WP 25/05

Spatial Clustering with Functional Trend Data: An Application to Italian Health Tax Detractions

Mauro Marè; Francesco Porcelli & Francesco Vidoli

May 2025

Spatial clustering with functional trend data: an application to Italian health tax detractions

Mauro Marè[a], Francesco Porcelli[b], Francesco Vidoli[c]

[a] *Tuscia University and Luiss Business School*
[b] *University of Bari*
[c] *University of Urbino Carlo Bo, Department of Economics, Society, Politics, Via Aurelio Saffi, 42, Urbino 61029, PU Italy*

Abstract

The integration of functional and spatial data in clustering methods is an increasingly explored topic in regional and urban economics. In the analysis of regional patterns, it is often essential to identify clusters of production units that are not only similar in terms of their temporal evolution, but also geographically coherent. In this paper, we introduce a novel spatial algorithm, validated through both simulated and real-world data, aimed at achieving regional segmentation based on territorial behavior over time. The real-world data, focusing on health tax detractions at the municipal level in Italy, uncovers distinct geographical clusters, within which the level and trend of detractions display significant similarity.

Keywords: Spatial clustering, Functional data analysis, Tax expenditures, Health economics
JEL Codes: C38, H51, H71

1. Introduction

Throughout the history of analytical thought, many philosophers have emphasized that knowledge rarely stems from a single, absolute criterion. Instead, it often arises from a dynamic equilibrium among multiple dimensions. Aristotle, in his *Metaphysics*, recognized the necessity of reconciling empirical observation with general principles to construct reliable knowledge (Aristotle, 1924). In modern philosophy, Immanuel Kant argued that knowledge emerges from the interplay between sensory intuition and conceptual categories, suggesting a fundamental ”trade-off” between data and structure

(Kant, 1781).
More recently, in the philosophy of science, Kuhn (1962) questioned the linear progress of knowledge and emphasized the incommensurability between paradigms , while Feyerabend (1975) went further by rejecting methodological monism, advocating for epistemological pluralism and theoretical diversity.

These perspectives converge on the idea that knowledge involves the balancing of multiple, sometimes conflicting, criteria.

It is within this epistemological framework that the present study proposes an empirical spatial clustering algorithm that integrates two distinct yet complementary dimensions: spatial proximity and similarity in temporal trends. The objective is to identify territorially coherent clusters that account for both geographical closeness and shared dynamic behavior over time. In the authors' view, such an approach provides a valuable tool for investigating spatial phenomena where neither spatial nor temporal information alone suffices to capture the complexity of the underlying structure (see Anselin, 1995; Chavent et al., 2009).
In the analysis of regional economic dynamics, it is, in fact, often essential to identify clusters of Decision Making Units (DMUs) that are not only similar in terms of their temporal evolution, but also geographically coherent. Integrating spatial information into the clustering process enhances the interpretability and practical relevance of the results. Indeed, local and regional policies are typically implemented within geographic boundaries, and targeting spatially contiguous areas that exhibit common trends enables more effective and equitable policy design (LeSage and Pace, 2009).
Moreover, we often expect parts of a territory to evolve in similar ways due to spatial spillovers and shared economic fundamentals, a notion supported by Marshallian economics. According to Marshall (1890), firms and regions benefit from agglomeration economies advantages arising from geographic proximity, such as knowledge diffusion, labor market pooling, and shared infrastructure. These mechanisms create self-reinforcing dynamics that promote similar patterns of development across neighboring areas (Krugman, 1991; Glaeser et al., 1992). However, local constraints, such as cultural differences, historical trajectories, and varying social norms, can lead to divergence, even among geographically close regions (Rodríguez-Pose, 2013). These factors shape local preferences, institutional settings, and innovation capacities, thus influencing how and to what extent regions benefit from agglomeration forces.

We contribute to the existing literature in two ways: from a methodological perspective, we propose an original algorithm that identifies homogeneous territorial areas exhibiting similar dynamics by integrating clustering techniques, spatial constraints, and functional data. From an empirical standpoint, using novel panel data at a highly granular territorial level, we show that healthcare tax detractions are primarily associated with the economic capacity of the areas rather than with actual needs. In other terms, higher levels of healthcare tax detractions tend to reflect the economic capacity of individuals or areas, rather than the actual healthcare needs they face.

The remainder of the article is structured as follows. Section 2 reviews the issue of spatially constrained clustering as addressed in the literature, while Section 3 presents a detailed description of the proposed algorithm. The algorithm is subsequently tested on simulated data (Section 4) and on Italian data concerning healthcare tax detractions at the municipal level (Section 5). Section 6 concludes the paper with a set of policy recommendations.

2. Related Literature

Conventional clustering techniques, predicated solely on attribute similarity, often fall short in capturing the intricate spatial dependencies inherent in socioeconomic phenomena; the integration of spatial data addresses this limitation by explicitly incorporating geographical relationships into the clustering process.

In recent years, the integration of spatial data into clustering techniques has become an increasingly prominent topic within the fields of regional science and urban economics, as researchers seek to develop methods that reflect both attribute similarity and spatial proximity.

