IRIS - Luiss Guido Carli Research Institutional Repository | Catalogo istituzionale della ricerca Luiss Guido Carli

Geographical clustering of the electoral flow between the Italian general elections of 2018 and 2022

Post-print version of the following publication: | Versione post-print della seguente pubblicazione:

Original Citation/Citazione:
D'Urso, Pierpaolo; De Giovanni, Livia; Federico, Lorenzo; Vitale, Vincenzina. (2026). Geographical clustering of the electoral flow between the Italian general elections of 2018 and 2022. SPATIAL STATISTICS, (ISSN: 2211-6753), 73: 1-8. Doi: 10.1016/j.spasta.2026.100977.

Availability/Disponibilità:
This version is available at: 11385/262480 since: 2026-05-24T15:43:54Z - Questa versione è disponibile alla pagina: 11385/262480 dal: 2026-05-24T15:43:54Z

Publisher/Casa editrice:
Elsevier

Published version/Pubblicato:
DOI: https://dx.doi.org/10.1016/j.spasta.2026.100977

License/Licenza:
Attribution-NonCommercial-NoDerivatives 4.0 International

Availability/Termini d'uso:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. Works made available under a Creative Commons license can be used according to the terms and conditions of said license. For all terms of use and more information see the publisher's website. | I termini e le condizioni relativi al riutilizzo della presente versione della pubblicazione sono disciplinati dalla politica editoriale. Le opere messe a disposizione con licenze Creative Commons possono essere utilizzate conformemente ai termini e alle condizioni previste da tali licenze. Per l'insieme delle condizioni di utilizzo e per ulteriori informazioni si rinvia al sito web dell'editore.

This item was downloaded from IRIS Luiss (https://iris.luiss.it/). When citing, please refer to the published version. | Questo documento è stato scaricato da IRIS Luiss (https://iris.luiss.it/). Per la citazione, fare riferimento alla versione pubblicata sul sito dell'editore.

(Article begins on next page | Il contributo inizia nella pagina successiva)

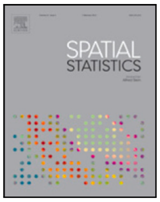
31 May 2026

Spatial Statistics 73 (2026) 100977

Contents lists available at ScienceDirect

Spatial Statistics

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



Geographical clustering of the electoral flow between the Italian general elections of 2018 and 2022☆

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

[a] Department of Social Science and Economics, Sapienza University, Piazzale Aldo Moro, 5, Rome, 00185, Lazio, Italy
[b] Department of Political Sciences, LUISS University, Viale Romania 32, Rome, 00197, Lazio, Italy

ARTICLE INFO

Keywords:
Robust Fuzzy C-medoids
Flow matrices
Italian general election
Spatial constraints

ABSTRACT

This paper aims to group Italian regions based on their similarity in voting flows between the major political coalitions during the 2018 and 2022 General Elections. We employ a Fuzzy C-Medoids clustering approach, using a tailored dissimilarity measure that captures the differences between matrices while also addressing the presence of outliers to mitigate their impact. The objective function integrates spatial constraints via fuzzy modularity, enabling the model to consider spatial relationships among regions. The methodology is applied to the ITANES dataset, a panel sample of 4696 respondents revealing interesting patterns, most notably a clear dichotomy between northern and southern Italy.

1. Introduction

Electoral flows refer to changes in voting behavior occurring across different elections, capturing shifts in party support, voter turnout, and abstention. Understanding these dynamics is essential for political scientists, sociologists, and policymakers seeking to analyze electoral volatility and voter loyalty.

In electoral studies, a significant body of literature has developed statistical methods to infer individual-level behaviors or relationships from aggregate-level data, such as precinct or district-level voting outcomes and demographic characteristics, a classic instance of the so-called ecological inference problem (Klima et al., 2015; Bracalente et al., 2026, and the references therein). While ecological inference has also been widely applied to analyze vote transitions across elections, the use of clustering techniques remains relatively underexplored in this context. An exception is the work by Puig and Ginebra (2014, 2015) who provide an example in which clustering contributes to ecological inference.

Given the complexity of electoral data, we argue that clustering methods can offer powerful tools to uncover latent structures, such as grouping geographic areas, demographic segments, or individual voters that exhibit similar voting transitions. In particular, we argue that unsupervised learning methods, such as the one presented here, are ideally suited to analyze voting flows, as they uncover latent patterns in complex transition matrices without relying on predefined categories.

More broadly, clustering in electoral studies has focused mainly on grouping voters based on stated preferences, often using mixture models (Kuriwaki, 2020; Jang and Hitchcock, 2022; Lee et al., 2017) or algorithms like k-means (Akarca and Başlevent, 2011; Wahyuni et al., 2023). From a spatial perspective, however, the dominant approach continues to rely on spatial autocorrelation statistics, such as Moran's I and Local Indicators of Spatial Association (LISA; Anselin (1995)), to detect geographic patterns in

---

☆ This article is part of a Special issue entitled: 'SPASTA_Electoral Studies' published in Spatial Statistics.
* Corresponding author.
E-mail address: lfederico@luiss.it (L. Federico).

https://doi.org/10.1016/j.spasta.2026.100977
Received 11 August 2025; Received in revised form 20 March 2026; Accepted 22 March 2026

electoral support (Rezaee et al., 2024; Jacintho et al., 2021; Shin and Agnew, 2007). For interesting works employing LISA in unsupervised approaches, see Appice and Malerba (2014), Mattera (2022).

