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

Alessia Benevento, Fabrizio Durante

Dipartimento di Matematica e Fisica

“Ennio De Giorgi”

Università del Salento

Roberta Pappadà

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche

“B. de Finetti”

Università degli Studi di Trieste

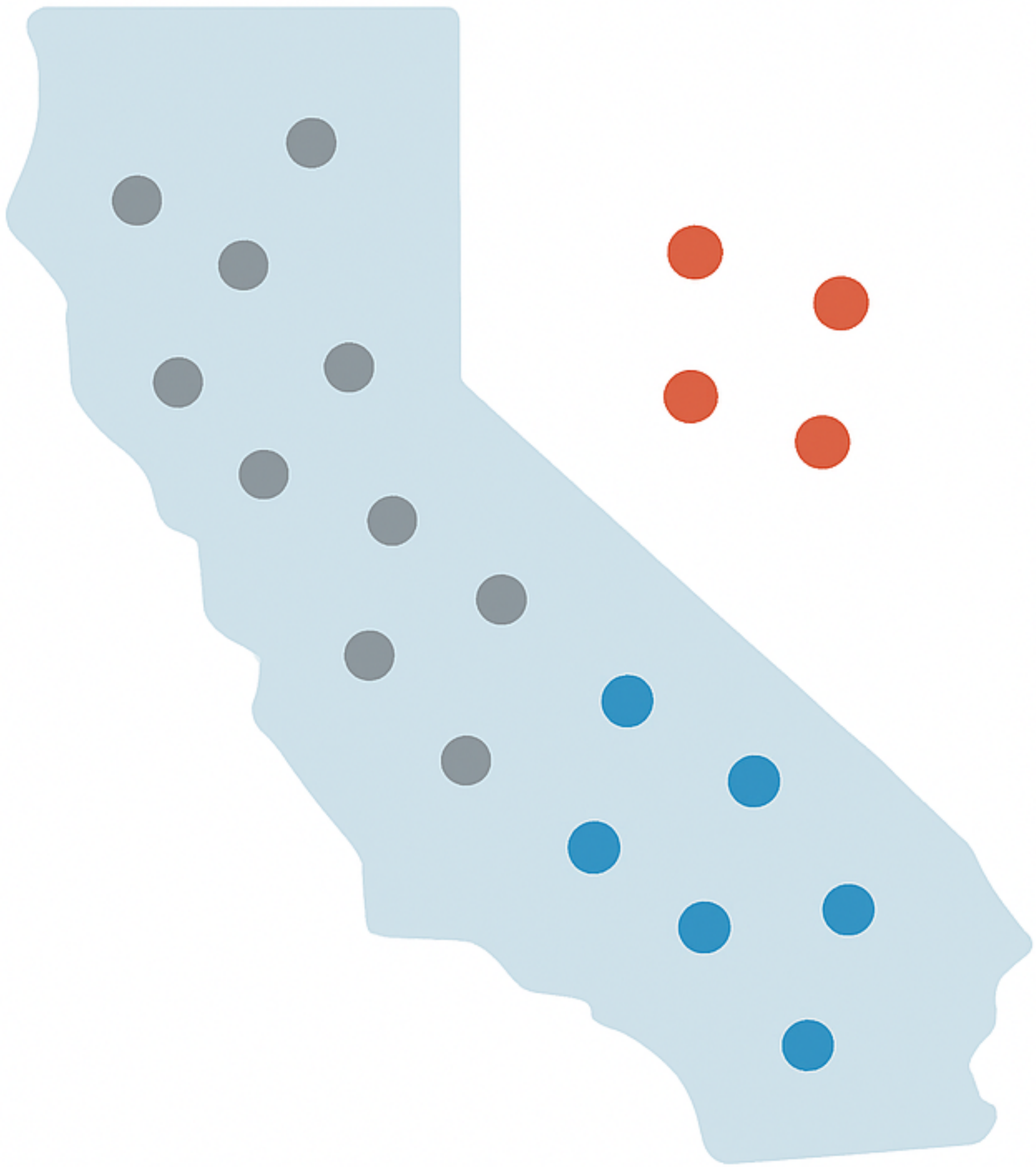
UNIVERSITÀ DEL SALENTO

AI-driven Data Engineering and Reusability for Earth and Space Sciences (DARES 2025)

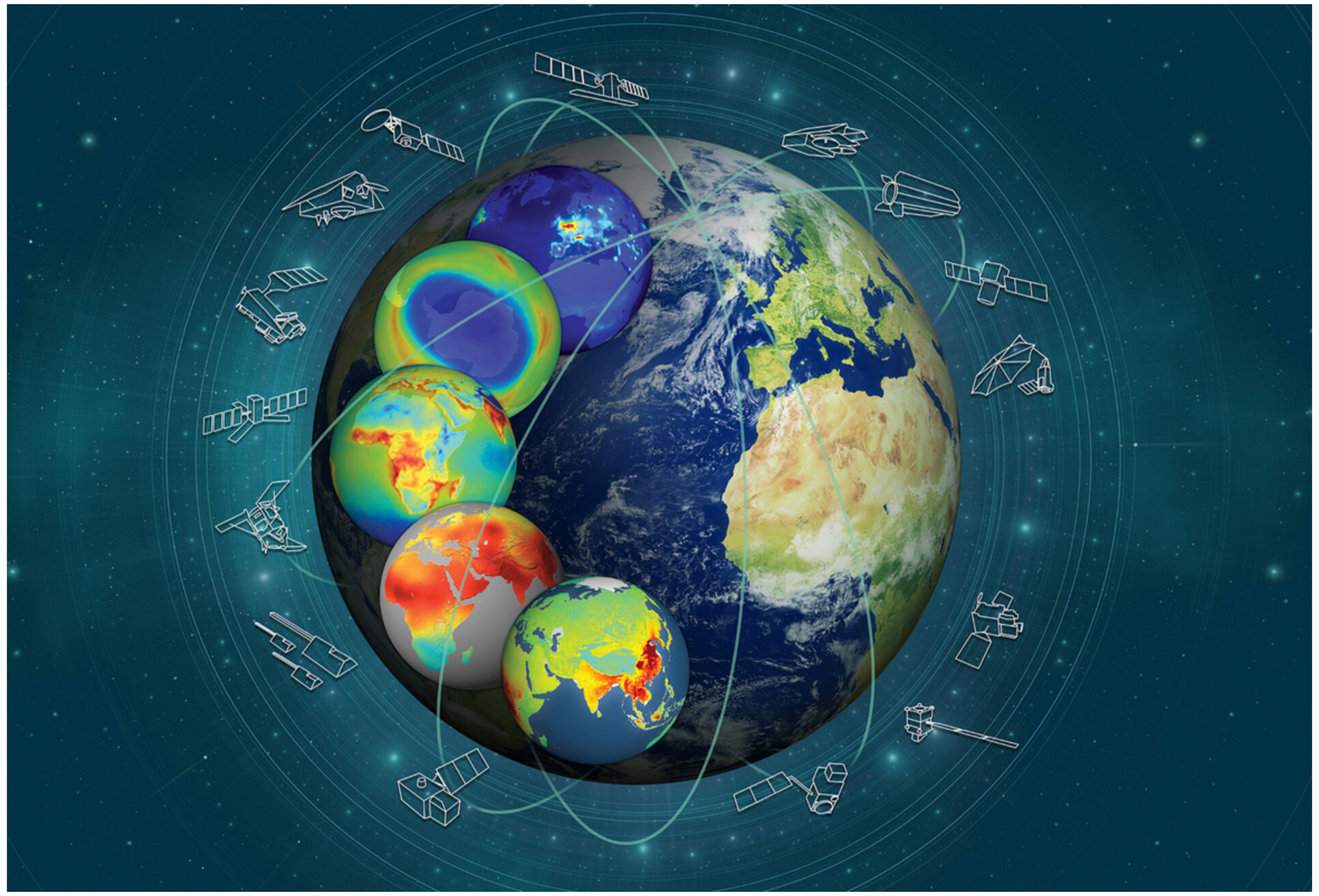
ecai 2025 BOLOGNA

Bologna, 25/10/2025

Introduction to Clustering Methods



Geo-referenced Time-series: why clustering?