Several important methods in spatial clustering explicitly incorporate spatial contiguity constraints. The SKATER algorithm (Assunção et al., 2006) builds a minimum spanning tree based on attribute dissimilarities and partitions it while preserving spatial adjacency, yielding spatially connected and homogeneous clusters. REDCAP (Guo, 2008) combines hierarchical clustering with dynamic programming to optimize clusters under both attribute and contiguity constraints, making it ideal for regionalization tasks requiring compact and homogeneous regions. The Max-p algorithm (Duque et al., 2012) treats clustering as a constrained optimization problem, aiming to maximize the number of contiguous regions while satisfying minimum thresholds (*e.g.*, population), ensuring regional balance. Lastly, ClustGeo (Chavent et al., 2018) offers a hierarchical approach that blends attribute dissimilarities with geographic distances, allowing users to adjust the balance between spatial cohesion and attribute similarity via a tuning parameter.

All these methods are, however, based on static approaches, relying on cross-sectional data to identify spatial clusters. However, none of them explicitly account for temporal dynamics, and therefore cannot capture clusters that are homogeneous in terms of their evolution over time.

From an economic perspective, adopting a clustering approach - particularly one that incorporates spatial dimensions - is of significant importance. This methodology allows for the identification of regions or spatial units that share similar socio-economic characteristics, facilitating more targeted and effective policy interventions. In other terms, by accounting for spatial proximity and contiguity, spatial clustering helps capture the underlying geographical structure of economic phenomena, revealing patterns of demand for homogeneous policies, inequality, or spatial spillovers that might otherwise remain hidden.

Functional Data Analysis (FDA), thoroughly reviewed by Jacques and Preda (2014), represents a highly promising methodological framework, particularly suited for capturing and analyzing temporal dynamics. In this approach, individual observations are represented as smooth functions or curves, rather than discrete time points, allowing for the modeling of complex temporal evolutions. These functional trajectories can then be compared using a variety of functional distance measures, providing a nuanced understanding of similarities and differences in the shape, amplitude, and timing of the processes under study.

A growing body of research has focused on methods for clustering spatially dependent functional data, data where each observation is a function and is spatially located, highlighting the importance of simultaneously considering both temporal evolution and spatial dependence. Notable contributions include the work of Ignaccolo et al. (2008), who applied functional clustering to air quality monitoring networks, demonstrating how this approach can reveal coherent spatial patterns in environmental processes over time. Similarly, Romano et al. (2015) evaluated the performance of two clustering techniques for spatial functional data, emphasizing the role of spatial structure in improving cluster reliability and interpretability.

Secchi et al. (2013) introduced a Bagging Voronoi classifier framework for clustering spatial functional data, showcasing the effectiveness of ensemble learning techniques in enhancing robustness and accuracy in the presence of spatial variability. Meanwhile, Giraldo et al. (2012) proposed a hierarchical clustering approach that explicitly accounts for spatial correlation among functional observations, allowing for the detection of geographically coherent clusters that evolve similarly over time.

Building upon these earlier contributions, Abramowicz et al. (2017) advanced the methodology further by introducing the Bagging Voronoi K-Medoid Alignment (BVKMA) algorithm, which simultaneously addresses three key challenges in spatial functional data analysis: clustering, curve misalignment, and spatial dependence. This integrated approach enhances the interpretability and precision of the clustering process, particularly in applications such as environmental monitoring or climate reconstruction, where temporal shifts and spatial continuity are both critical.

Collectively, these methodologies underline the analytical and practical value of incorporating spatial and functional dimensions in clustering processes. They allow researchers and policymakers to identify not only groups of units with similar dynamic behaviors but also regions that form spatially contiguous and interpretable patterns. This dual consideration is especially beneficial in applications where spatial contiguity enhances the relevance of findings, such as regional policy design, environmental planning, and public health surveillance, ensuring that interventions can be tailored to the specific spatio-temporal characteristics of each cluster.

Nevertheless, while these approaches are undoubtedly valuable, they are often challenging to describe comprehensively and do not always make explicit the inherent trade-off between enforcing spatial contiguity and preserving dynamic homogeneity across clusters.

3. Methodological framework

Building on a critical assessment of the strengths and limitations of existing approaches, we propose an original algorithm that balances spatial constraints with similar dynamic patterns. The proposed method, compared to spatial clustering methods on functional data, aims to be simpler and easier to interpret, clearly highlighting the inherent trade-off in empirical applications between geographical proximity and similarity in dynamics.
The first necessary step concerns the calculation of the two distance matrices representing these dimensions of analysis:

- D_0: the **geographical distance matrix**, measuring the physical (spatial) distance between DMUs based on their geographical coordinates and

- D_1: the **functional distance matrix**, measuring dissimilarities in the temporal trends.

More in depth, the geographical distance matrix $D_0 \in \mathbb{R}^{n \times n}$ (where n is the number of DMUs) is computed for each DMU identified by a pair of coordinates $(\text{lat}_i, \text{lon}_i)$ using the pairwise Euclidean distances[1] between all DMUs:

$$D_0(i,j) = \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{lon}_i - \text{lon}_j)^2} \tag{1}$$

while the symmetric distance matrix $D_1 \in \mathbb{R}^{n \times n}$ is calculated in two steps: first, the trends are smoothed using a B-spline basis of m functions (Eilers and Marx, 1996):

$$y_i(t) \approx \sum_{k=1}^{m} c_{ik} \phi_k(t) \tag{2}$$

where $y_i(t)$ is the temporal trend for DMU i over time t, $\phi_k(t)$ are B-spline basis functions, and c_{ik} the estimated coefficients; by using B-spline estimates instead of trends based on raw data, in fact, the algorithm benefits from smoother and more flexible representations of temporal dynamics, reducing the impact of noise and enhancing the comparability across spatial units. As a second step, for each pair of DMUs (i,j), the algorithm computes the squared L^2 distance between their functional representation[2]:

$$D_1(i,j) = \|y_i(t) - y_j(t)\|_{L^2} = \sqrt{\int (y_i(t) - y_j(t))^2 \, dt} \tag{3}$$

Once it has been calculated D_1, the functional distance matrix quantifying differences in tax detraction trends between DMUs and D_0 the spatial distance matrix, both matrices[3] have to be converted into generalized inertia matrix using the formula outlined in equation (4) in order to align with

[1] Alternative distance measures are certainly possible.