In this work, we use a fuzzy clustering technique in which the objects to be clustered are not, as usual, vectors but matrices — specifically, the transition matrices representing voter flows from the 2018 Italian general election to the 2022 election for the Chamber of Deputies, regarding the main coalitions. The transition matrices have been estimated using an ad hoc survey conducted by ITANES, designed to be representative of the target population based on stratification variables and appropriate sample weights.

Moreover, given the presence of autonomous regions, which exhibit distinct voting transition patterns due to the influence of local parties, we implement a robust clustering technique based on a suitable exponential transformation of the dissimilarity measure, as first proposed by Wu and Yang (2002) to mitigate potential disruptions caused by anomalous units. This way, we ensure that regions with anomalous electoral dynamics do not affect the groups' structure. To this purpose, it is worth mentioning the work of Yamin et al. (2022) proposing an unsupervised density-based clustering algorithm to detect potential anomalies in election data, applied to the 2020 U.S. presidential election in Fulton County, Georgia.

The spatial relationships between objects are also considered; building on prior advances in fuzzy clustering with spatial constraints, a penalty function is integrated into the objective function (Pham, 2001; Coppi et al., 2010; Oriona et al., 2022; D'Urso et al., 2023, 2025; Cerqueti et al., 2025). An exogenous parameter is used to weight the importance of spatial contiguity in the clustering criterion. The higher the value of this tuning parameter, the more significant the role of spatial proximity in defining the clusters. In this work, following Cangemi et al. (2025), the characterization of spatial penalty is derived from the fuzzy modularity measure, borrowed from community detection (Nepusz et al., 2008). Although not applied to matrices, the work of Wu (2022), which uses spatial fuzzy C-means clustering to analyze the factors related to COVID-19 and the U.S. presidential election in the Rustbelt states in 2020, represents the most closely related work. From a methodological standpoint, the work of D'Urso et al. (2022) aligns most closely with the proposed approach. We also argue that LISA and spatially constrained clustering both utilize a contiguity matrix to define adjacency, but they pursue distinct analytical aims. LISA-based analysis computes a local test statistic (e.g., Local Moran's I or Getis–Ord Gi) for each spatial unit and uses permutation tests to identify statistically significant hot-spots, cold-spots, and spatial outliers. In contrast, our spatial clustering approach applies a contiguity constraint to group all units into a fixed number of homogeneous, spatially connected clusters by optimizing within-cluster similarity. Thus, LISA pinpoints where local autocorrelation is significant, while our method provides complete partitioning of the spatial units into coherent clusters.

The structure of the paper is organized as follows. In Section 2.1, we describe the ITANES dataset; Section 2.2 introduces a dissimilarity measure specifically designed for flow matrices; and Section 2.3 outlines our methodological framework, including the proposed model and algorithm. Section 3 reports the application and the main findings. Finally, the concluding remarks are presented in Section 4.

2. Methodology

In this section, we begin by presenting the ITANES dataset—its sampling design and pre-processing steps—then introduce the dissimilarity measure for flow matrices and outline the proposed clustering model.

2.1. The ITANES survey

ITANES dataset (Vezzoni et al., 2023c) was collected around the 2022 Italian General Elections. Italy is a parliamentary republic, where in the general election, a new parliament is voted in. A government is formed afterwards by a prime minister who has to secure a majority in both chambers of parliament based on either pre- or post-election coalitions between parties (while in theory possible, no single party has ever obtained a majority on its own in both chambers since the establishment of the republic). A total of $n = 4696$ Italian citizens were administered a pre- and a post-election survey, with several questions about their vote in the general elections of 2018 and 2022, together with several further questions about their socioeconomic status, political preferences and opinions on the most relevant current news topics at the time of survey (Sept–Oct 2022) such as the CoViD pandemic and ongoing conflicts. Full description of the survey can be found in Vezzoni et al. (2023a).

Our goal is to understand the geographic patterns in the electoral flows, so the information from this dataset that we use in this study is

- Vote declaration in the 2018 general elections, as registered in column `D10_3` of the database.
- Vote declaration in the 2022 general elections as registered in column `Q10LHa` (whether they voted) and `Q10LHb` (what they voted) of the database.
- Region of residence of the polled electors as registered in column `regione` of the database.
- Trimmed weights based on socioeconomic parameters to correct the imbalances in the sample (column `wsociodem_trim` in the database, see Vezzoni et al. (2023b) for the details).

In particular, in our model, we aim at clustering $K = 20$ statistical units, corresponding to the 20 Italian Regions (NUTS 2 level units and the largest administrative divisions in the country), based on both their geographic contiguity and the similarity of the electoral flows between the general elections of 2018 and 2022.

The geographic contiguity is encoded in the adjacency matrix $\mathbf{A} = \{a_{kl}, l, k \leq 20\}$ where $a_{kl} = 1$ if regions k and l share a land border and 0 otherwise.