Credit: ESA

Vast and complex Earth Observation dataset

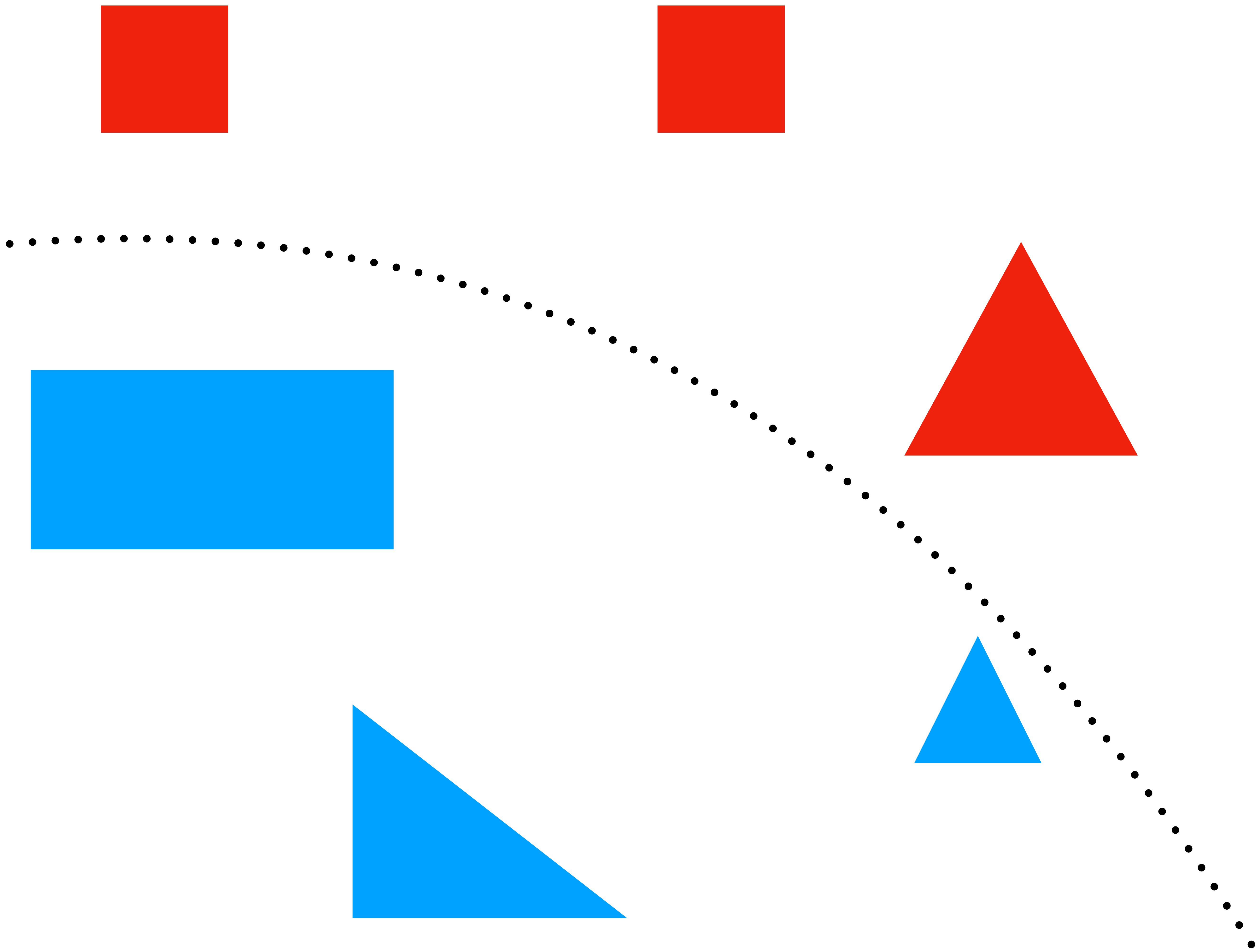
high dimensional

spatially heterogeneous

clustering

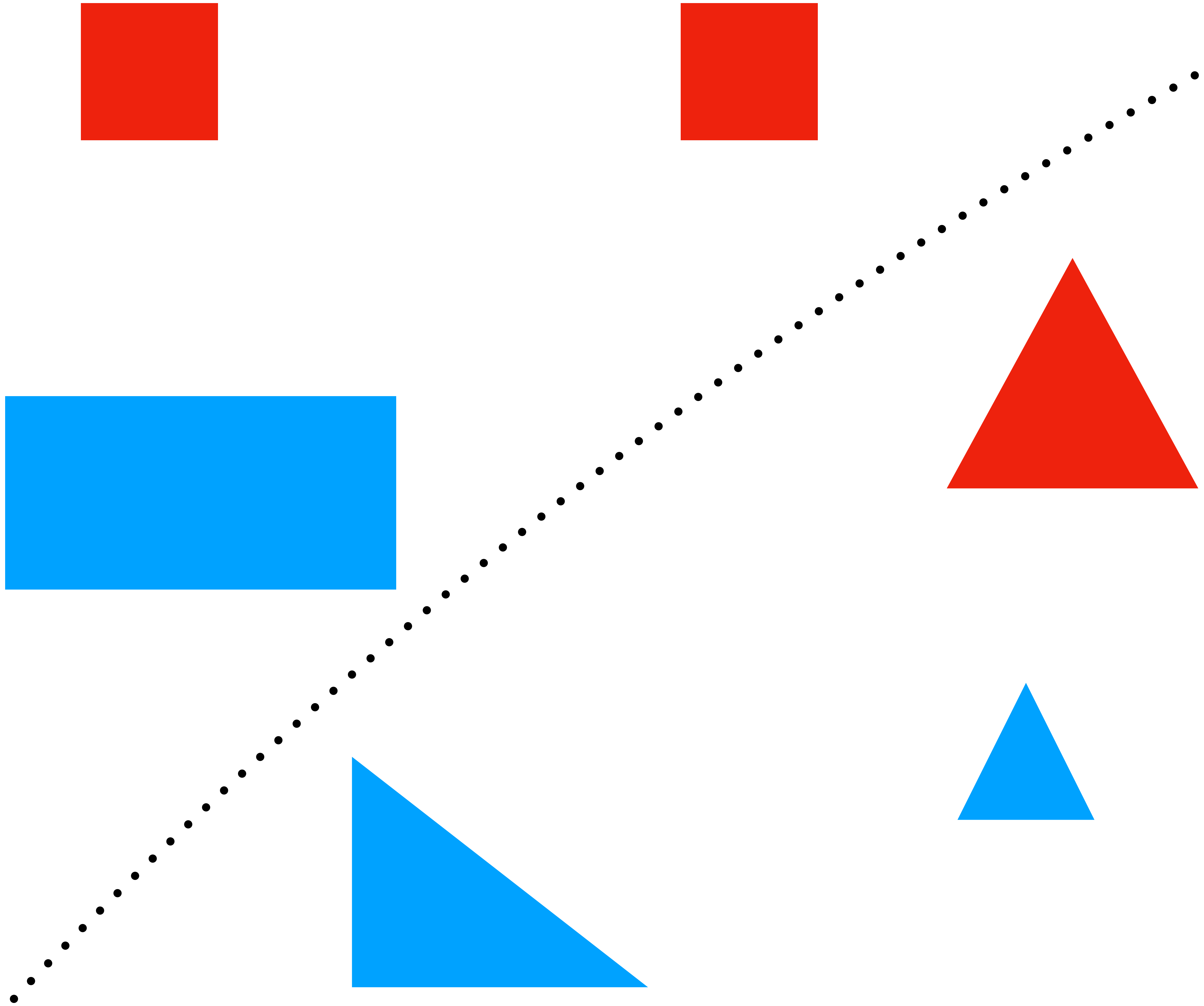
Extract structures

Clustering as unsupervised learning



... the task of partitioning a set of objects in such a way that objects in the same subset (called a **cluster**) are more similar (in some specific sense defined by the analyst) to each other than to those in other clusters.

Clustering as unsupervised learning



... the task of partitioning a set of objects in such a way that objects in the same subset (called a **cluster**) are more similar (in some specific sense defined by the analyst) to each other than to those in other clusters.

Clustering as unsupervised learning

Given

- $\mathcal{X} = \{X_1, \ldots, X_n\}$, a set of $n \geq 2$ objects
- a **dissimilarity function** d that is symmetric in its arguments and assigns a non-negative value to any pair $(\mathcal{X}_1, \mathcal{X}_2)$ of non-empty subsets of $\mathcal{X}$;

the goal is to find a partition $\mathscr{C}$ of $\mathcal{X}$ into K non-overlapping sets $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$, called **clusters**, that solves a minimization problem of type

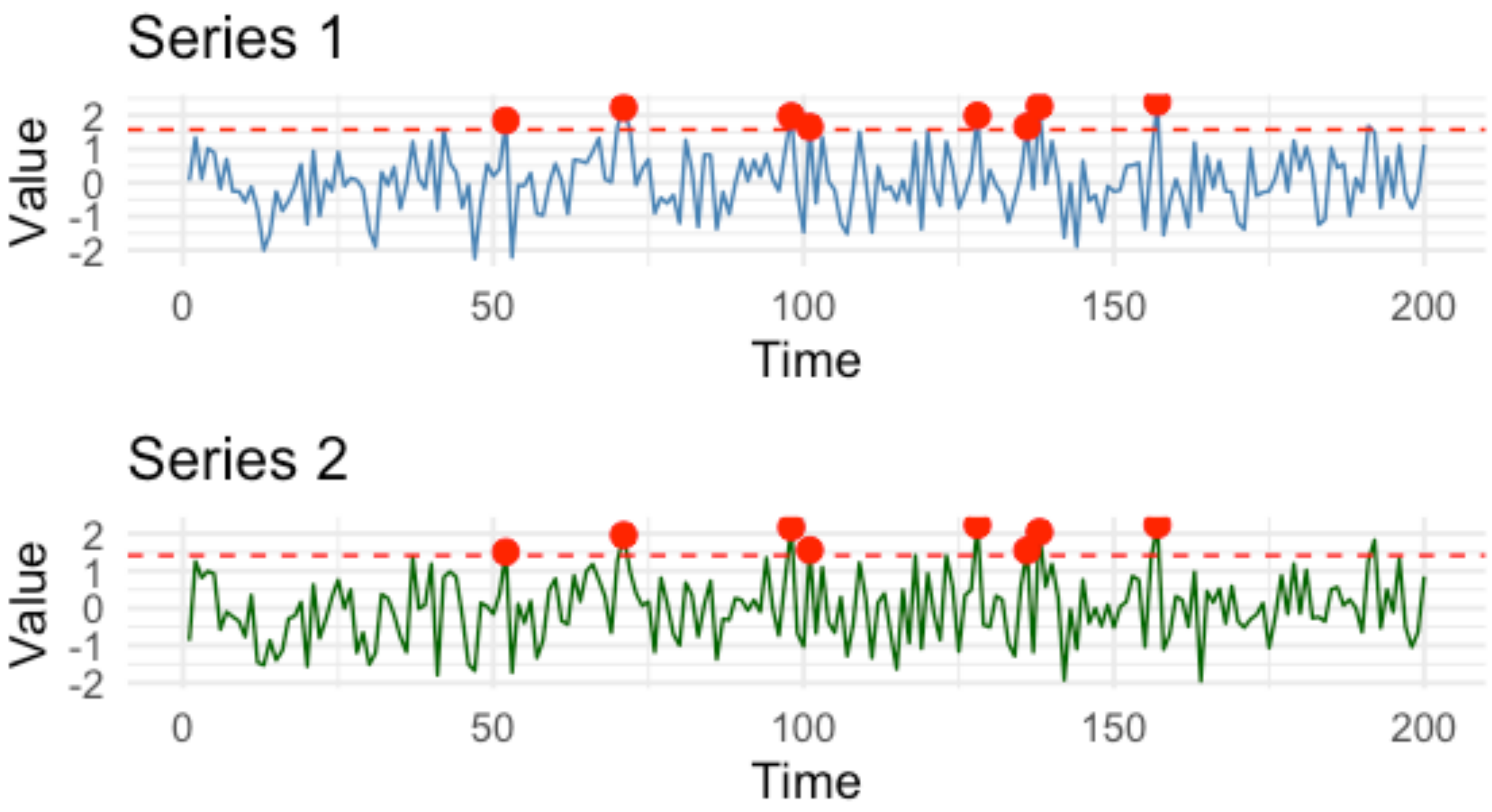
$$\mathscr{C} = \arg\min_{\mathscr{C}'} \varphi(\mathscr{C}'; d)$$

over the class of all possible $K-$partitions $\mathscr{C}'$ of $\mathcal{X}$.