[2] In this case as well, alternative distance measures, such as Dynamic Time Warping (DTW) or elastic metrics (Srivastava and Klassen, 2016), are clearly viable options.

[3] Each matrix is normalized by its maximum value to ensure comparable ranges.

Ward's clustering criterion (Ward, 1963) which minimizes within-cluster variance:

$$\delta_{ij} = \frac{D_{ij}^2}{2n} \tag{4}$$

At this point, the existence of two separate inertia matrices prompts the need to determine an appropriate method for combining them into a single, unified matrix. This process requires a careful balancing act that takes into account the particular goals and context of the specific application.

In other terms, the use of a mixing parameter allows analysts to control the trade-off between functional similarity and spatial proximity. This flexibility makes it possible to explore different clustering configurations, striking a balance between the internal homogeneity of clusters and their geographical consistency. As a result, spatial clustering with trend data proves particularly valuable in fields such as public health, urban planning, and territorial policy, where space and time jointly shape the dynamics of socio-economic phenomena.

Given these premises, the degree of influence from each matrix may be controlled by a mixing parameter $\alpha \in [0, 1]$ following this simple formula:

$$\delta_\alpha = (1 - \alpha) \cdot \delta_0 + \alpha \cdot \delta_1 \tag{5}$$

From an applied perspective, the optimal value of α can be identified as the point of balance between the relative loss or gain in explained inertia for each dissimilarity matrix, as α varies - each normalized to its respective best-case scenario.

In more formal terms, this means finding the α that best balances the contributions of both δ_0 and δ_1, by evaluating Q_0, the proportion of inertia explained by δ_0 for each $\alpha \in [0, 1]$, and Q_1, the proportion of inertia explained by δ_1. These values are then normalized : $Q_0^{norm} = Q_0/Q_0[\alpha = 0]$, which reflects how Q_0 evolves relative to its maximum (at $\alpha = 0$) and $Q_1^{norm} = Q_1/Q_1[\alpha = 1]$, which reflects how Q_1 evolves relative to its maximum (at $\alpha = 1$).

The optimal α is achieved when these normalized values are equal, indicating a balanced trade-off between the two sources of dissimilarity.

The resulting combined dissimilarity matrix δ_α is used as input for the hierarchical clustering algorithm using the Ward method (please see Murtagh and Legendre, 2014 for in-depth technical details), which aims to minimize the total within-cluster variance at each step of the agglomeration process.

Finally, the algorithm exhibits quadratic time complexity, *i.e.*, $\mathcal{O}(n^2)$ (please see Appendix B), which is consistent with the expected behavior of procedures involving pairwise operations over the set of units. While this may pose computational challenges for very large datasets, it remains manageable in many practical applications, especially considering the improved clustering accuracy and interpretability it offers.

4. Simulated data

To assess the behavior and properties of the proposed algorithm, four scenarios will be designed, ranging from the simplest to the most complex to identify. These scenarios will differ both in terms of spatial configuration, ranging from tightly clustered to more dispersed unit locations, and in the presence or absence of a shared underlying trend among the units.

4.1. Distinct proximity, distinct trends

In the first scenario, we generate a dataset consisting of two spatially distinct groups, Group A and Group B, each characterized by both geographical separation and divergent temporal trends. Each group contains 200 observational units, each assigned a unique set of geographical coordinates (latitude and longitude) according to the following specifications:

- **Group A**: Units are uniformly distributed in latitude between 43 and 45, and longitude between 11 and 13.
- **Group B**: Units are uniformly distributed in latitude between 45 and 46, and longitude between 13 and 14.

Each unit i in the generic group g is associated with a time series y_{it} observed yearly from 2010 to 2019. The data generation model is:

$$y_{it} = \alpha_i + \beta_g \cdot (t - 2010) + \varepsilon_{it}, \quad \text{for } t = 2010, \ldots, 2019 \tag{6}$$

where: $\alpha_i \sim \mathcal{U}(10, 50)$ is the unit-specific intercept; $\beta_g \sim \mathcal{U}(0.5, 1.5)$ for Group A (increasing trend), and $\beta_g \sim \mathcal{U}(-1.5, -0.5)$ for Group B (decreasing trend) and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$, with $\sigma = 2$, represents Normally distributed noise.
This setting, therefore, represents the more favorable scenario for the proposed algorithm, as the units are organized into two spatially distinct groups

(Figure 1, top left panel) that are internally homogeneous and clearly differentiated from one another in terms of their temporal trends (Figure 1, top right panel).

By applying the spatial clustering method based on trend similarity, as introduced in the previous section, we obtain a value of α very close to 0.5 (Figure 1, bottom right panel). This result indicates an optimal balance between geographic proximity and trend similarity in this specific case, leading to a correct classification of all units under analysis (Figure 1, bottom left panel). In this specific case, the correlation reaches a perfect value of 1, as all 400 units have been accurately and consistently imputed.