To represent the electoral flows, we built for each region a 4 × 4 *flow matrix* $\mathbf{F}^{(k)}, k = 1, \dots, 20$.

Between the general elections of 2018 and 2022 a significant shift happened in Italian politics. The 2018 elections resulted in a hung parliament, with neither of the three major players, the center-right and center-left coalitions, and the Movimento 5 Stelle (5 Star Movement), an anti-establishment party, being able to secure an outright majority in either of the chambers. As is common in the Italian parliamentary system, this resulted in the rearrangement of coalitions after the election to secure the majority needed in both chambers for a vote of confidence in the new government. Between 2018 and 2022, three different governments were sworn in, the first two led by Giuseppe Conte and the third by Mario Draghi as prime ministers, every time with significant shifts in the composition of the majority supporting the government. Finally, in 2022, a snap election was called as it was no longer feasible to find an agreement to form a new majority. The 2022 elections resulted in a very different outcome, with the center-right coalition obtaining a decisive victory, securing a majority in both chambers, and Fratelli d'Italia (Brothers of Italy), previously a minor party, emerging as the dominant force within the winning coalition. The leader of Fratelli d'Italia, Giorgia Meloni, was sworn in as the first female Prime Minister in the history of the Italian Republic without the need for a post-election agreement. Here, the first 3 rows and columns correspond to the 3 major coalitions that ran in both elections:

CS Center-left coalition. Partito Democratico, and +Europa in 2018, Partito Democratico, Impegno Civico, +Europa, and Alleanza Verdi Sinistra in 2022.

CD Center-right coalition. Forza Italia, Lega, and Fratelli d'Italia in 2018, Forza Italia, Lega, Noi Moderati, and Fratelli d'Italia in 2022.

5S Movimento 5 stelle, running on its own in both 2018 and 2022.

The fourth row and column, labeled **NO**, represented electors who did not cast a valid vote. This includes both the electors who did not participate at all in the election, and those who cast a blank or invalid ballot. This category does not include electors who did not disclose their vote or voted for secondary coalitions, which were not considered in the analysis. This is because in the highly volatile Italian election system, where many parties are created and dissolved between consecutive elections, adding a residual category would be confounding, as it could have very different meanings in different elections, and adding all the secondary parties and coalitions would make the flow matrices way too sparse for a reliable analysis.

2.2. Dissimilarity for transition matrices

Given a sample of n units (interviewed individuals), each with an associated sample weight w_s such that $\sum_{s=1}^{n} w_s = n$, let t_1 and t_2 denote two distinct time points (in this application, t_1 and t_2 are the 2018 and 2022 general election, respectively). Let $\mathcal{G}$ be the set of possible categories (political parties, in this specific application), and let k, for $k = 1, \dots, K$, index the objects to be clustered (the Italian regions, in this case). For each region k, the entry $f_{ij}^{(k)}$ denotes the weighted proportion of individuals who were in category $i \in \mathcal{G}$ at time t_1 and transitioned to category $j \in \mathcal{G}$ at time t_2. More formally,

$$f_{ij}^{(k)} = \frac{\sum_{s=1}^{n} w_s\, I_{s,i,t_1}\, I_{s,j,t_2}}{\sum_{s=1}^{n} \sum_{j' \in \mathcal{G}} w_s\, I_{s,i,t_1}\, I_{s,j',t_2}},$$

where I_{s,i,t_1} (respectively, I_{s,j,t_2}) is an indicator function equal to 1 if unit s belonged to category $i \in \mathcal{G}$ at time t_1 (respectively, to category $j \in \mathcal{G}$ at time t_2), and 0 otherwise.

In cases where the formula results in an indeterminate form—because there were no elements from that category at time t_1—we set $f_{ij}^{(k)} = 0$. This ensures that each row is either entirely zero or sums to 1, representing the weighted distribution of transitions between categories across time.

Let $\mathbf{F}^{(k)} = \left(f_{ij}^{(k)}\right)$ denote the $|\mathcal{G}| \times |\mathcal{G}|$ matrix associated with object k, to quantify the dissimilarity between two flow matrices $\mathbf{F}^{(k)}$ and $\mathbf{F}^{(l)}$, we define their L_1 distance as:

$$d_1(\mathbf{F}^{(k)}, \mathbf{F}^{(l)}) = \sum_{i,j} \left| f_{ij}^{(k)} - f_{ij}^{(l)} \right|. \tag{1}$$

This way, if both matrices have no 0-lines, they are both row-stochastic and the distance is proportional to the sum of the total variation distances of the distributions in each line. If instead a matrix contains a line which is all 0s, said line contributes the same value to the distance of the unit from all the other units for which that line is non-0.

