68.2	48.7
68	**Caserta**	0	0	0	1	1	0	0	0	1	512.2	277.6	67.8	54.2
69	Fermo	0	0	0	0	1	0	0	0	0	532.0	349.2	70.4	65.6
70	Benevento	0	0	0	0	1	0	0	1	1	450.7	299.2	59.7	66.4
71	Macerata	0	0	0	1	1	0	0	0	1	460.9	343.2	61.0	74.5
72	Roma	0	0	0	0	1	0	1	1	1	578.6	265.4	76.6	45.9
73	Viterbo	0	0	0	0	1	0	0	1	1	408.0	227.5	54.0	55.8
74	Savona	0	0	0	0	0	0	0	1	1	533.8	218.7	70.7	41.0
75	L'Aquila	0	0	0	0	0	0	1	0	1	494.3	204.6	65.4	41.4
76	Pescara	0	0	0	0	0	0	0	0	1	524.6	245.9	69.5	46.9
77	Frosinone	0	0	0	0	1	0	0	1	1	495.2	344.9	65.6	69.6
78	Chieti	0	0	0	0	1	0	0	0	1	509.6	350.2	67.5	68.7
79	Teramo	0	0	0	0	0	0	0	0	1	411.3	298.3	54.4	72.5
80	Campobasso	0	1	1	0	1	0	0	1	1	401.7	178.0	53.2	44.3
81	Napoli	0	0	0	0	1	0	0	0	1	564.0	227.9	74.7	40.4
82	Cosenza	0	0	1	0	1	0	0	0	1	425.6	255.4	56.3	60.0
83	Reggio di Calabria	0	0	0	0	1	0	0	1	1	395.8	162.9	52.4	41.2
84	Salerno	0	0	0	0	0	0	0	0	1	455.6	295.2	60.3	64.8
85	Catanzaro	0	0	0	0	0	0	0	0	1	427.5	293.3	56.6	68.6
86	Bari	0	0	0	0	0	0	0	1	1	554.8	221.9	73.4	40.0
87	Taranto	0	0	0	0	0	0	0	0	0	549.0	153.1	72.7	27.9
88	Brindisi	1	1	1	1	1	0	0	0	0	541.8	245.5	71.7	45.3
89	Barletta	0	0	0	0	0	0	0	0	0	450.6	304.2	59.7	67.5
90	Andria	0	0	0	0	0	0	0	0	0	446.2	275.5	59.1	61.7
91	Trani	0	0	0	0	0	0	0	0	1	464.0	343.3	61.4	74.0
92	Potenza	0	0	0	0	0	0	0	0	0	413.4	250.8	54.7	60.7
93	Matera	1	1	1	1	0	0	0	0	1	416.5	303.1	55.1	72.8
94	Vibo Valentia	0	0	0	0	1	0	0	1	1	450.4	314.8	59.6	69.9
95	Crotone	0	0	0	0	0	0	0	0	0	510.1	109.3	67.5	21.4
96	Palermo	0	0	0	0	0	0	0	1	1	558.4	84.7	73.9	15.2
97	Siracusa	0	0	0	0	1	0	0	0	1	516.5	260.4	68.4	50.4
98	Ragusa	0	0	0	1	1	0	0	0	1	488.7	344.9	64.7	70.6
99	Enna	0	0	0	0	1	0	0	0	1	411.6	276.7	54.5	67.2
100	Catania	0	0	0	0	0	0	0	0	1	733.4	161.3	97.1	22.0
101	Agrigento	0	0	0	0	1	0	0	1	1	491.4	335.9	65.1	68.4
102	Caltanissetta	0	0	0	1	0	0	1	0	1	485.6	280.6	64.3	57.8
103	Trapani	0	0	0	0	1	0	0	0	1	526.4	343.0	69.7	65.2
104	**Messina**	0	0	0	0	0	0	0	1	1	453.2	242.3	60.0	53.5
105	Nuoro	0	0	0	0	1	0	0	1	1	425.7	356.8	56.4	83.8
106	Sassari	0	0	0	0	1	0	1	1	1	489.4	306.1	64.8	62.6
107	Carbonia	0	0	0	0	1	0	0	0	0	447.2	342.1	59.2	76.5
108	Oristano	0	0	0	0	1	1	0	0	0	510.0	411.1	67.5	80.6
109	Cagliari	0	0	0	0	1	0	0	0	0	468.5	350.5	62.0	74.8