A Copula-based approach

Emerging areas: identifying sets of time series that exhibit **comovement**, or **tail dependence** behavior, regardless of marginal modeling.

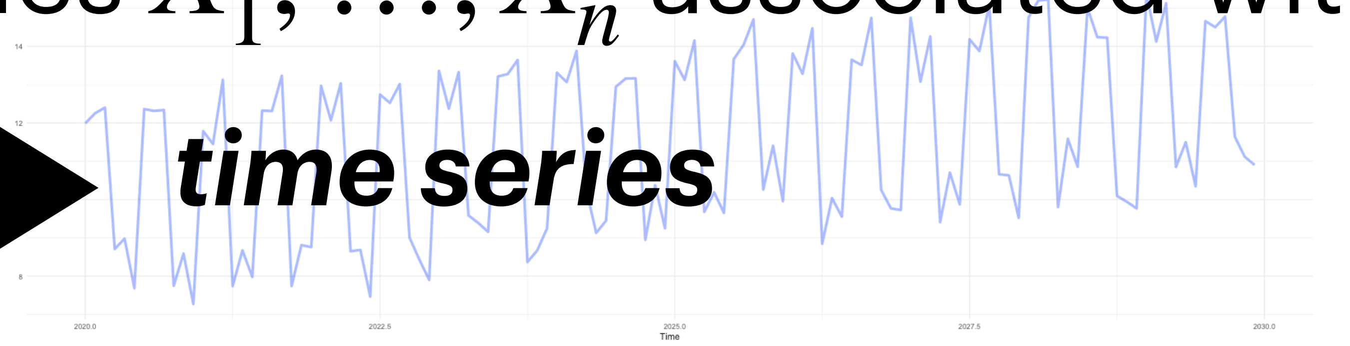
Examples in environmental sciences: analyzing joint extremes such as maxima of precipitations, temperature, or modeling flood risks.



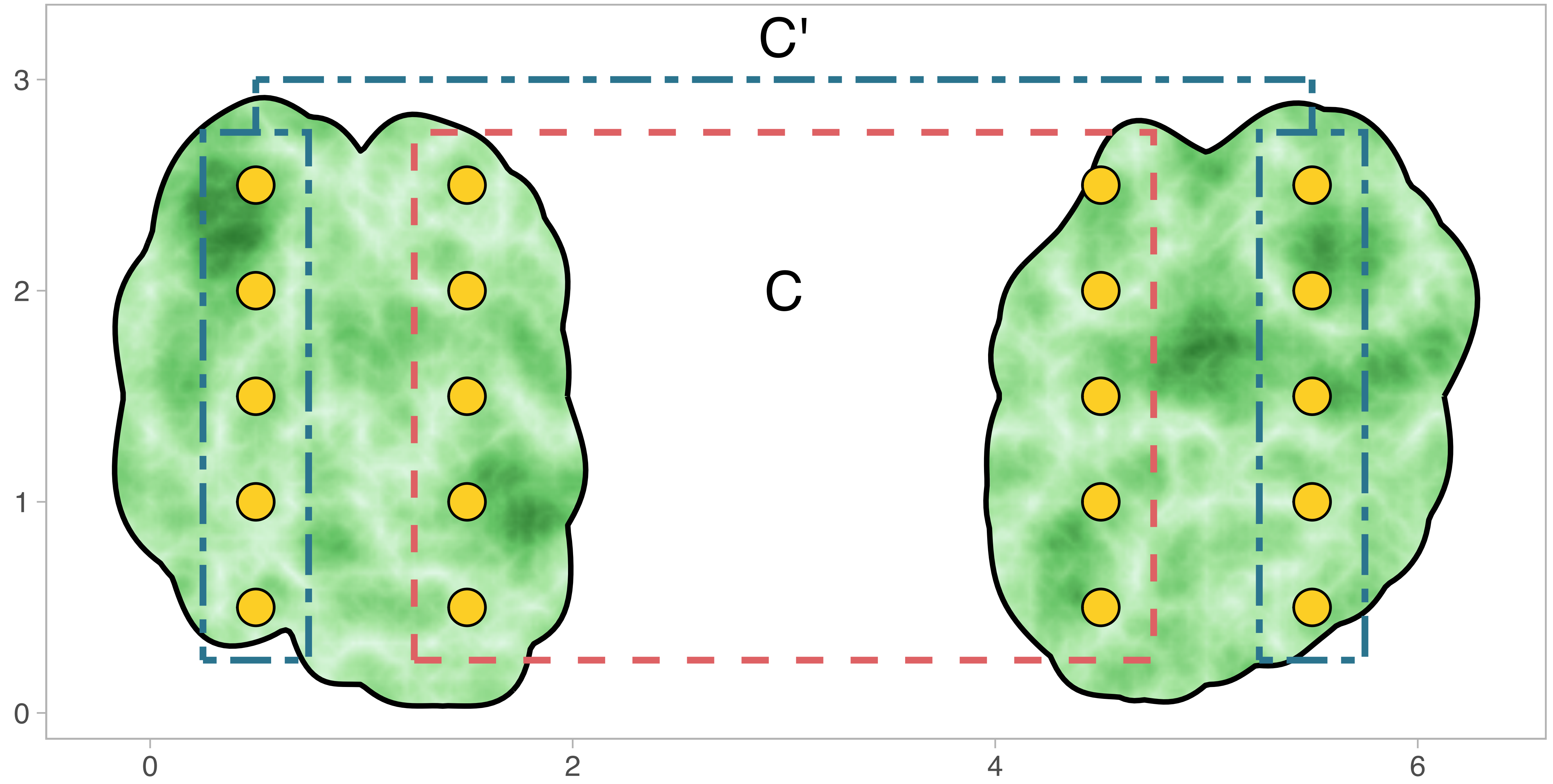
Step C1: Copula-based dependence

A copula-based variable clustering assumes that:

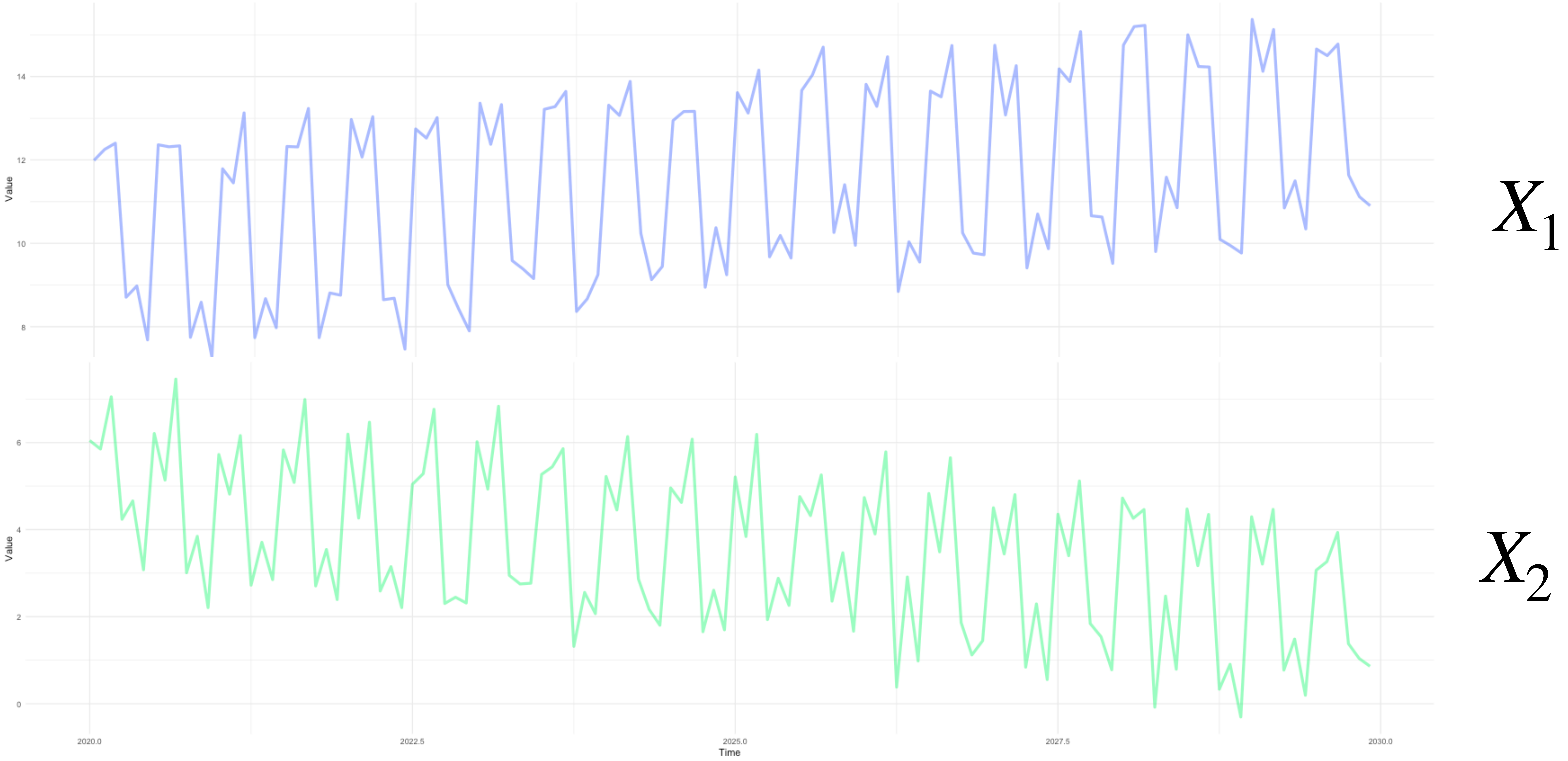
- a set $\mathscr{X}$ of real-valued continuous random variables $X_1, \ldots, X_n$ associated with an iid sample $(x_{ti})_{t=1,\ldots,T} \sim X_i$ for every $i = 1,\ldots,n$; $\longrightarrow$ ***time series***