The robust version of Eq. (1) is based on the following exponential transformation (Wu and Yang, 2002; D'Urso et al., 2015) that lies in $[0, 1]$:

$$d_{exp}^2(\mathbf{F}^{(k)}, \mathbf{F}^{(l)}) = 1 - \exp\{-\beta \cdot d_1(\mathbf{F}^{(k)}, \mathbf{F}^{(l)})^2\}, \tag{2}$$

$$\text{with} \quad \beta = \frac{K}{\min_k \sum_l d_1(\mathbf{F}^{(k)}, \mathbf{F}^{(l)})^2}.$$

The main advantage of this approach is that it helps to mitigate the impact of large distances, potentially preventing outliers or distant objects from influencing the clustering procedure. The parameter β is typically chosen as the inverse of a measure of data variability. Therefore, when the data exhibit low variability (corresponding to a high value of β), larger dissimilarities are assigned lower weights compared to situations with high variability (for further details, refer to Wu and Yang (2002)). This way, whenever $d_1(\mathbf{F}^{(k)}, \mathbf{F}^{(l)})$ is very large, we have $d_{exp}^2(\mathbf{F}^{(k)}, \mathbf{F}^{(l)}) \sim 1$.

The most important characteristic to note is that the use of $d^2_{\text{exp}}(\mathbf{F}^{(k)}, \mathbf{F}^{(l)})$ results in outliers being assigned to the C clusters with approximately equal membership degrees (around $1/C$), effectively treating them as fuzzy units.

In the application to Italian data, this is particularly relevant for handling the regions of Trentino-Alto Adige, and Valle d'Aosta, where the presence of parties representing local linguistic minorities skews heavily the results and, to a lesser extent, Molise and Basilicata, where the amount of data available was quite low.

2.3. Fuzzy C-Medoids clustering with modularity spatial correction

The Fuzzy C-Medoids clustering with modularity-based spatial correction (FCMd-MSC) introduced in Cangemi et al. (2025) is here extended to incorporate robustness and a dissimilarity measure specifically tailored for complex objects such as flow matrices. The adoption of a fuzzy clustering approach is motivated by the nature of the data. Some regional electoral flows cannot be meaningfully assigned to a single, mutually exclusive group, as they often exhibit ambiguous or mixed behavioral patterns that share characteristics with more than one cluster. The fuzzy framework accommodates this complexity by allowing partial membership across clusters, thereby enhancing both the flexibility of the model and the interpretability of the results. In this setting, the inherent vagueness of cluster membership is explicitly modeled rather than ignored, leading to a more informative and realistic representation of regional heterogeneity. The FCMd-MSC method was created to bring together fuzzy entropic clustering and community detection on networks. Its goal is to produce a fuzzy partition matrix $\mathbf{U} := \{u_{k,c} : k \le K, c \le C\}$ where $u_{k,c}$ represents the level of membership of unit k to cluster c, together with a set of C prototype units $(\tilde{\mathbf{F}}^{(1)}, \dots, \tilde{\mathbf{F}}^{(C)})$, called medoids, which represent the "most typical" unit in each cluster. To this end, it employs the dissimilarity measure d^2_{exp} in conjunction with an adjacency matrix $\mathbf{A}$, which, in this context, represents a spatial weight matrix. The entries of $\mathbf{A}$ are binary (0–1), indicating whether pairs of units share a land border (1) or not (0). Specifically, the method aims to optimize a linear combination of two components: the objective function of fuzzy entropic clustering applied to the units using the chosen dissimilarity, and the fuzzy modularity, defined in the same manner as in Nepusz et al. (2008). he fuzzy modularity is computed from the partition matrix $\mathbf{U}$ and the network structure encoded by the adjacency matrix $\mathbf{A}$.

To this purpose, we introduce the modularity matrix $\mathbf{B} := \{b_{k,l}, k, l \le K\}$:

$$b_{k,l} = a_{k,l} - \frac{s_k s_l}{L} \tag{3}$$

where $s_k = \sum_{l=1}^K a_{k,l}$, and the total strength of the network as $L = \sum_{k=1}^K s_k$.

We can thus define, up to a constant, the fuzzy modularity of the partition $\mathbf{U}$ with respect to the network $\mathbf{A}$ as

$$Q(\mathbf{U}, \mathbf{A}) = \sum_{k=1}^K \sum_{l=1}^K b_{k,l}(1 - \delta_{k,l}) \sum_{c=1}^C u_{k,c} u_{l,c}. \tag{4}$$

Here, $\delta_{k,l}$ is the Kronecker δ, the indicator that $k = l$.

The full optimization problem is then:

$$\begin{aligned}
\min_{\mathbf{U},\mathbf{F}_c} J_{p,C,\gamma}(\mathbf{U}, \mathbf{F}_c) := {} & (1-\gamma) \sum_{k=1}^K \sum_{c=1}^C u_{k,c} d^2_{exp}(\mathbf{F}^{(k)}, \tilde{\mathbf{F}}^{(c)}) + p \sum_{k=1}^K \sum_{c=1}^C u_{k,c} \log(u_{k,c}) \\
& - \frac{\gamma}{2} \sum_{k=1}^K \sum_{c=1}^C \sum_{l=1}^K u_{k,c} b_{k,l} u_{k,c} (1 - \delta_{k,l}) \\
= {} & (1-\gamma) \sum_{k=1}^K \sum_{c=1}^C u_{k,c} \left(1 - \exp\{-\beta \cdot d_1(\mathbf{F}^{(k)}, \tilde{\mathbf{F}}^{(c)})^2\}\right) \\
& + p \sum_{k=1}^K \sum_{c=1}^C u_{k,c} \log(u_{k,c}) \\
& - \frac{\gamma}{2} \sum_{k=1}^K \sum_{c=1}^C \sum_{l=1}^K u_{k,c} b_{k,l} u_{k,c} (1 - \delta_{k,l})
\end{aligned} \tag{5}$$

