Comovement in Geo-referenced Time Series: A Copula-Based Approach for Clustering

Alessia Benevento[1,*,†], Fabrizio Durante[1,†] and Roberta Pappadà[2,†]

[1]Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento, Lecce, Italy
[2]Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "B. de Finetti", Università degli Studi di Trieste, Trieste, Italy

Abstract

Time series clustering plays a crucial role in managing and extracting knowledge from the vast and complex Earth Observation (EO) datasets such as satellite-derived temperatures, precipitation levels, or soil-related variables. This study explores copula-based clustering techniques that focus on temporal dependence structures among time series, rather than their marginal behavior, to detect patterns of comovement in environmental variables. Applied to summer maximum temperatures in Italy, the approach reveals spatially coherent clusters that reflect underlying climatic regimes. However, when applied to monthly maximum precipitation data, clustering based solely on temporal dependence yields fragmented and geographically inconsistent results. To address this, we introduce a method that incorporates spatial proximity via soft constraints, combining temporal and spatial-based dissimilarities through a tunable mixing parameter. Our results demonstrate that including spatial information can significantly improve cluster coherence and interpretability, particularly for variables with strong geographic variability. Applications are based on EO data from the Copernicus Climate Data Store.

Keywords

Time Series Clustering, Spatio-Temporal Data, Dependence Modeling, Copula Models

1. Introduction

Time series clustering plays a crucial role in managing and extracting knowledge from the vast and complex Earth Observation (EO) datasets such as satellite-derived temperatures, precipitation levels, or soil-related variables. These data are often collected almost continuously over time and across thousands of spatial locations. The growing availability of high-resolution EO data provides unique opportunities for understanding complex environmental processes. However, this data ore often high-dimensional or spatially heterogeneous, presenting significant challenges for automated analysis and reuse.

In this context, time series clustering emerges as a key unsupervised learning strategy to manage complexity and extract meaningful structure from large-scale temporal datasets. By grouping time series with similar temporal behavior, clustering helps uncover regional patterns, reduce dimensionality, and support the design of interpretable models. This is particularly valuable in EO applications, where thousands of gridded time series must be analyzed jointly, such as temperature or precipitation over different locations.

One increasingly relevant direction in this area involves clustering based on cross-sectional dependence: identifying sets of time series that exhibit comovement, meaning they tend to increase or decrease together over time, even if their individual marginal behaviors differ. This type of dependence is particularly important for capturing joint climate dynamics, especially related to joint extremes, e.g., maxima of precipitations [1] and temperatures [2], or to model flood risks [3]. Copula-based clustering methods for time series (see, e.g., [4] and references therein) have naturally appeared in this context to

Workshop on AI-driven Data Engineering and Reusability for Earth and Space Sciences (DARES'25), co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025), Bologna, Italy, October 25, 2025
*Corresponding author.
†These authors contributed equally.
alessia.benevento@unisalento.it (A. Benevento); fabrizio.durante@unisalento.it (F. Durante); rpappada@units.it (R. Pappadà)
0009-0003-7346-4922 (A. Benevento); 0000-0002-4899-1080 (F. Durante); 0000-0002-4852-0561 (R. Pappadà)

focus exactly on the dependence among different time series regardless of the marginal behavior, rather than comparing raw values or global features.

These methods rely on a rank-invariant dissimilarity measure that quantifies how close the underlying copula is to the comonotonic case, representing perfect positive dependence.

Additionally, beyond time-depending features, EO data comes with different deterministic (spatial) information such as latitude, longitude or altitude. When general-purpose clustering methods are used for clustering geo-referenced time series, the resulting clusters are scattered over the spatial domain of the study. Thus, numerous studies have emphasized the importance of including spatial information for geographically referenced data since, in these scenarios, forming clusters that also reflect geographical proximity can significantly improve the interpretability of the results [5, 6].

In the copula framework, clustering methods with deterministic constraints have been proposed in [7] to incorporate non-temporal proximity information into the clustering process. Importantly, these methods do not enforce strict adherence to proximity constraints: the existing algorithms may cluster time series that are geographically distant if their dependence structure justifies it [8]. This flexibility arises from the use of soft proximity constraints [9], which contrast with hard constraints that require strict spatial coherence, as explored, e.g., in [10, 11].

Here, we present two applications of clustering with and without spatial information on climatological data downloaded from the Copernicus Climate Data Store[1].

2. Copula-based Clustering for Comovement Detection

Firstly, we show how to adopt a copula-based clustering approach that captures the dependence structure among variables to identify patterns of comovement in environmental time series.