- the rv's are **continuous**, so that any copula among them is unique;

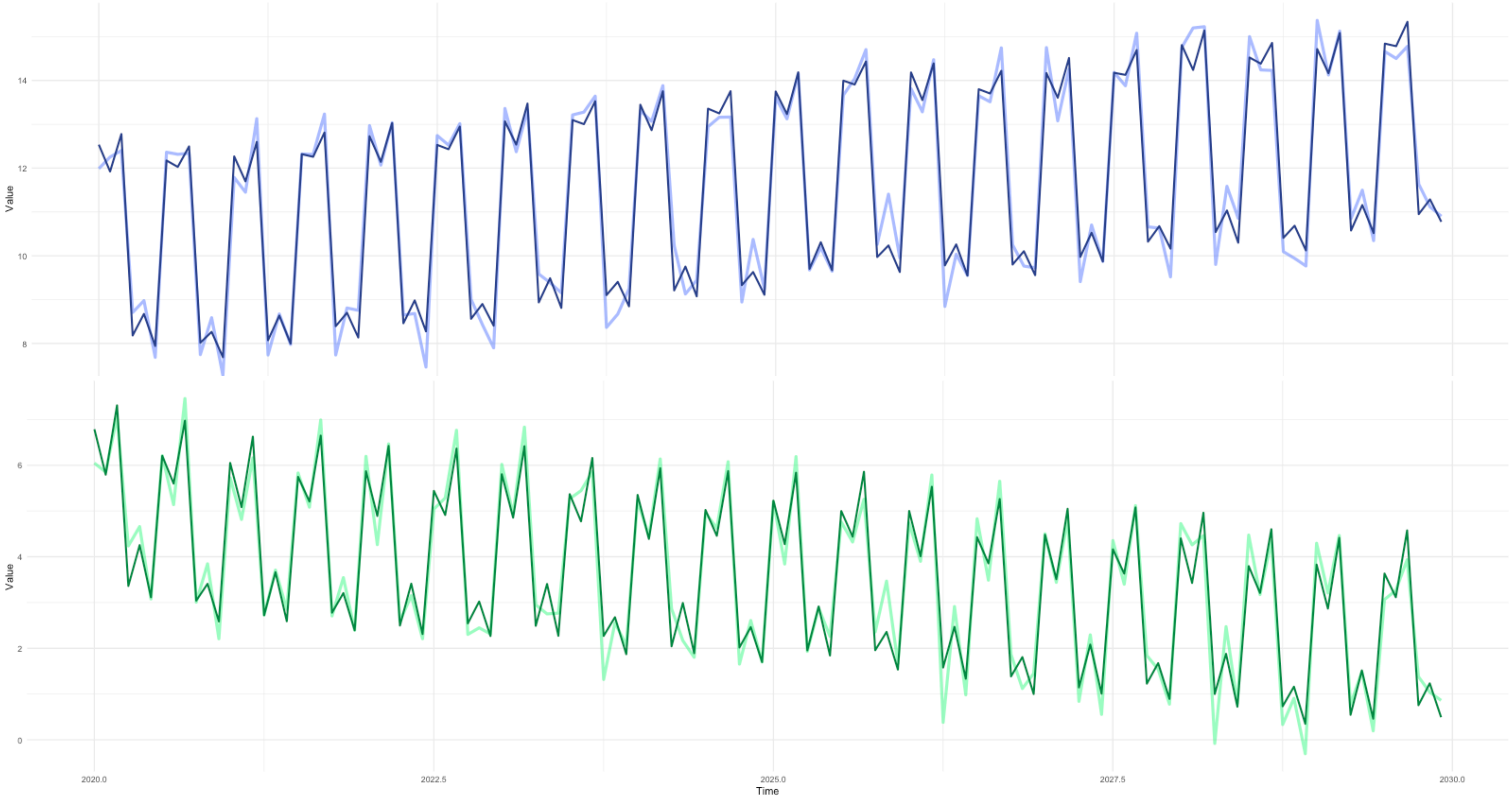


Dependence Retrieval



Patton (2012), Neumeyer et al. (2019)

Dependence Retrieval

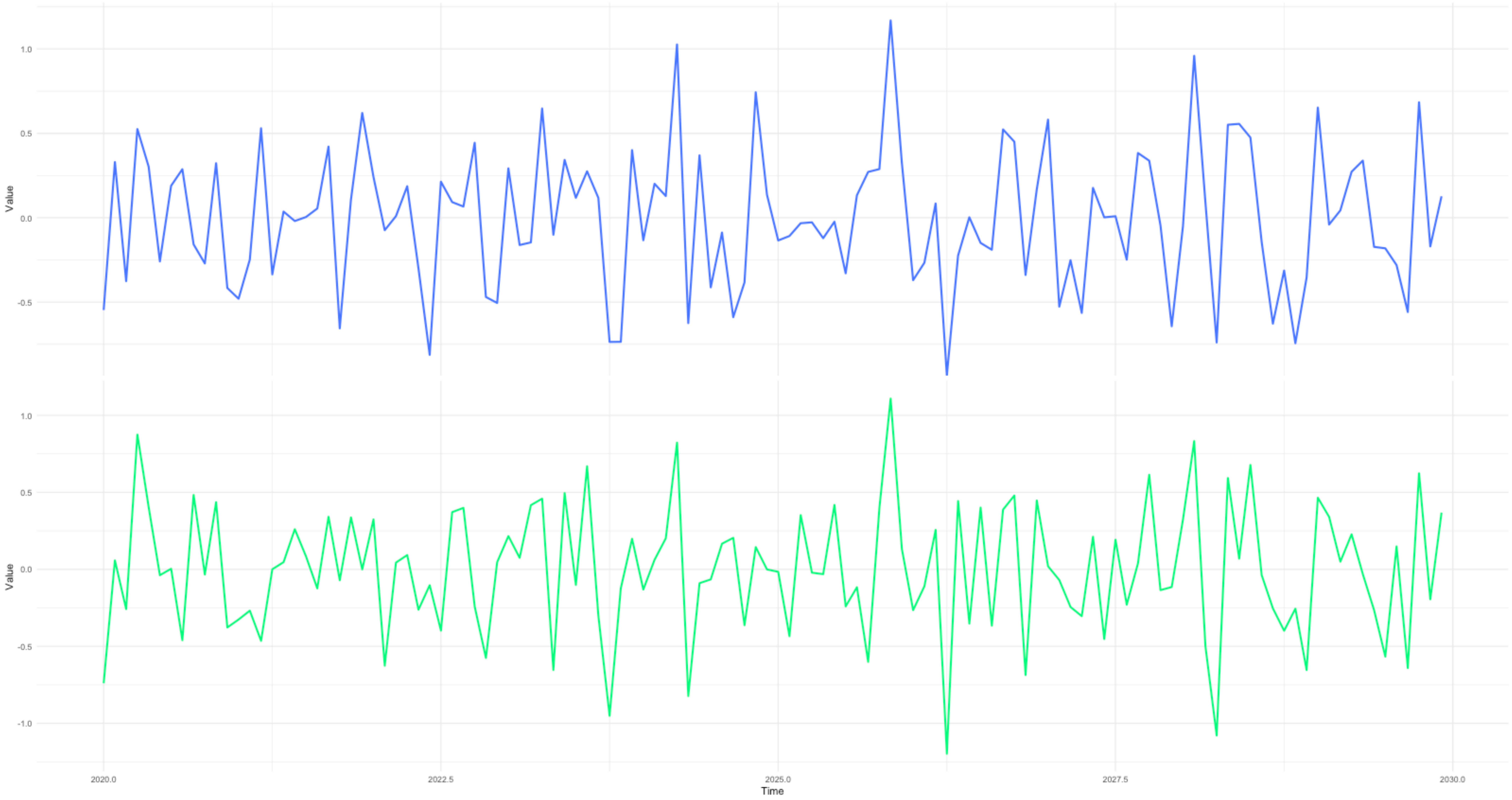


$\hat{X}_1$

$\hat{X}_2$

Patton (2012), Neumeyer et al. (2019)