Here, $\gamma \in [0, 1]$ is a parameter that represents the relative importance of the dissimilarity between electoral flows and geographic contiguity, and $p > 0$ is the tuning parameter that regulates the weight given to the Shannon entropy of the partition and thus its desired degree of fuzziness (higher values of p correspond to fuzzier partitions). If $\gamma = 0$, the algorithms considers only the attributes of the units, ignoring the adjacency structure, while for $\gamma = 1$ it considers only the adjacency matrix, ignoring the attributes. The modularity term is divided by 2 to account for the fact that the adjacency between two units k and l is considered twice in the sum, once as b_{kl} and once as b_{lk}. We find the local minimum of the objective function by iteratively optimizing with respect to the membership matrix $\mathbf{U}$ and the medoids $(\tilde{\mathbf{F}}^{(1)}, \dots, \tilde{\mathbf{F}}^{(C)})$. For the first step, we apply the Lagrange multiplier method. The optimization with respect to the medoids is done by linear search restricted for each cluster to the units with the corresponding maximal membership. The full details of the algorithm can be found in Cangemi et al. (2025, Section 2.2) where the Fuzzy C-Medoids clustering with modularity spatial correction was originally defined. While there is no guarantee that the algorithm converges to the global minimum, under basic conditions on the convexity of the objective function (cfr. Hathaway and Bezdek (1988)) the algorithm converges to a local minimum. To increase the chances of finding the global optimum, we restart the algorithm multiple times and choose the best solution.

Table 1
Values of the fuzzy silhouette for $\gamma = 0.15$, for different values of C and p. Optimal value is for $C = 2$, $p = 0.3$.

C/p	0,1	0,2	**0,3**	0,4	0,5
2	0,068	0,077	**0,087**	0,083	0,083
3	−0,113	−0,12	−0,108	−0,094	−0,084
4	−0,123	−0,126	−0,125	−0,118	−0,129

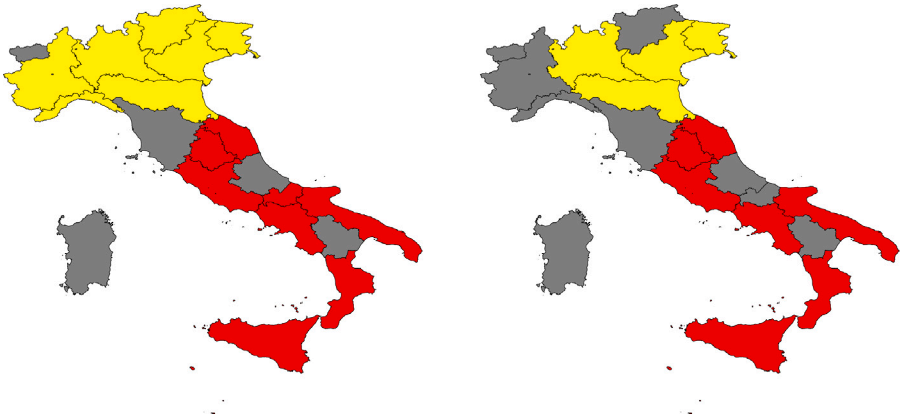


Fig. 1. Map of the cluster memberships with $\gamma = 0.15$ (left) and $\gamma = 0$ (right). Yellow indicates cluster 1, red cluster 2 and gray fuzzy membership (no membership above 0.7).

3. Application and results

We run the clustering algorithm described in Section 2.3 on the data as described in Section 2.1.[1] We fix the value of the parameter $\gamma = 0.15$ that regulates the relative importance between the spatial information and the distance between flow matrices. This is done because, even if in Cangemi et al. (2025) was given a method for the optimization of the clustering with respect to γ based on an appropriate validity function, in this setting, the most relevant information is the one relative to the electoral flows, and the spatial information represents the geography of Italy, which is well known and not a point of interest of the present analysis, but only a correction term. We thus do not think it would be wise to run an optimization with respect to γ, considering that even if the geography-based communities were better-separated than the attribute clusters, optimizing the matrix $\mathbf{U}$ mainly around them would be of little interest. We use the fuzzy silhouette as a validity measure to optimize a posteriori the values of the number of clusters C and the fuzziness parameter p. We choose, based on the results in Table 1, $C = 2$ and $p = 0.3$.

We next show the results of the clustering algorithm applied with the optimal values of C and p in Table 2 and compare it with the model with the same hyperparameters, except with $\gamma = 0$, that is, no spatial correction.

In Fig. 1, which shows the clusters on a map, we observe that both models reveal the well-known north–south divide commonly seen in analyses of Italian social, economic, and political phenomena. However, as expected, the spatially corrected model yields geographically more coherent clusters, with fewer fuzzy units

In the partition based on the spatial constraint, regions such as Abruzzo, Basilicata, and Molise exhibit stronger memberships to the southern cluster compared to the model with no spatial correction, where their memberships are more ambiguous. This suggests that geographical proximity plays a role in refining cluster assignments. In both models, the medoid of the northern cluster is Lombardia and that of the southern cluster is Lazio. Table 3 presents the corresponding flow matrices for these two medoids, and the associated alluvial plots are displayed in Fig. 2.