We begin by considering a set of monthly maximum temperatures of the summer months (JJA) in Italy, derived from ERA5 reanalysis data, covering multiple spatial points across the Italian territory and spanning the period from 1960 to 2024. Although ERA5 provides a globally homogeneous grid, we restrict our attention to land areas by excluding sea points, selecting a subset of $n = 105$ grid points.

The aim is to cluster the relative time series not based on absolute temperature levels, but on how strongly their fluctuations are statistically dependent over time, for example, locations that tend to heat up simultaneously, even if the temperature magnitudes differ.

The approach proceeds in three main steps. First, the data are filtered by suitable univariate time series models, like seasonal ARIMA models, in order to remove the seasonality and the auto-correlation. The temperature series are then transformed into pseudo-observations, using their empirical marginal distributions. This operation removes the effect of differing scales or distributions and retains only the information relevant to dependence structure. From these pseudo-observations, we compute pairwise copulas for each pair of locations. The full collection of these bivariate copulas defines the Copula Matrix, a compact representation of the dependencies among all time series. An alternative strategy would be to rely on multivariate copulas to construct the Copula Matrix; however this approach may substantially increase computational complexity when dealing with large datasets.

Next, we define a dissimilarity measure over this matrix, which evaluates how far each pairwise copula deviates from the comonotonicity case, i.e., perfect positive dependence. This dissimilarity is rank-invariant, meaning that it is robust to monotonic transformations and unaffected by outliers or marginal variability. The resulting Dissimilarity Matrix is then used to perform Partitioning Around Medoids (PAM) clustering, yielding groups of locations whose time series exhibit similar comovement patterns. We present the results using $k = 8$ clusters, as this value represents a reasonable compromise between the optimal Average Silhouette Index [12] and the need for a clear and interpretable spatial visualization. Choosing $k = 8$ avoids overly complex maps with too many colors (which would hinder interpretation) while still offering more informative structure than overly simplistic solutions with only two or three regions. Although we have a finite set of representative stations, the maps are displayed

[1] https://cds.climate.copernicus.eu/

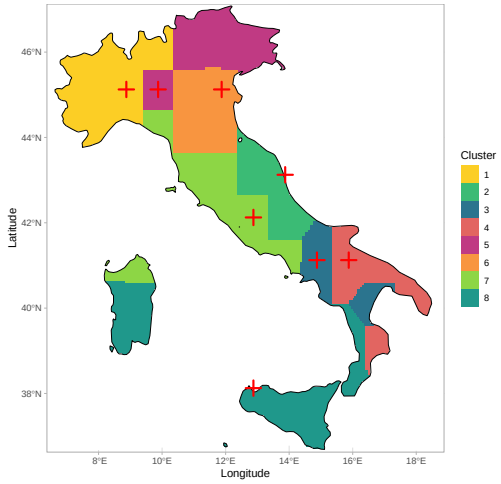

Figure 1: Cluster compositions of summer maximum temperatures. Stations marked with red crosses indicate the medoids of each cluster.

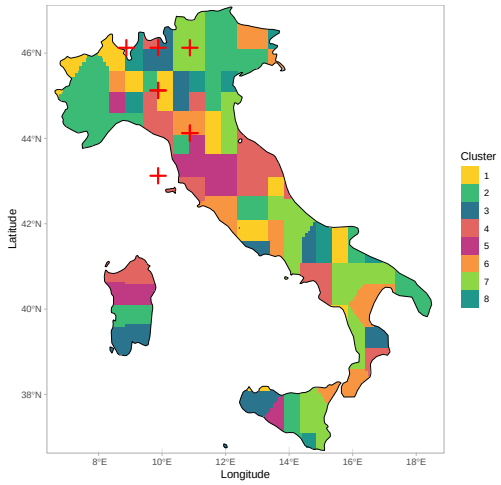

Figure 2: Cluster compositions of monthly maximum precipitations. Stations marked with red crosses indicate the medoids of each cluster.

using continuous colors to provide a full spatial representation across Italy. Each station is associated with a rectangular area (pixel), allowing the entire map to be fully covered.

Applied to the Italian dataset, this method reveals several meaningful clusters as shown in Fig.1. For instance, stations located in the Alps tend to cluster together, reflecting a shared temporal behavior likely driven by mountain air masses. Coastal and southern regions, influenced by different climatic regimes, form separate groups.