Dependence Retrieval

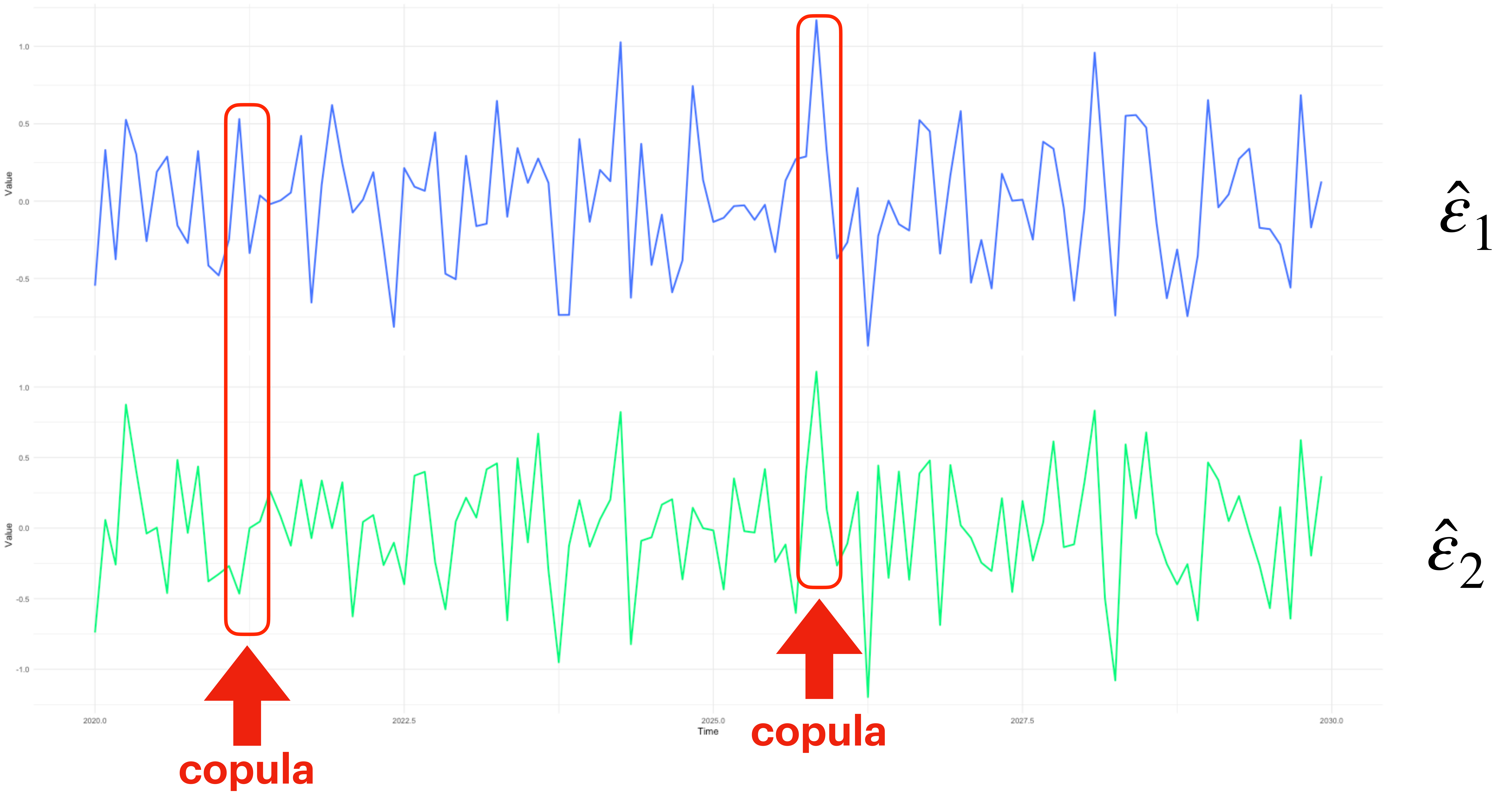


$\hat{\varepsilon}_1$

$\hat{\varepsilon}_2$

Patton (2012), Neumeyer et al. (2019)

Dependence Retrieval



Step C2: Detection of comovements

A copula-based variable clustering assumes that:

- the dissimilarity function $d = d(\mathscr{X}_1, \mathscr{X}_2)$
  - only depends on the copula of $(\mathscr{X}_1, \mathscr{X}_2)$, regardless any possible permutation of the elements in a cluster;
  - is related to the degree of **comonotonicity** among rv's (i.e. closeness to the upper bound of the Fréchet class).

Only dependence
No marginals

Step C2: Detection of comovements

A copula-based variable clustering assumes that:

- the dissimilarity function $d = d(\mathscr{X}_1, \mathscr{X}_2)$
  - only depends on the copula of $(\mathscr{X}_1, \mathscr{X}_2)$, regardless any possible permutation of the elements in a cluster;
  - is related to the degree of **comonotonicity** among rv's (i.e. closeness to the upper bound of the Fréchet class).

Only dependence
No marginals

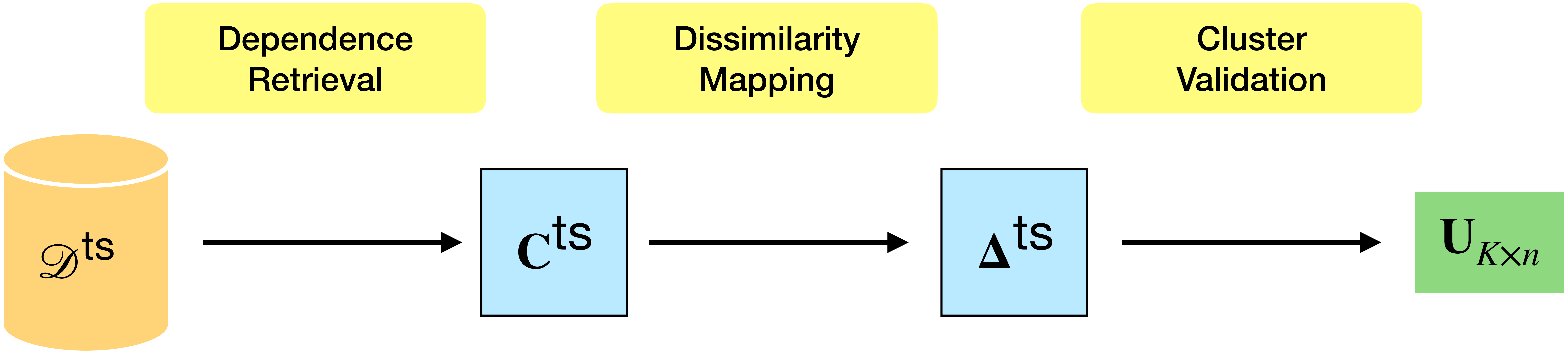
Given a copula matrix $\mathbf{C}$ each pairwise copula C_{ij} is transformed into a numerical dissimilarity Δ_{ij}

Comonotone-based clustering requires $d^{1,1}$ to satisfy:

$$d^{1,1}(C) = 0 \text{ if } C = M, \qquad \text{where } M(u, v) = \min\{u, v\}.$$

$d^{1,1}$ evaluates how far C_{ij} is from the (perfect) comonotonicity copula M (Fuchs et al. 2021)

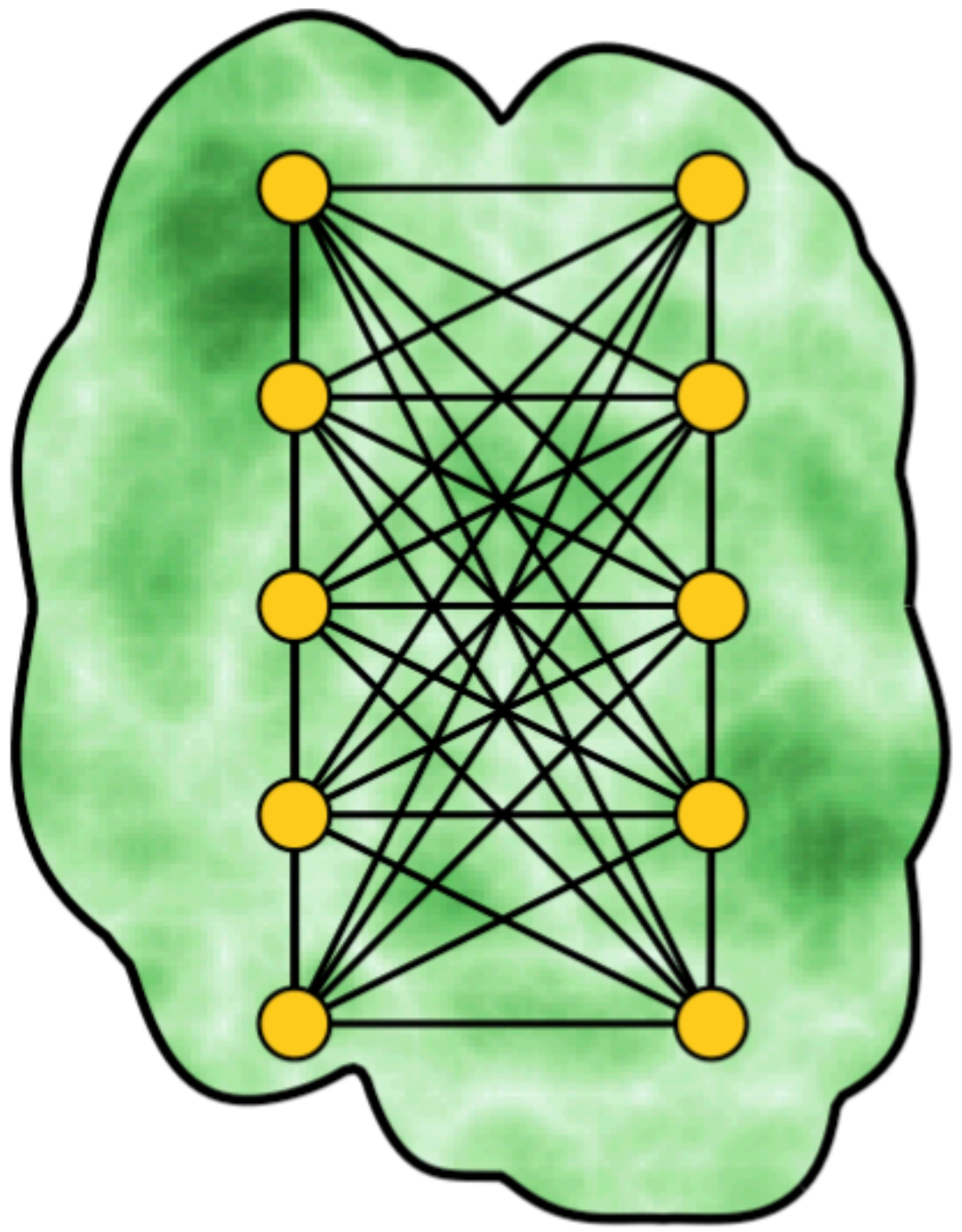
Copula-based time series clustering – model architecture