Looking at them, clear north–south differences in voting flows exist. We observe that in the two medoid regions, the electoral flows for voters who voted for the CD coalition are similar, with most of them voting again for the same coalition. Lombardia shows a much higher flow of voters from the 5S to the other two coalitions than Lazio, even if in both 5S is the group that lost the most voters to both other parties and abstention. On the other hand, Lazio has a much higher flow of CS voters to the other coalitions. Indeed, the 2022 elections saw a big success for the CD coalition, which obtained an absolute majority of the seats in both chambers of parliament. The progressive vote was split between the CS coalition, which was more popular in the northern regions, and the 5S, which was more popular in the south. These patterns reflect broader regional trends: northern areas have stronger party loyalty, while southern areas show more protest-vote persistence and higher volatility in traditional party support.

---

[1] We remark that here, "flow" denotes aggregate voter transitions between political options across two elections at the regional level, not individual voting trajectories. Accordingly, the proposed clustering reflects similarities in the structural patterns of electoral change between regions rather than individual-level dynamics.

Table 2
Results of the clustering algorithm with $C = 2$, $p = 0.3$.

	$\gamma = 0.15$		$\gamma = 0$	
	C1	C2	C1	C2
Abruzzo	0,352	0,648	0,657	0,343
Basilicata	0,343	0,657	0,615	0,385
Calabria	0,174	0,826	0,160	0,84
Campania	0,105	0,895	0,266	0,734
Emilia Romagna	0,927	0,073	0,812	0,188
Friuli Venezia Giulia	0,984	0,016	0,979	0,021
Lazio	**0.000**	**1.000**	**0.000**	**1.000**
Liguria	0,816	0,184	0,699	0,301
Lombardia	**1.000**	**0.000**	**1.000**	**0.000**
Marche	0,116	0,884	0,159	0,841
Molise	0,194	0,806	0,481	0,519
Piemonte	0,813	0,187	0,533	0,467
Puglia	0,145	0,855	0,247	0,753
Sardegna	0,612	0,388	0,631	0,369
Sicilia	0,119	0,881	0,087	0,913
Toscana	0,502	0,498	0,657	0,343
Trentino Alto Adige	0,836	0,164	0,538	0,462
Umbria	0,053	0,947	0,084	0,916
Valle D'Aosta	0,575	0,425	0,495	0,505
Veneto	0,988	0,012	0,958	0,042

Table 3
Flow matrices of the two medoids.

2018\2022	Lombardia				Lazio			
	CS	CD	5S	No	CS	CD	5S	No
CS	0,864	0,019	0,021	0,095	0,714	0,082	0,092	0,112
CD	0,037	0,851	0,029	0,083	0,017	0,829	0,042	0,111
5S	0,168	0,242	0,432	0,158	0,060	0,140	0,585	0,216
No	0,119	0,142	0,065	0,674	0,076	0,115	0,172	0,637

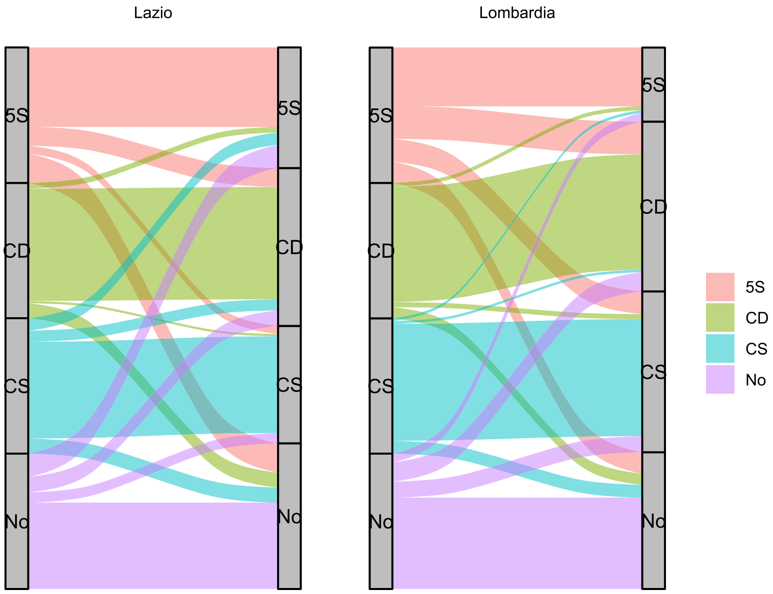


Fig. 2. Alluvial plot for the medoids Lazio and Lombardia.