While the copula-based clustering algorithm provides coherent and geographically interpretable groups when applied to summer maximum temperature time series, the same methodology proves less effective when used with other variables, such as precipitation extremes. In particular, when clustering monthly maximum precipitations time series from January 2011 till November 2023 across the same $n = 105$ locations in Italy, the resulting clusters appear highly fragmented and spatially inconsistent as shown in Fig.2 where some stations in northwestern Italy are grouped together with others located in the southeastern part of the country, despite their clear geographic and climatic differences. Moreover, the fact that all cluster medoids are located in the northern regions further indicates that the clustering fails to capture the spatial heterogeneity of precipitation patterns across the country.

These results suggest that, for certain environmental variables, temporal dependence alone is insuf-

ficient to meaningfully characterize comovement structures. In such cases, it becomes necessary to explicitly incorporate spatial information into the clustering process. By doing so, we aim to balance dependence-driven similarity with geographic proximity, producing clusters that are both statistically coherent and spatially interpretable. This motivates the introduction of clustering with spatial constraints frameworks, where soft proximity constraints enforce spatial cohesion among time series. Given one matrix capturing temporal dependence among the time series and another representing spatial proximity, several approaches exist to merge these into a single dissimilarity matrix suitable for clustering [13, 14, 7]. Specifically, introducing a parameter α to balance the two dependencies ensures that the spatial constraint is imposed in a soft, rather than hard, form. Such α plays the role of a regularization parameter, calibrating the trade-off between temporal dependence and spatial proximity, and thereby influencing the number and shape of the resulting clusters. A common approach to merge the dissimilarities that does not leverage on copulas is, e.g., [9]. In the current copula-based framework, we adopt a convex combination controlled by the parameter α, where $\alpha = 0$ corresponds to purely temporal dependence and $\alpha = 1$ to purely spatial dependence. The clustering procedure relies on a dissimilarity measure defined as the distance between the copula-based dependence structure in the data and a target matrix M representing the perfect comonotonicity. This dissimilarity is itself constructed as a convex combination of temporal and spatial components, in line with the strategy proposed in [7]. The tuning parameter α thus acts as a weight that regulates the relative contribution of the two sources of information, balancing dependence-driven similarity with geographic proximity. In Figure 3, we illustrate how gradually incorporating the spatial component into the dependence model leads to increasingly compact clusters.

The selection of the optimal parameter α is far from straightforward. Both its selection and the determination of the optimal number of clusters K can significantly influence the resulting clusters, and therefore deserve careful consideration.

Acknowledgments

AB and FD have been supported by MUR-PRIN 2022 PNRR, Project "Stochastic Modeling of Compound Events" (No. P2022KZJTZ) funded by European Union – Next Generation EU. The work of FD has been carried out with partial financial support from ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by EU – Next Generation EU (CUP F83C22000740001). RP has been supported by MUR-PRIN 2022, Project "Modelling Non-standard data and Extremes in Multivariate Environmental Time series" (No. 20223CEZSR) funded by European Union – Next Generation EU.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] E. Bernard, P. Naveau, M. Vrac, O. Mestre, Clustering of maxima: Spatial dependencies among heavy rainfall in France, J. Clim. 26 (2013) 7929–7937.

[2] M. Bador, P. Naveau, E. Gilleland, M. Castellà, T. Arivelo, Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe, Weather Clim. Extremes 9 (2015) 17–24.

[3] R. Pappadà, F. Durante, G. Salvadori, C. De Michele, Clustering of concurrent flood risks via hazard scenarios, Spat. Stat. 23 (2018) 124–142.