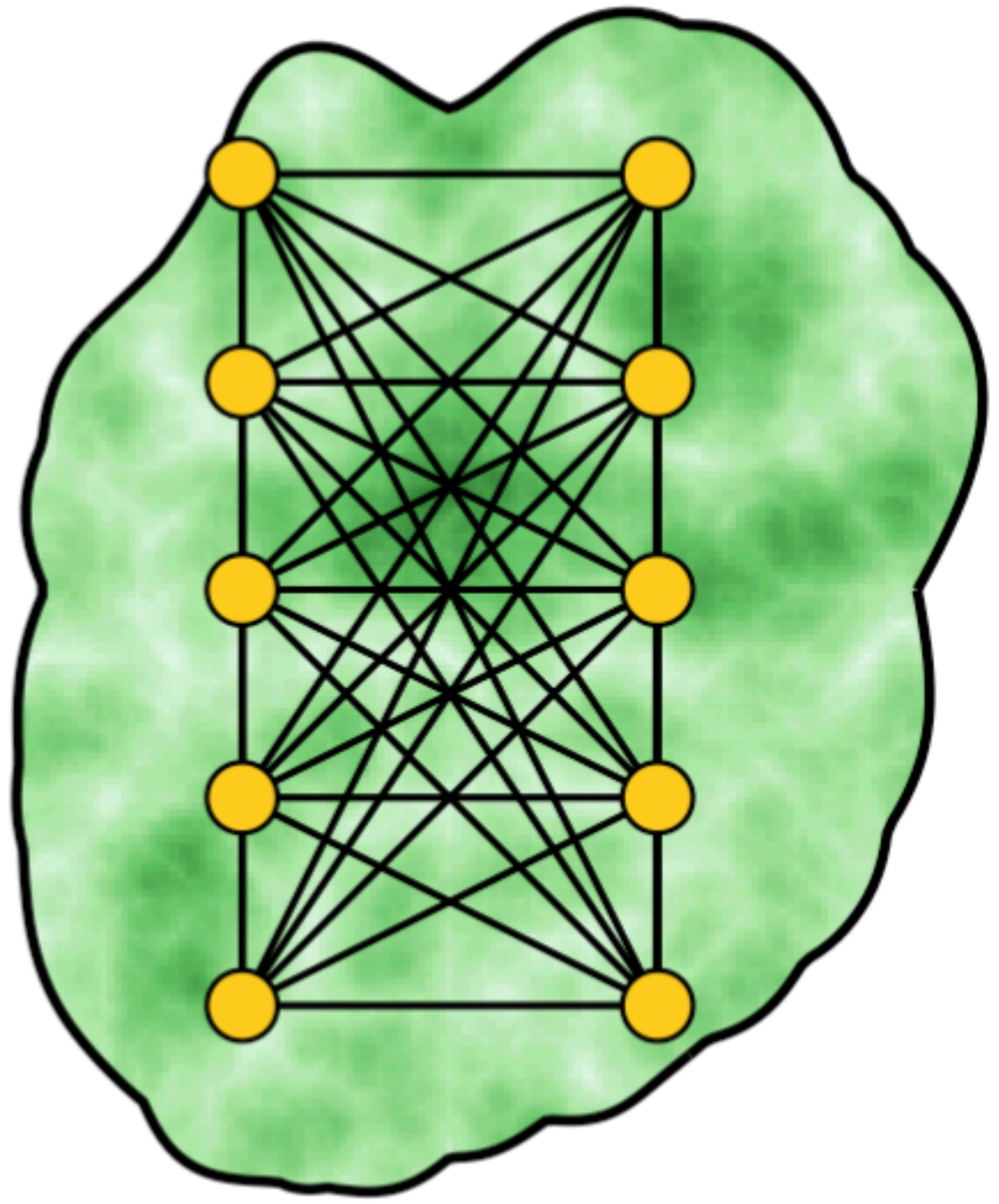


The procedure will be of **algorithmic nature and data-driven** (eventually with some working model assumptions).

Each partition in K clusters is represented by a membership matrix U of order $(K \times n)$ so that each entry U_{ki} belongs to {0, 1} (or $[0,1]$ in a Fuzzy/soft context) and the sum of the entries in each column is 1

Clustering with spatial constraints



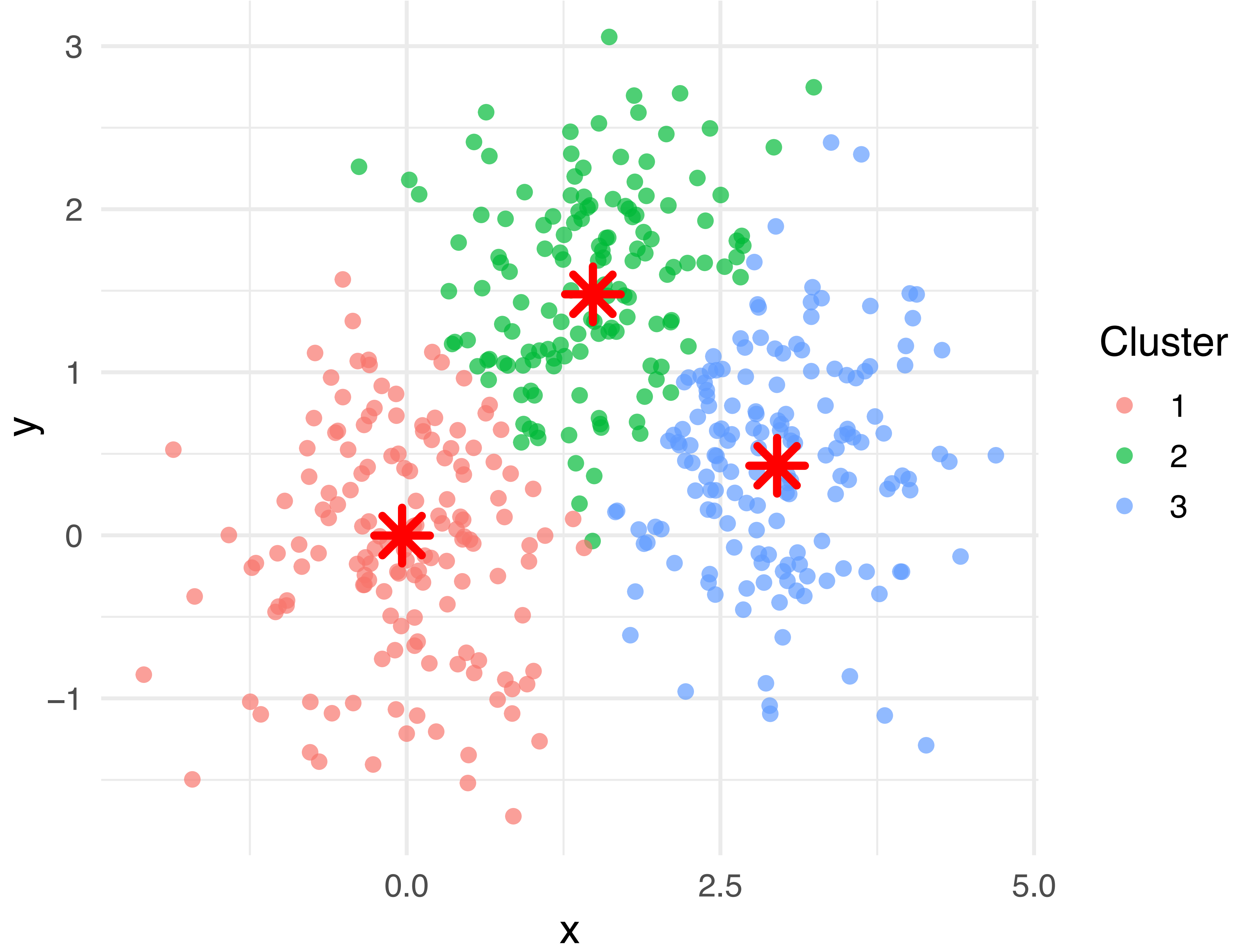


First application: maximum temperatures

- **Time series:** monthly maximum temperatures of the summer months (JJA), derived from ERA5 reanalysis data, spanning the period from 1960 to 2024.
- **Locations:** multiple spatial points across the Italian territory. We restrict our attention to land areas by excluding sea points, selecting a subset of n=105 grid points.
- **Goal:** 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.