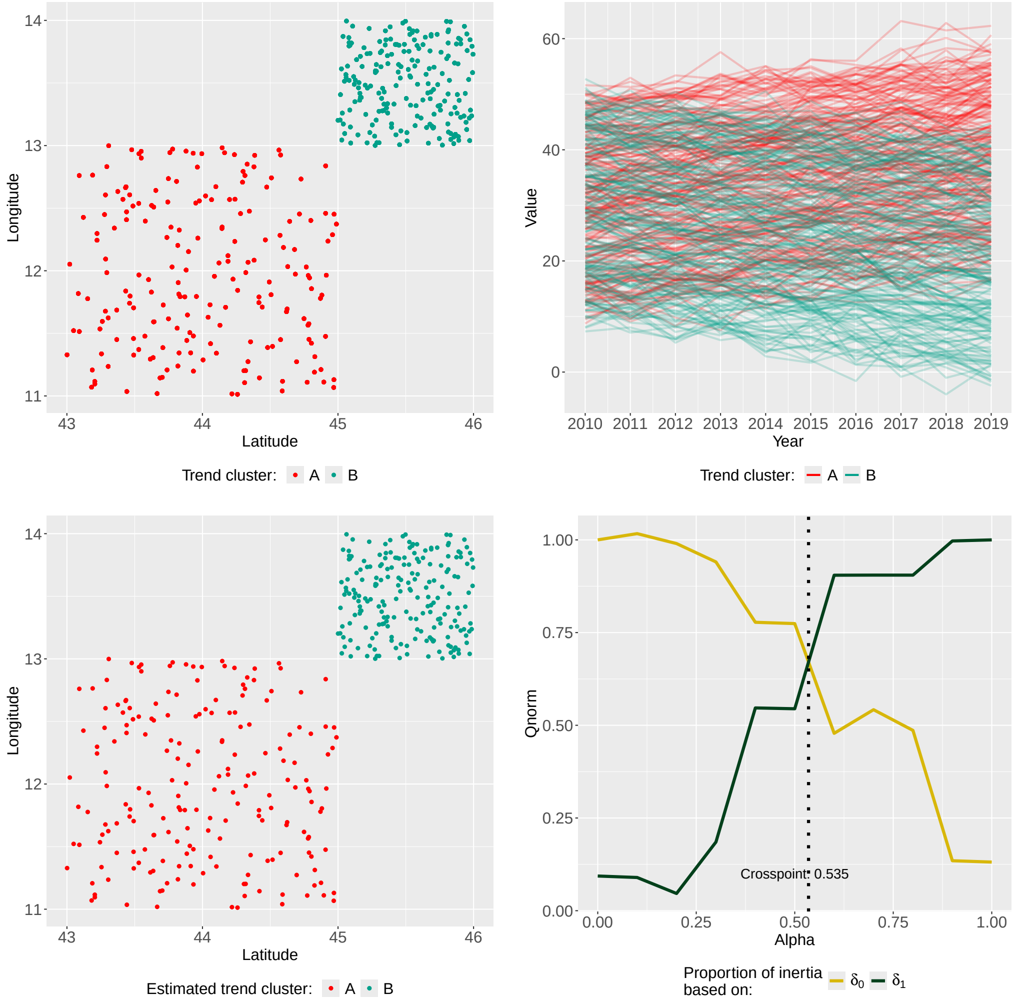


Figure 1: Simulation and estimation, distinct proximity and distinct trends case

4.2. Indistinct proximity, distinct trends

To assess the robustness and limitations of the proposed algorithm, it is essential to move beyond the more favorable scenario and systematically relax the baseline assumptions outlined in Subsection 4.1.

Let us now consider the case in which the units are not geographically distinct, or are only partially so (Figure 2, top left panel), while still exhibiting clear differentiation in their temporal dynamics (Figure 2, top right panel).