References

Abdali, E., Valadan Zoej, M.J., Taheri Dehkordi, A., Ghaderpour, E., 2023. A parallel-cascaded ensemble of machine learning models for crop type classification

in Google earth engine using multi-temporal sentinel-1/2 and landsat-8/9 remote sensing data. Remote Sens. 16 (1), 127.
Ado, M., Amitab, K., Maji, A.K., Jasińska, E., Gono, R., Leonowicz, Z., Jasiński, M., 2022. Landslide susceptibility mapping using machine learning: A literature survey. Remote Sens. 14 (13), 3029.
Ahmad, A., Dey, L., 2007. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng. 63 (2), 503–527.
Ambroise, C., Dang, M., 2009. Spatial data clustering. Data Analysis 289–318.
Antoni, L., Krajči, S., Krídlo, O., Macek, B., Pisková, L., 2014. On heterogeneous formal contexts. Fuzzy Sets and Systems 234, 22–33.
Bezdek, J., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic, Norwell, MA, USA.
Coppi, R., D'Urso, P., Giordani, P., 2010. A fuzzy clustering model for multivariate spatial time series. J. Classification 27 (1), 54–88.
Deza, M.M., Deza, E., 2009. Encyclopedia of distances. In: Encyclopedia of Distances. Springer, pp. 1–583.
D'Urso, P., 2015. Fuzzy clustering. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (Eds.), Handbook of Cluster Analysis. Chapman and Hall, pp. 545–573.
D'Urso, P., De Giovanni, L., Disegna, M., Massari, R., 2019. Fuzzy clustering with spatial–temporal information. Spatial Stat. 30, 71–102.
D'Urso, P., De Giovanni, L., Federico, L., Vitale, V., 2023a. Fuzzy clustering of spatial interval-valued data. Spatial Stat. 57, 100764.
D'Urso, P., De Giovanni, L., Massari, R., 2014. Time series clustering by a robust autoregressive metric with application to air pollution. Chemometr. Intell. Lab. Syst. 141, 107–124.
D'Urso, P., De Giovanni, L., Vitale, V., 2022. Spatial robust fuzzy clustering of COVID 19 time series based on B-splines. Spatial Stat. 49, 100518.
D'Urso, P., De Giovanni, L., Vitale, V., 2023b. A robust method for clustering football players with mixed attributes. Ann. Oper. Res. 325 (1), 9–36.
D'Urso, P., Maharaj, E., 2009. Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets and Systems 160 (24), 3565–3589.
D'Urso, P., Massari, R., 2013. Fuzzy clustering of human activity patterns. Fuzzy Sets and Systems 215, 29–54.
D'Urso, P., Massari, R., 2019. Fuzzy clustering of mixed data. Inform. Sci. 505, 513–534.
Everitt, B.S., 1988. A finite mixture model for the clustering of mixed-mode data. Stat. Probab. Lett. 6 (5), 305–309.
Everitt, B., Landau, S., Leese, M., Stahl, D., 2011. Cluster analysis, fifth ed. John Wiley & Sons, Ltd, London.
Fu, K., Albus, J., 1977. Syntactic Pattern Recognition. Springer-Verlag.
García-Escudero, L.A., Gordaliza, A., 1999. Robustness properties of k means and trimmed k means. J. Amer. Statist. Assoc. 94 (447), 956–969.
García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., 2010. A review of robust clustering methods. Adv. Data Anal. Classif. 4, 89–109.
Gilbert, E.N., 1959. Random graphs. Ann. Math. Stat. 30 (4), 1141–1144.
Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 857–871.
Guha, S., Rastogi, R., Shim, K., 1999. ROCK: A robust clustering algorithm for categorical attributes. In: Data Engineering, 1999. Proceedings., 15th International Conference on. IEEE, pp. 512–521.
Heiser, W., Groenen, P., 1997. Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. Psychometrika 62 (1), 63–83.
Hu, T., Sung, S., 2006. A hybrid EM approach to spatial clustering. Comput. Statist. Data Anal. 50 (5), 1188–1205.
Hüllermeier, E., Rifqi, M., Henzgen, S., Senge, R., 2012. Comparing fuzzy partitions: A generalization of the Rand index and related measures. IEEE Trans. Fuzzy Syst. 20 (3), 546–556.
Hwang, H., Desarbo, W., Takane, Y., 2007. Fuzzy clusterwise generalized structured component analysis. Psychometrika 72 (2), 181–198.
Kaufman, L., Rousseeuw, P., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. WileyBlackwell, ISBN: 0471735787.
Krishnapuram, R., Joshi, A., Nasraoui, O., Yi, L., 2001. Low-complexity fuzzy relational clustering algorithms for web mining. IEEE Trans. Fuzzy Syst. 9 (4), 595–607.
McBratney, A., Moore, A., 1985. Application of fuzzy sets to climatic classification. Agricult. Forest. Meterol. 35 (1–4), 165–185.
Páez, A., Scott, D., 2005. Spatial statistics for urban analysis: A review of techniques with examples. GeoJournal 61, 53–67.
Torabi, M., 2016. Hierarchical multivariate mixture generalized linear models for the analysis of spatial data: An application to disease mapping. Biom. J. 58 (5), 1138–1150.
Viroli, C., 2011. Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comput. 21 (4), 511–522.
Wedel, M., Kamakura, W., 2000. Market Segmentation: Conceptual and Methodological Foundations, vol. 8, Springer.
Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 13 (8), 841–847.
Yeung, D.S., Wang, X., 2002. Improving performance of similarity-based clustering by feature weight learning. IEEE Trans. Pattern Anal. Mach. Intell. 24 (4), 556–561.