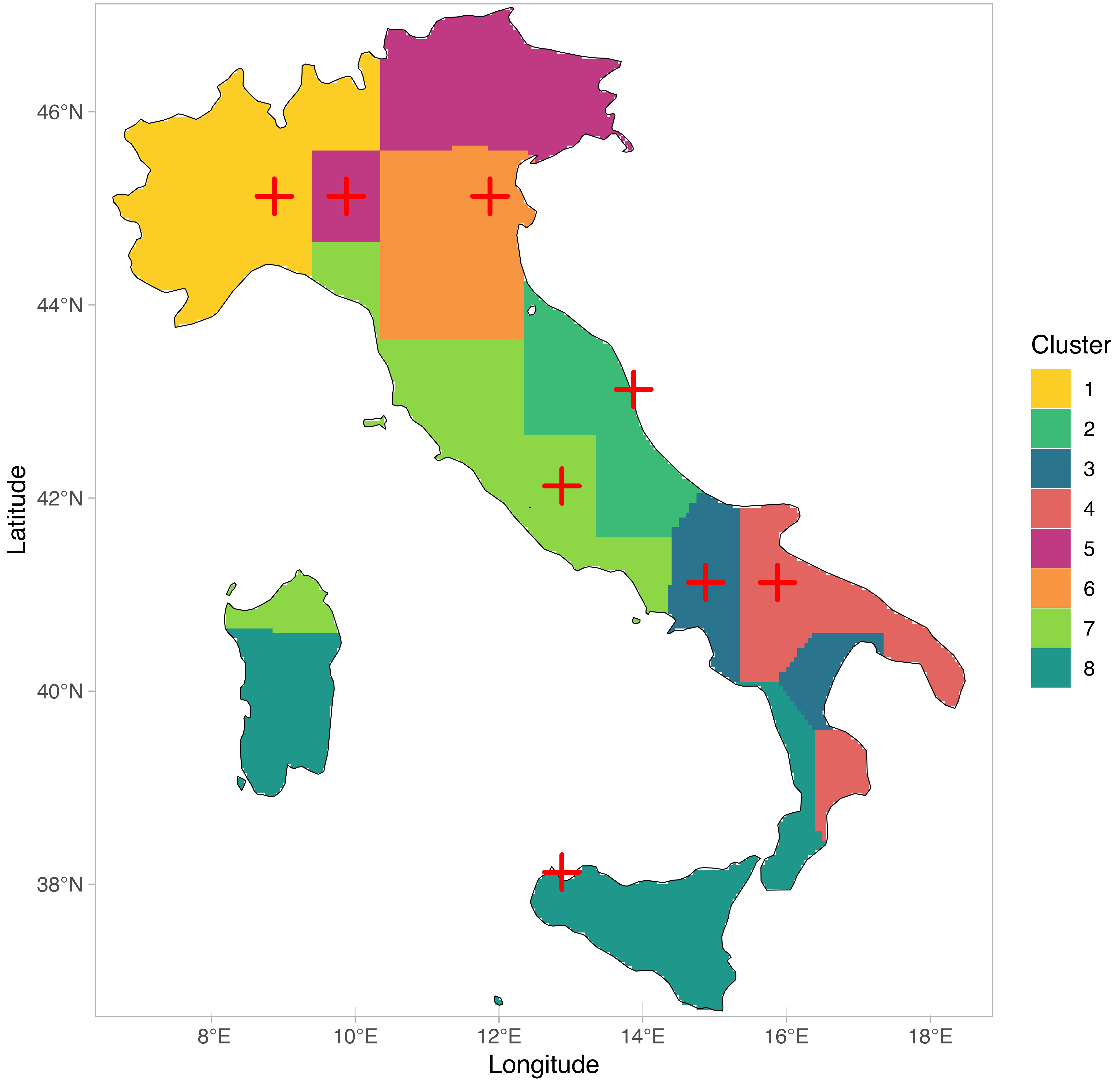
Partitioning Around Medoids (PAM)

- PAM is a clustering algorithm that groups data points into *k* clusters.
- Each cluster is represented by a *medoid*, which is a real data point acting as the most central or representative object of the cluster.
- The goal is to make each cluster as compact and consistent as possible, minimizing the total dissimilarity between points and their medoid.



PAM finds representative data points (medoids) that best describe *k* compact and well-separated clusters

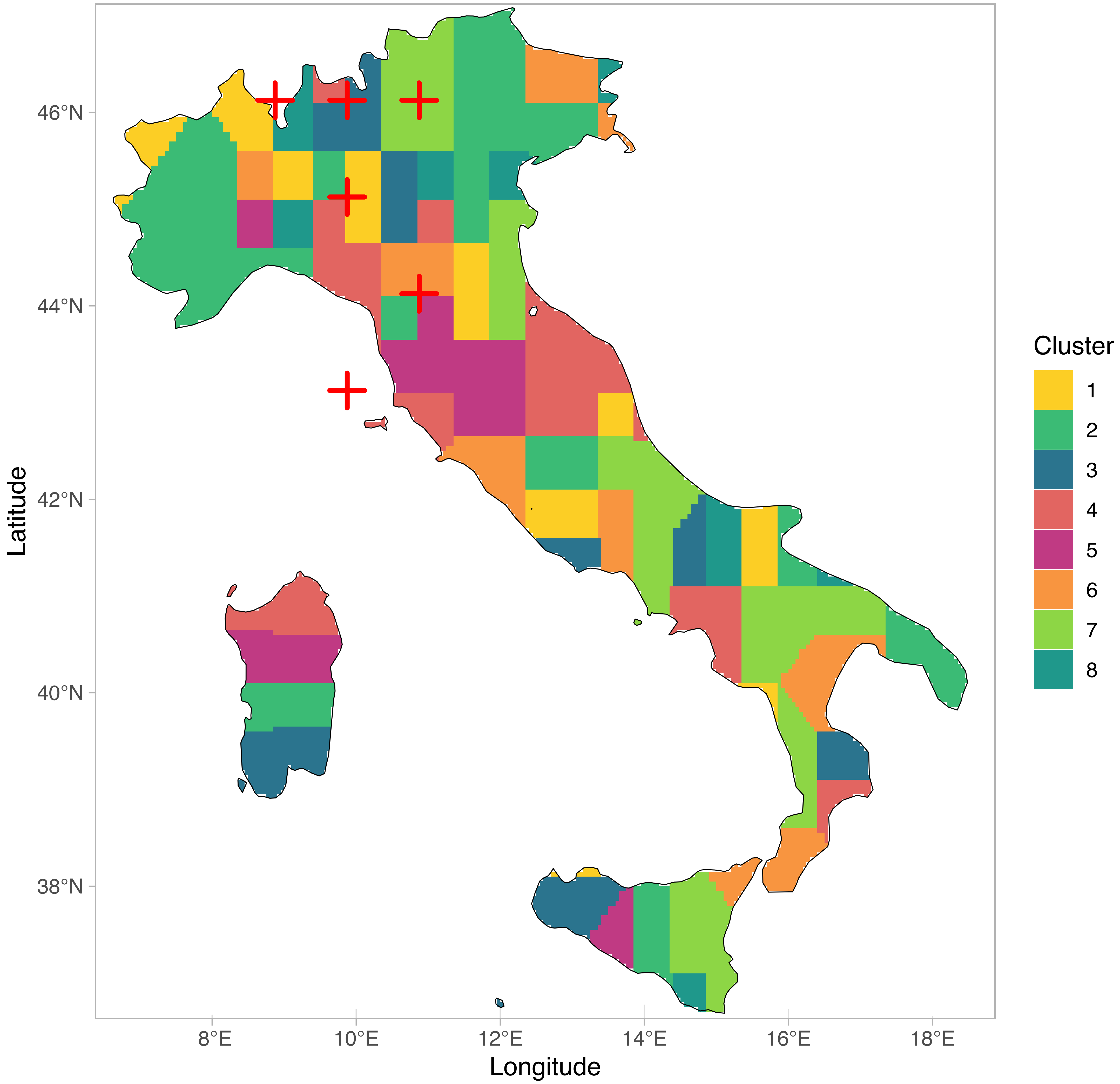
First application: maximum temperatures



Temperature maxima

- Partitioning Around Medoids (PAM)
- K = 8 clusters: reasonable compromise between the optimal Average Silhouette Index and the need for a clear and interpretable spatial visualization
- Groups of locations whose time series exhibit similar comovement patterns