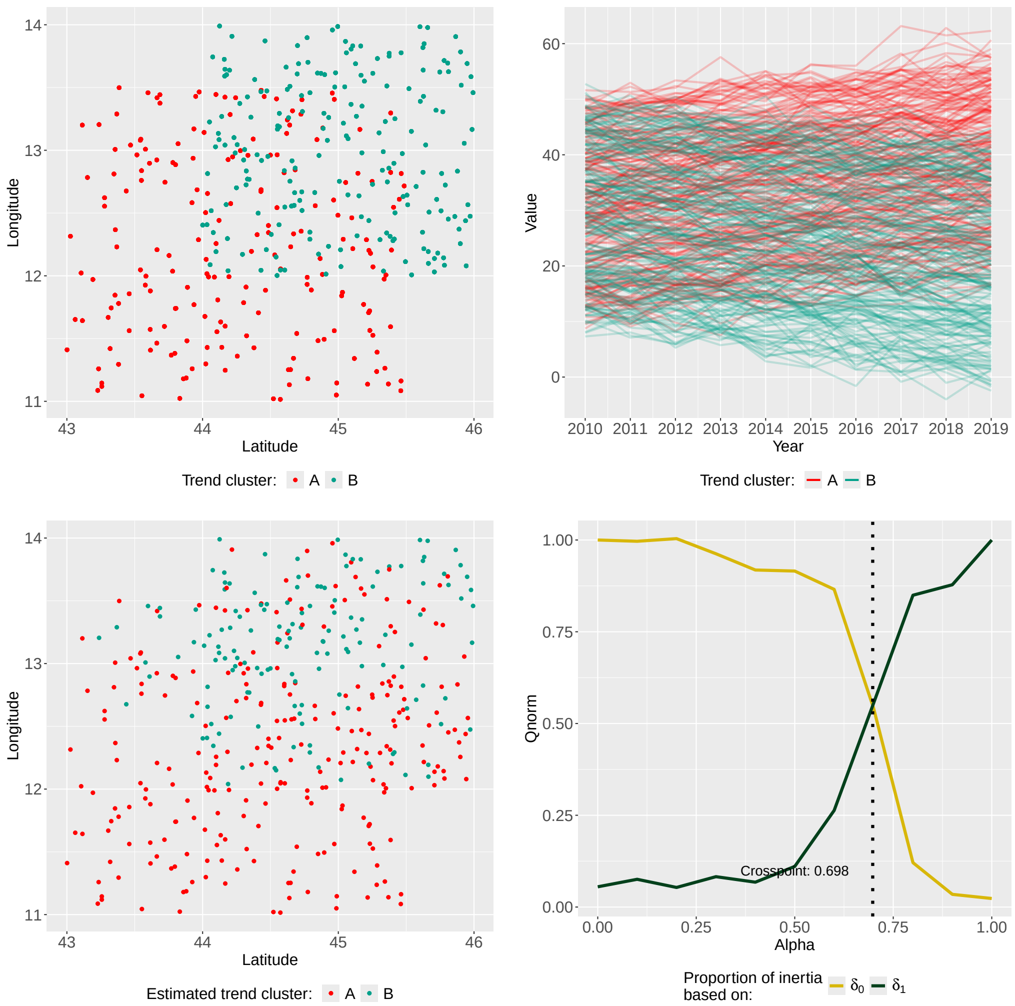


Figure 2: Simulation and estimation, indistinct proximity and distinct trends case

In this case, the optimal value of α, estimated at the intersection point of the inertia proportions as a function of α, is 0.698 (Figure 2, bottom right panel). According to equation (5), this value reflects a greater influence of temporal dynamics relative to spatial homogeneity in the clustering process, consistent with the data-generating mechanism.

In this case, the algorithm achieves a reasonably accurate assignment of units to their true clusters, with approximately 70% of the units correctly imputed, as shown in Figure 2, bottom left panel. The clustering result is further supported by a moderate positive correlation of 0.406, indicating a meaningful alignment between the estimated and actual groupings.

4.3. Distinct proximity, indistinct trends

The final scenario represents the most challenging case, one in which, despite the presence of spatially distinct and internally homogeneous groups (Figure 3, top left panel), the units do not exhibit any differences in their temporal dynamics (Figure 3, top right panel). In such a context, we do not expect to identify distinct clusters, as the similarity in trends across units correctly prevents the occurrence of differentiated clusters, even in the presence of a marked geographical partition.

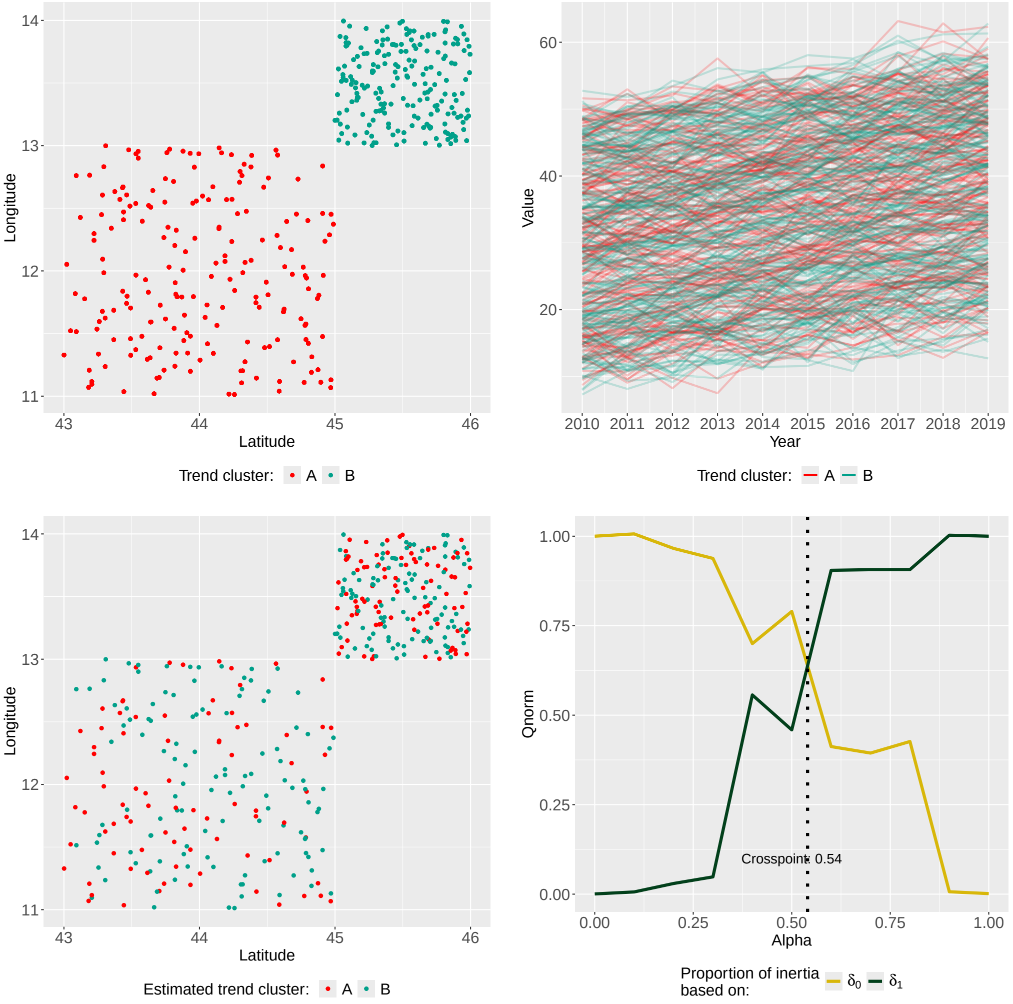


Figure 3: Simulation and estimation, distinct proximity and indistinct trends case

Figure 3, bottom left panel, shows a clustering result that closely approximates a random allocation, with a correlation of only 0.111 between real and estimated assignment. In this case, geographical proximity alone proves insufficient to identify clusters of units that are not only spatially close, but also have similar temporal dynamics.