4. Conclusions

The proposed approach provides a coherent framework for clustering electoral transition patterns based on matrix-valued data. By combining dissimilarity between flow matrices, entropy-based fuzziness, and spatial modularity, the model captures different aspects of the data structure while ensuring robustness to extreme patterns. The empirical application shows that allowing for fuzzy memberships is particularly appropriate in this context, as electoral flows often exhibit overlapping characteristics that cannot be adequately represented by crisp partitions. In addition, the incorporation of spatial information improves the geographical coherence of the clusters without altering the main electoral patterns. Overall, the results highlight the usefulness of the proposed framework in identifying meaningful and spatially consistent structures in electoral flow matrices. The main findings of the analysis can be summarized as follows:

- Electoral flow between the 2018 and 2022 Italian general elections exhibit clear spatial heterogeneity, with a north–south pattern emerging from the clustering results
- The analysis of the flow matrices shows that abstention represents a relevant component of voter transitions, particularly for voters originating from the Five Star Movement
- The proposed spatially-constrained clustering framework captures the complexity of electoral dynamics by allowing for overlapping patterns across regions through fuzzy memberships

5. Discussion

The proposed framework provides a structured approach for analyzing electoral flow and its spatial structure. However, it is important to emphasize that the method is not designed to identify the causal mechanisms underlying voter transitions as clustering is inherently an exploratory technique. While our method does not explicitly incorporate additional covariates and is not intended to directly explain why electoral flows occur, the spatial patterns identified in the analysis appear consistent with well-established territorial structures in Italian electoral behavior. The results reaffirm the well-documented north–south divide in Italian politics, highlighting distinct electoral dynamics across macro-regions that may be related to structural socio-economic differences, varying levels of political engagement, and territorially specific political traditions. Alluvial plots of the medoid regions, in particular, reveal how voter retention and transitions differ markedly between coalitions, highlighting a pronounced north–south divide in Italy's electoral dynamics. In Lombardia, for instance, the wide diagonal flows for both the Center-Right and Center-Left blocs indicate strong voter loyalty to traditional party blocs, whereas the Five Star Movement shows substantial outflows, both to the major coalitions and to abstention. By contrast, in Lazio, the Five Star Movement holds on to a larger share of its base, and the Center-Left loses more voters to the Center-Right than it does in Lombardia. These patterns reflect longstanding socio-political divides. Northern regions benefit from strong voter loyalty to large traditional coalitions and economic stability, while southern regions display a higher persistence of protest-vote alignments (Five Stars Movement) and greater volatility in traditional center-left support, consistent with deeper disengagement from politics. As already highlighted, the use of a robust dissimilarity measure has also played a crucial role in preserving the main electoral parties within the clustering results. By mitigating the influence of extreme patterns, it ensures that the main political alignments remain well-represented. This further strengthens the interpretation of the clusters, allowing for a clearer distinction between different voting behaviors while maintaining consistency with known spatial electoral patterns. However, the current modeling framework does not allow us to disentangle the relative impact of different factors that may influence voting behavior, such as economic developments, institutional dynamics, or external shocks (e.g., the COVID-19 pandemic). Similarly, the model does not explicitly account for the determinants of party support or abstention which are treated here as components of the observed flow. A more comprehensive analysis of the mechanisms driving electoral flows would require extending the proposed framework in several directions. In particular, this would involve integrating socio-economic and contextual covariates at the regional level, combining aggregate data with individual-level survey information and adopting dynamic approaches that consider multiple elections over time. Indeed, the model is designed with the study of electoral flow between two consecutive elections but it would be interesting to design a model that can account for a longer time series of electoral flows. This would require tackling significant new challenges, like the fact that, in the very unstable Italian electoral system, parties and coalitions can change significantly over time, and thus the flow matrices could have significantly different shapes. Future research could also explore alternative spatial weighting strategies to refine the balance between geographical and electoral information. Also, we used the L_1 distance between matrices due to its connection to total variation distance, but other distances could be considered, such as the L_2 distance induced by the Frobenius norm.

References

Akarca, A.T., Başlevent, C., 2011. Persistence in regional voting patterns in Turkey during a period of major political realignment. Eur. Urban Reg. Stud. 18, 184–202.

Anselin, L., 1995. Local indicators of spatial association—lisa. Geogr. Anal. 27, 93–115. http://dx.doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Appice, A., Malerba, D., 2014. Leveraging the power of local spatial autocorrelation in geophysical interpolative clustering. Data Min. Knowl. Discov. 28, 1266–1313. http://dx.doi.org/10.1007/s10618-014-0349-0.

Bracalente, B., Forcina, A., Falocci, N., 2026. Determinants of vote transitions by ecological inference within small areas. Spat. Stat. 71, 100950.

Cangemi, D., D'Urso, P., De Giovanni, L., Federico, L., Vitale, V., 2025. Fuzzy clustering and community detection: an integrated approach. J. Classification 1–31.

Cerqueti, R., D'urso, P., Mattera, R., 2025. Fuzzy group fixed-effects estimation with spatial clustering. AStA Adv. Stat. Anal. 1–32.

Coppi, R., D'Urso, P., Giordani, P., 2010. A fuzzy clustering model for multivariate spatial time series. J. Classification 27, 54–88.

D'Urso, P., De Giovanni, L., Federico, L., Vitale, V., 2023. Fuzzy clustering of spatial interval-valued data. Spat. Stat. 100764, http://dx.doi.org/10.1016/j.spasta.2023.100764.

D'Urso, P., De Giovanni, L., Federico, L., Vitale, V., 2025. Fuzzy clustering of mixed data with spatial constraints. Spat. Stat. 100874, http://dx.doi.org/10.1016/j.spasta.2024.100874.