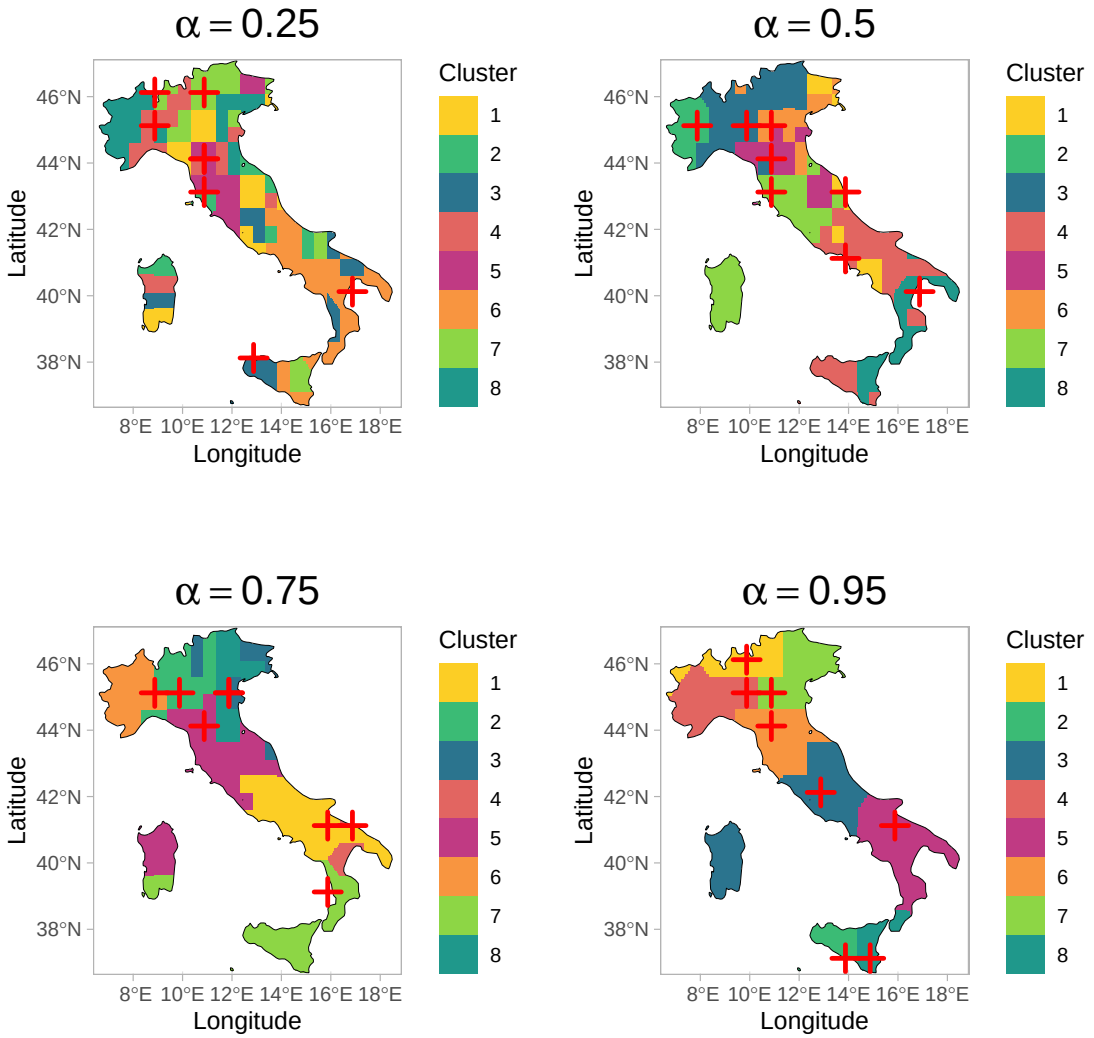

Figure 3: Cluster compositions of monthly maximum precipitations. The parameter α controls the influence of the spatial component, with higher values assigning increasing importance to spatial proximity.

[4] F. M. L. Di Lascio, A. Menapace, R. Pappadà, A spatially-weighted amh copula-based dissimilarity measure for clustering variables: An application to urban thermal efficiency, Environ. 35 (2024) e2828.

[5] F. Fouedjio, Clustering of multivariate geostatistical data, Wiley Interdiscip. Rev.: Comput. Stat. 12 (2020) e1510.

[6] K. Kopczewska, Spatial machine learning: new opportunities for regional science, Ann. Reg. Sci. 68 (2022) 713–755.

[7] M. Disegna, P. D'Urso, F. Durante, Copula-based fuzzy clustering of spatial time series, Spat. Stat. 21 (2017) 209–225.

[8] T. Romary, F. Ors, J. Rivoirard, J. Deraisme, Unsupervised classification of multivariate geostatistical data: Two algorithms, Comput. & Geosci. 85 (2015) 96–103.

[9] M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco, ClustGeo: an R package for hierarchical clustering with spatial constraints, Comput. Stat. 33 (2018) 1799–1822.

[10] Y. Pawitan, J. Huang, Constrained clustering of irregularly sampled spatial data, J. Stat. Comput. Simul. 73 (2003) 853–865.

[11] G. Guénard, P. Legendre, Hierarchical clustering with contiguity constraint in R, J. Stat. Softw.

103 (2022) 1–26.
[12] C. Hennig, M. Meila, F. Murtagh, R. Rocci, Handbook of cluster analysis, Chapman Hall/CRC Handb. Mod. Stat. Methods, Boca Raton, FL: CRC Press, 2016.
[13] A. Benevento, F. Durante, R. Pappadà, Tail-dependence clustering of time series with spatial constraints, Environ. Ecol. Stat. 31 (2024) 801–817.
[14] M. de Carvalho, R. Huser, R. Rubio, Similarity-based clustering for patterns of extreme values, Stat 12 (2023) e560.