5. Application to Italian health tax detractions

The evolution of health tax expenditures in Italy between 2017 and 2022 serves as the empirical starting point of this section, with the aim of verifying whether there are areas of the country that systematically differ from others in terms of their dynamics. This period is marked by a clear upward trend in the use of tax detractions, often employed, perhaps inappropriately, as tools of political economy. The significant and persistent increase, both in terms of the number of measures and their fiscal impact, has led tax expenditures to account for approximately 6.3% of foregone revenue relative to Italy's GDP (year 2022 , Marè et al., 2024).

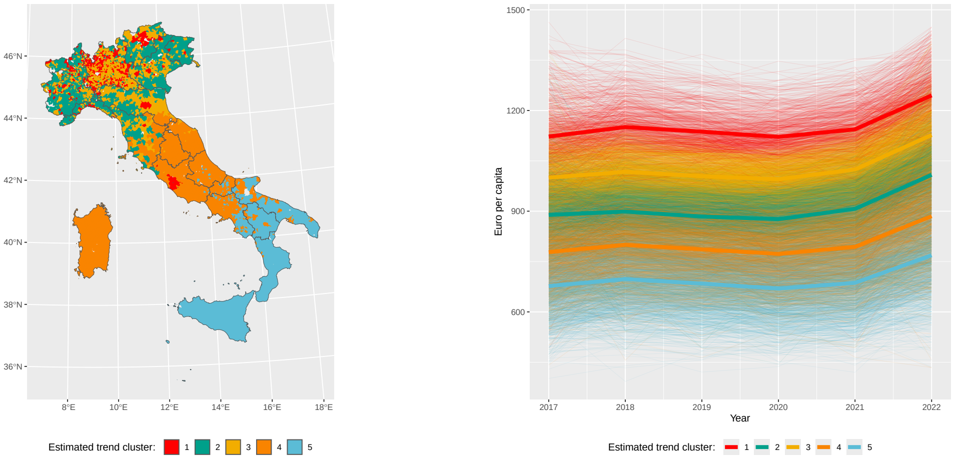
This increase poses a hidden challenge to the soundness of public finances: its fiscal implications are seldom made explicit in the budgetary process, which makes it all the more necessary to analyze, evaluate, and, where appropriate, reduce such measures.

Reining in the proliferation of tax expenditures or determining how to manage their continued growth is inherently difficult for several reasons. These include electoral sensitivities, the presence of multiple and often well-organized interest groups, the lack of clear justifications accompanying many detractions introduced in the past, and the absence of transparent criteria to assess the relative importance or priority of individual provisions (Surrey and McDaniel, 1985). Moreover, tax expenditures can raise concerns related to territorial equity. Since they are typically linked to income, they may disproportionately benefit wealthier individuals or more economically developed regions (OECD, 2010).

To investigate whether certain areas of the Italian territory are systematically advantaged or disadvantaged, both in terms of the level and dynamics of such benefits, this study relies on unpublished data provided by the Department of Finance of the Ministry of Economy and Finance at the municipal level. Specifically, the analysis focuses on personal income tax detractions for healthcare expenses (*i.e.*, the 19% detraction on the portion of medical and specific assistance expenses exceeding euro 129.11), covering the years 2017 to 2022.

The average levels of the per-capita tax detractions in the 7,637 Italian municipalities, for which data were available throughout the entire 2017-2022 period, were analyzed using the proposed algorithm, with the aim of identifying whether there existed significantly large areas of the territory in which detractions followed a similar trend, both in level and/or in shape.

After verifying a substantial balance between the spatial criterion and the dynamic one through the optimal α parameter (see Figure A.1), we analyzed the dendrogram derived from Ward's algorithm applied to the dissimilarity matrix δ_α, identifying five clusters.



(a) Map

(b) Single and mean trends

Figure 4: Estimated trend clusters

What emerges from the data clearly highlights the existence of multiple Italia's, revealing stark disparities in healthcare expenditure across income groups. The territories with higher incomes tend to spend significantly more on their health, while those with lower incomes spend considerably less with a clear North-South regularity. This pattern might appear natural if health were treated as a market commodity, accessible in proportion to one's financial means. However, the Italian Constitution enshrines the principle of universal healthcare, guaranteeing equal access to medical services regardless of income, emphasizing their nature as merit goods with significant positive externalities.

The data, therefore, expose a troubling contradiction: a reality in which access to health is, in practice, still heavily influenced by economic status, contrary to the foundational ideals of our welfare system.

Income levels and access to care, proxied by tax-deductible health expenditures, appear to be inversely related to health care needs, as illustrated in Table 1, which reports key health and socioeconomic indicators by estimated cluster. In other terms, where the need is greatest, due to the increasing shift towards out-of-pocket healthcare expenses, individuals are unable to afford care. The result is a higher cancer rate, lower life expectancy, and an increased number of people forgoing treatment.