D’Urso, P., De Giovanni, L., Massari, R., 2015. 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. Spat. Stat. 49, http://dx.doi.org/10.1016/j.spasta.2021.100518.

Hathaway, R.J., Bezdek, J.C., 1988. Recent convergence results for the fuzzy c-means clustering algorithms. J. Classification 5, 237–247.

Jacintho, L., da Silva, T.P., Parmezan, A.R.S., Batista, G., 2021. Analysing spatio-temporal voting patterns in Brazilian elections through a simple data science pipeline. J. Inf. Data Manag. 12.

Jang, J., Hitchcock, D.B., 2022. Model-based cluster analysis of democracies. J. Data Sci. 10, 297–319. http://dx.doi.org/10.6339/JDS.2012.10(2).1018.

Klima, A., Thurner, P.W., Molnar, C., Schlesinger, T., Küchenhoff, H., 2015. Estimation of voter transitions based on ecological inference: an empirical assessment of different approaches. AStA Adv. Stat. Anal. 100, 133–159.

Kuriwaki, S., 2020. A clustering approach for characterizing voter types: An application to high-dimensional ballot and survey data. http://dx.doi.org/10.31219/osf.io/v3rhz, OSF Preprints, Center for Open Science.

Lee, L., Zhang, S., Yang, V.C., 2017. Do two parties represent the us? clustering analysis of us public ideology survey. arXiv preprint arXiv:1710.09347.

Mattera, R., 2022. A weighted approach for spatio-temporal clustering of covid-19 spread in Italy. Spat. Spatio-Temporal Epidemiol. 41, 100500. http://dx.doi.org/10.1016/j.sste.2022.100500.

Nepusz, T., Petróczi, A., Négyessy, L., Bazsó, F., 2008. Fuzzy communities and the concept of bridgeness in complex networks. Phys. Rev. E—Stat. Nonlinear Soft Matter Phys. 77, 016107.

Oriona, A.L., D’Urso, P., Vilar, J., Lafuente, B., 2022. Spatial weighted robust clustering of multivariate time series based on quantile dependence with an application to mobility during covid-19 pandemic. IEEE Trans. Fuzzy Syst. 30, 3990–4004. http://dx.doi.org/10.1109/TFUZZ.2021.3136005.

Pham, D.L., 2001. Spatial models for fuzzy clustering. Comput. Vis. Image Underst. 84, 285–297.

Puig, X., Ginebra, J., 2014. A cluster analysis of vote transitions. Comput. Statist. Data Anal. 70, 328–344.

Puig, X., Ginebra, J., 2015. Ecological inference and spatial variation of individual behavior: National divide and elections in Catalonia. Geogr. Anal. 47, 262–282.

Rezaee, M., Alamdar, M., Taheri, E., Badiee Azandahi, M., 2024. Exploring spatial analysis of the voting patterns in the Afghanistan president elections of 2019. GeoJournal 89, 146. http://dx.doi.org/10.1007/s10708-024-11150-2.

Shin, M.E., Agnew, J., 2007. The geographical dynamics of Italian electoral change, 1987–2001. Elect. Stud. 26, 287–302.

Vezzoni, C., Basile, L., Garzia, D., Maggini, N., Mancosu, M., Paparo, A., 2023a. Itanes 2022 - descrizione dello studio_release 01.pdf. In: Itanes 2022 - Italian National Election Study 2022 (Release 01, July 2023). UNIMI Dataverse, http://dx.doi.org/10.13130/RD_UNIMI/JV77WR/JVI2V4.

Vezzoni, C., Basile, L., Garzia, D., Maggini, N., Mancosu, M., Paparo, A., 2023b. Itanes 2022 - istruzioni per i pesi_release 01.pdf. In: Itanes 2022 - Italian National Election Study 2022 (Release 01, July 2023). UNIMI Dataverse, http://dx.doi.org/10.13130/RD_UNIMI/JV77WR/5CDTKV.

Vezzoni, C., Basile, L., Garzia, D., Maggini, N., Mancosu, M., Paparo, A., 2023c. Itanes 2022 - Italian national election study 2022 (release 01, July 2023). http://dx.doi.org/10.13130/RD_UNIMI/JV77WR.

Wahyuni, S.N., Khanom, N.N., Astuti, Y., 2023. K-means algorithm analysis for election cluster prediction. JOIV: Int. J. Inform. Vis. 7, 1–6.

Wu, S., 2022. Spatial fuzzy c-means clustering analysis of U.S. presidential election and covid-19 related factors in the rustbelt states in 2020. Axioms 11, 401. http://dx.doi.org/10.3390/axioms11080401, URL: https://www.mdpi.com/2075-1680/11/8/401.

Wu, J., Yang, S.S., 2002. Alternative objective functions for fuzzy clustering: a review. Pattern Recognit. 35, 2863–2878. http://dx.doi.org/10.1016/S0031-3203(01)00226-0.

Yamin, K., Oswald, M., Jadali, N., Xie, Y., Zegura, E., Nazzal, D., 2022. An unsupervised density based clustering algorithm to detect election anomalies: Evidence from Georgia’s largest county. In: Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies. pp. 313–323.