Estimated cluster	Per capita income	Life expectancy at birth - Women	Standardized disease rate for cancer age 25–64	Number of medical examination drop-outs per capita
1	17,530	85.73	71.42	34.02
2	15,520	85.59	72.68	31.05
3	16,761	85.75	71.49	33.82
4	11,706	85.40	73.77	62.28
5	9,527	84.75	74.75	54.66

Table 1: Key health and socioeconomic indicators by estimated cluster

More in particular, in fact, Clusters 1 and 2, encompassing the wealthiest urban centers and Lombardy, show the highest socioeconomic status and health outcomes, with Cluster 1 at the top. Cluster 3, representing peripheral northern regions, maintains a strong socioeconomic profile but with some health deterioration. In contrast, Cluster 4, corresponding to Central Italy, sees a notable decline in both income and health indicators, while Cluster 5, representing Southern Italy, exhibits the greatest disadvantages across all dimensions, marked by the lowest income, shortest life expectancy, and highest health risks.

Finally,the territorial clusters also appear to be clearly delineated, with a few notable exceptions that may, at least in part, reflect the presence or absence of private healthcare providers. Particularly striking is the distinct position of Rome within the Lazio region, as well as the case of Lombardy, which emerges as the region with the highest growth in per capita healthcare expenditure. Overall, the average dynamics of expenditure are more sharply differentiated in terms of magnitude rather than in their underlying patterns or trends.

6. Concluding remarks

This paper introduces an original spatial clustering algorithm that integrates both geographical proximity and temporal trend similarity, providing a nuanced framework for analyzing regional dynamics. By applying this methodology to healthcare tax detractions in Italian municipalities, we uncover distinct territorial patterns that reflect deep-rooted socio-economic disparities.

The findings reveal that healthcare tax benefits are not uniformly distributed across the Italian territory. Rather, they align closely with the economic capacity of the population, raising important questions about equity in access to healthcare-related fiscal support. While the algorithm successfully delineates coherent regional clusters, the results expose a troubling contradiction: despite constitutional principles that guarantee universal access to healthcare, financial capacity remains a key determinant of tax benefit utilization.

The growing fiscal weight of such measures calls for mechanisms that assess not only their economic impact but also their distributive consequences. Targeting tax relief more effectively, particularly in disadvantaged areas, may help reconcile fiscal policy with the principles of fairness and territorial cohesion. In other terms, the analysis highlights how seemingly neutral financial instruments, such as tax detractions, can inadvertently reinforce territorial inequality, exacerbating pre-existing regional disparities; this underscores the urgent need to recognize spatial injustices and to ensure that regional policies are designed with a territorial lens, especially when such disparities run counter to foundational principles of equity.

The simulation exercises confirm the robustness and flexibility of the proposed algorithm, too, demonstrating its ability to adapt to different spatial and temporal configurations and to effectively disentangle the relative contributions of proximity and dynamic similarity in the clustering process.

Finally, the methodological approach proposed here offers a replicable tool for analyzing spatial-temporal phenomena across various domains, from public health to education and infrastructure. Its capacity to balance spatial contiguity with trend similarity opens new avenues for both academic research and evidence-based policy design.

References

Abramowicz, K., Arnqvist, P., Secchi, P., Luna, S.S.d., Vantini, S., Vitelli, V., 2017. Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction. Stochastic Environmental Research and Risk Assessment 31, 71–85.

Anselin, L., 1995. Local indicators of spatial association—lisa. Geographical Analysis 27, 93–115.

Aristotle, 1924. Metaphysics. Oxford University Press.

Assunção, R.M., Neves, M.C., Câmara, G., Freitas, C.C., 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science 20, 797–811.

Chavent, M., Kuentz, V., Labenne, A., Saracco, J., 2018. Clustgeo: an r package for hierarchical clustering with spatial constraints. Computational Statistics 33, 1799–1822.

Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J., 2009. Clustgeo: an r package for hierarchical clustering with spatial constraints. Computational Statistics 34, 2099–2120.

Duque, J.C., Anselin, L., Rey, S.J., 2012. The max-p-regions problem. Journal of Regional Science 52, 397–419.

Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with b-splines and penalties. Statistical science 11, 89–121.

Feyerabend, P., 1975. Against Method. Verso.

Giraldo, R., Delicado, P., Mateu, J., 2012. Hierarchical clustering of spatially correlated functional data. Statistica Neerlandica 66, 403–421.

Glaeser, E.L., Kallal, H.D., Scheinkman, J.A., Shleifer, A., 1992. Growth in cities. Journal of Political Economy 100, 1126–1152.

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

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

Jacques, J., Preda, C., 2014. Functional data clustering: A survey. Advances in Data Analysis and Classification 8, 231–255.

Kant, I., 1781. Critique of Pure Reason. Macmillan.

Krugman, P., 1991. Increasing returns and economic geography. Journal of Political Economy 99, 483–499.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press.

LeSage, J.P., Pace, R.K., 2009. Introduction to Spatial Econometrics. Statistics: A Series of Textbooks and Monographs, CRC Press, Boca Raton.

Marshall, A., 1890. Principles of Economics. Macmillan, London.

Marè, M., Porcelli, F., Vidoli, F., 2024. Does private supply drive personal health choices? A spatial approach of health tax detractions at municipal level. Technical Report. HEDG - Health Econometrics and Data Group, University of York. York, UK.

Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? Journal of Classification 31, 274–295.

OECD, 2010. Tax Expenditures in OECD Countries. OECD Publishing, Paris.

Rodríguez-Pose, A., 2013. Do institutions matter for regional development? Regional Studies 47, 1034–1047.

Romano, E., Mateu, J., Giraldo, R., 2015. On the performance of two clustering methods for spatial functional data. AStA Advances in Statistical Analysis 99, 467–492.

Secchi, P., Vantini, S., Vitelli, V., 2013. Bagging voronoi classifiers for clustering spatial functional data. International journal of applied earth observation and geoinformation 22, 53–64.

Srivastava, A., Klassen, E., 2016. Functional and Shape Data Analysis. Springer, New York.

Surrey, S.S., McDaniel, P.R., 1985. Tax Expenditures. Harvard University Press, Cambridge, MA.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

Appendix A. Choice of mixing parameter α

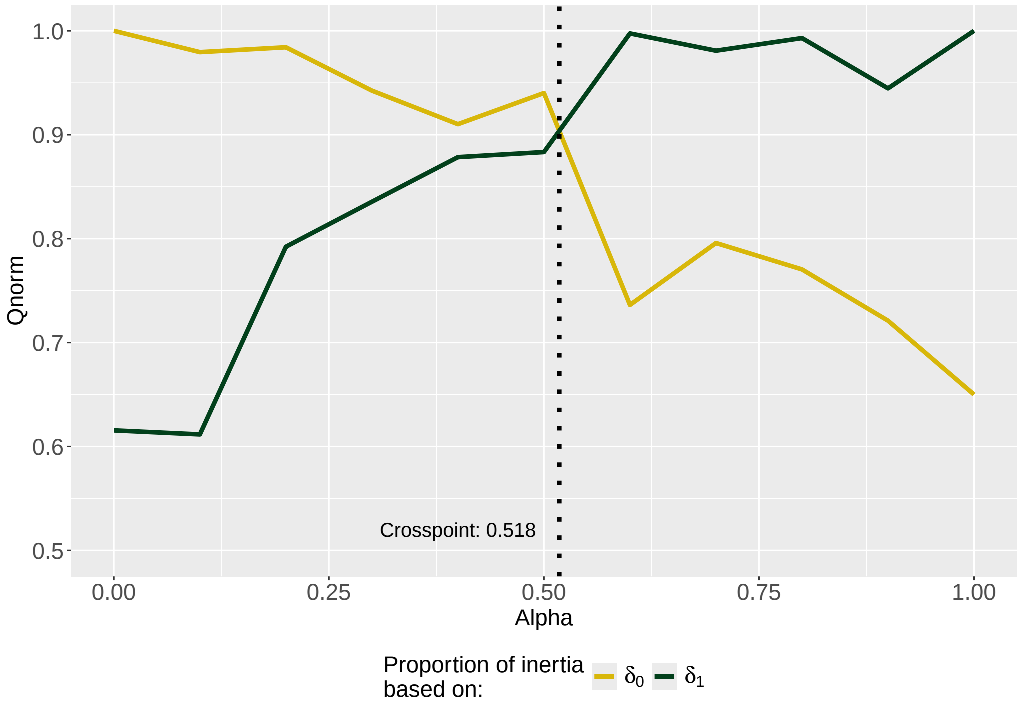


Figure A.1: Choice of mixing parameter α

Appendix B. Computational Complexity

To assess the computational scalability of the proposed algorithm, a set of simulation experiments was conducted by systematically varying the number of processed units. The observed execution times indicate a nonlinear relationship with the number of units.

A second-order polynomial regression model was fitted to the data:

$$\text{Duration}_i = \beta_0 + \beta_1 \cdot \text{units}_i + \beta_2 \cdot \text{units}_i^2 + \varepsilon_i, \tag{B.1}$$

where $\texttt{Duration}_i$ denotes the execution time corresponding to the i-th configuration. Estimation results show that the quadratic term is highly significant ($p < 0.001$), while the linear term is not statistically significant, suggesting a dominant quadratic growth pattern. The coefficient of determination is close to unity ($R^2 \approx 1$), indicating an excellent model fit.

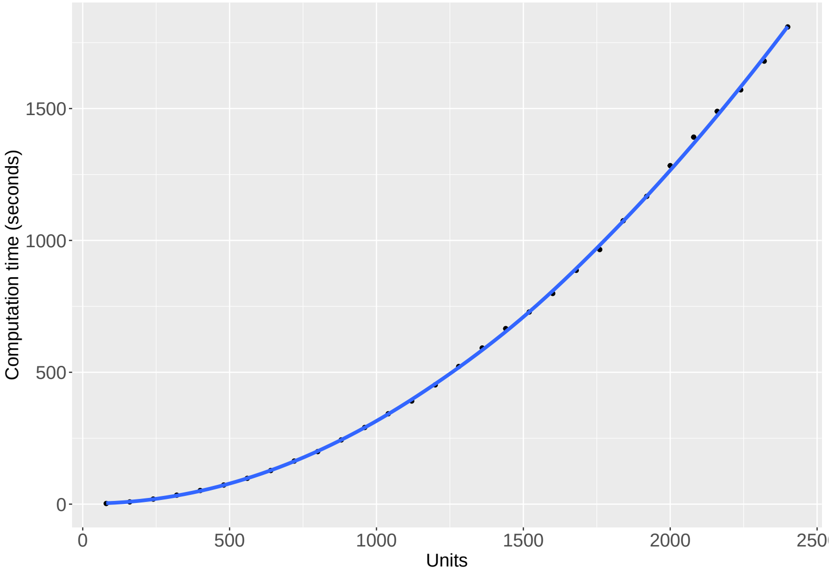


Figure B.2: Relationship between the number of units and computation time