Second application: maximum precipitations

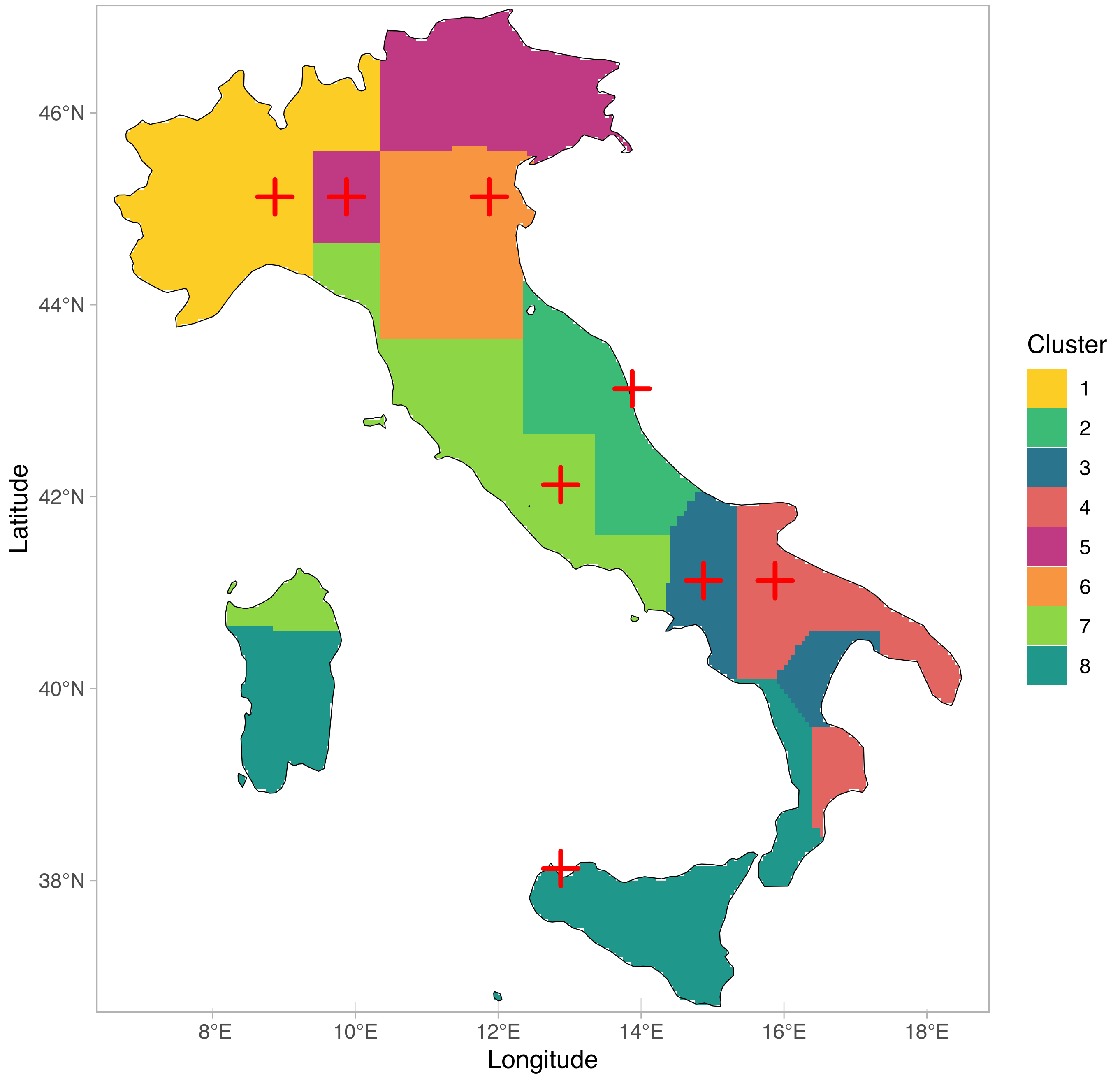


Precipitation maxima

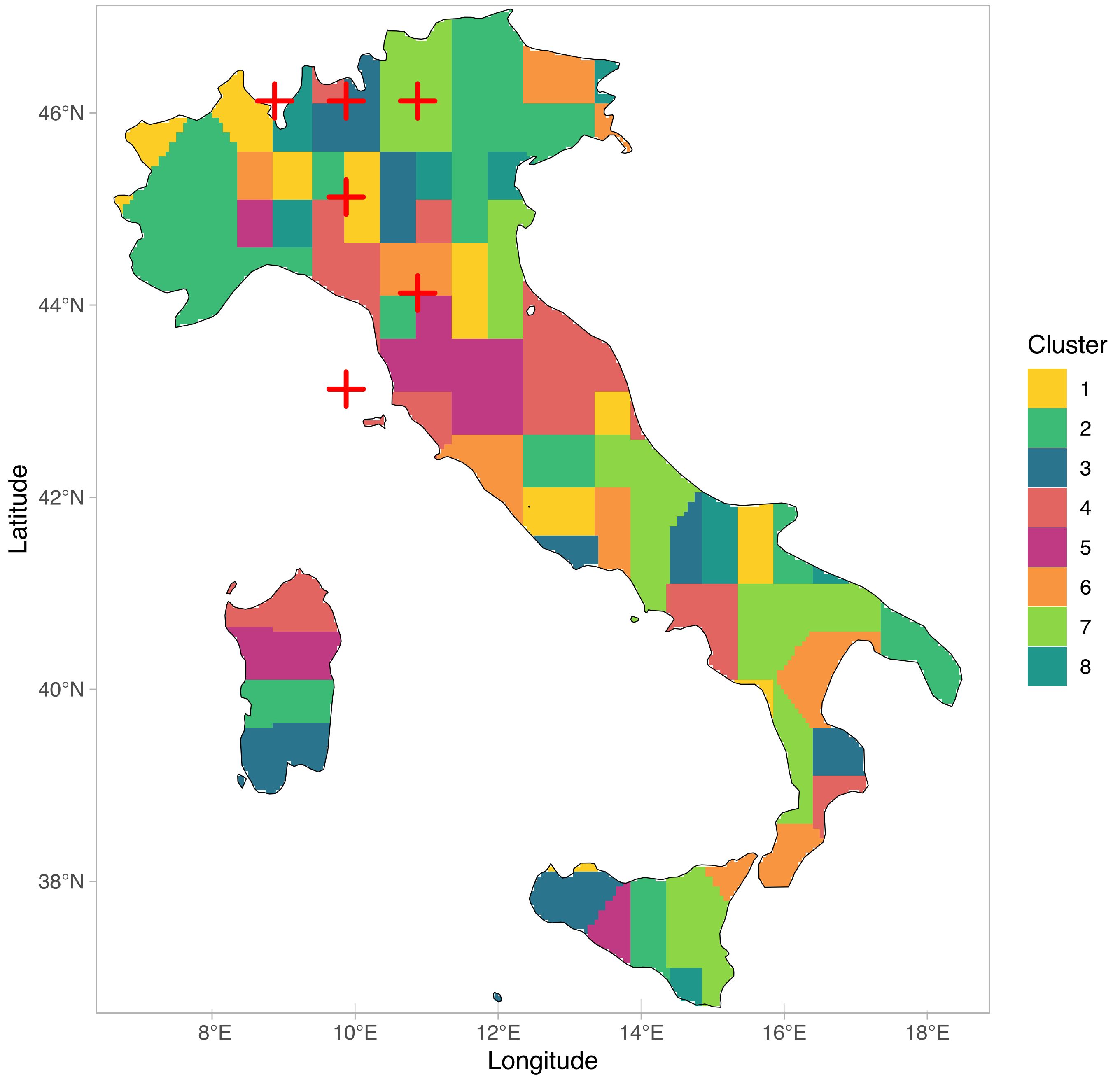
- Partitioning Around Medoids (PAM)
- K = 8 clusters: reasonable compromise between the optimal Average Silhouette Index and the need for a clear and interpretable spatial visualization
- Groups of locations whose time series exhibit similar comovement patterns

Spatial heterogeneity!

Using only temporal information can lead to heterogeneous clusters that are hard to interpret:



Temperature maxima



Precipitation maxima

Solution: introduce spatial constraints

Semi-supervised learning algorithms

Feature Information

Given

- a set of real-valued continuous random variables $X_1, \ldots, X_n$ associated with an iid sample $(x_{ti})_{t=1,\ldots,T} \sim X_i$ for every $i = 1,\ldots,n$;
- a p-dimensional vector $\mathbf{s}_i^\top$ associated with each X_i for every $i = 1,\ldots,n$, that represent the spatial (e.g. geographic) location where X_i is observed;
- a dissimilarity function d for rv's;

Spatial Information

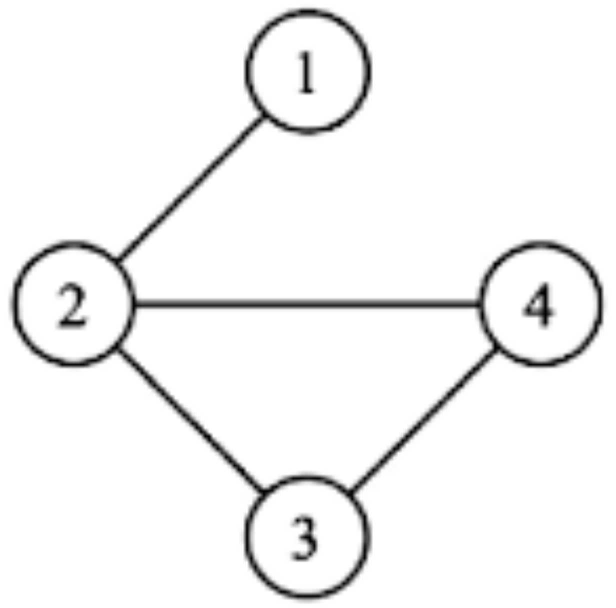
the goal is to find an algorithm to group variables which are similar in their statistical attributes as well as in their spatial location.

Spatial proximity retrieval

Additional information is associated with the multivariate time series in the form of a set $D^{sp} = \{s_1, \ldots, s_n\}$
D^{sp} is converted to an element in $Diss(n)$.

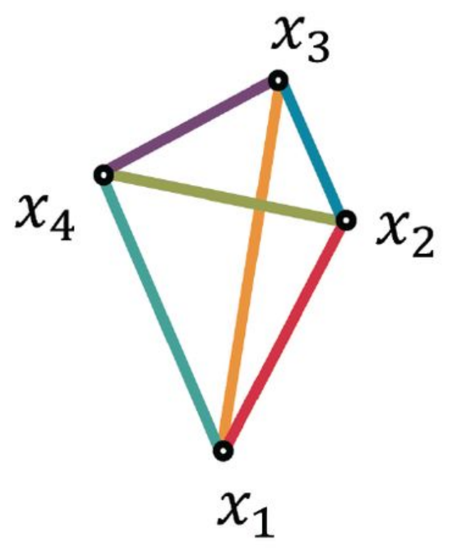
Two main types of matrices can be obtained:

- an incidence matrix $\Delta^{sp} = (\Delta^{sp}_{ij}) \in \{0,1\}^n$, where each entry Δ^{sp}_{ij} indicates whether the i–th and j–th time series components are related (value equals to 1) or not (value equal to 0).



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

- a distance matrix $\Delta^{sp} = (\Delta^{sp}_{ij}) \in [0, +\infty]^n$, where each entry Δ^{sp}_{ij} only depends on the (Euclidean) distance between s_i and s_j



D

	x_1	x_2	x_3	x_4
x_1	0	9	9	9
x_2	9	0	1	1
x_3	9	1	0	1
x_4	9	1	1	0

Semi-supervised learning algorithms with soft constraints

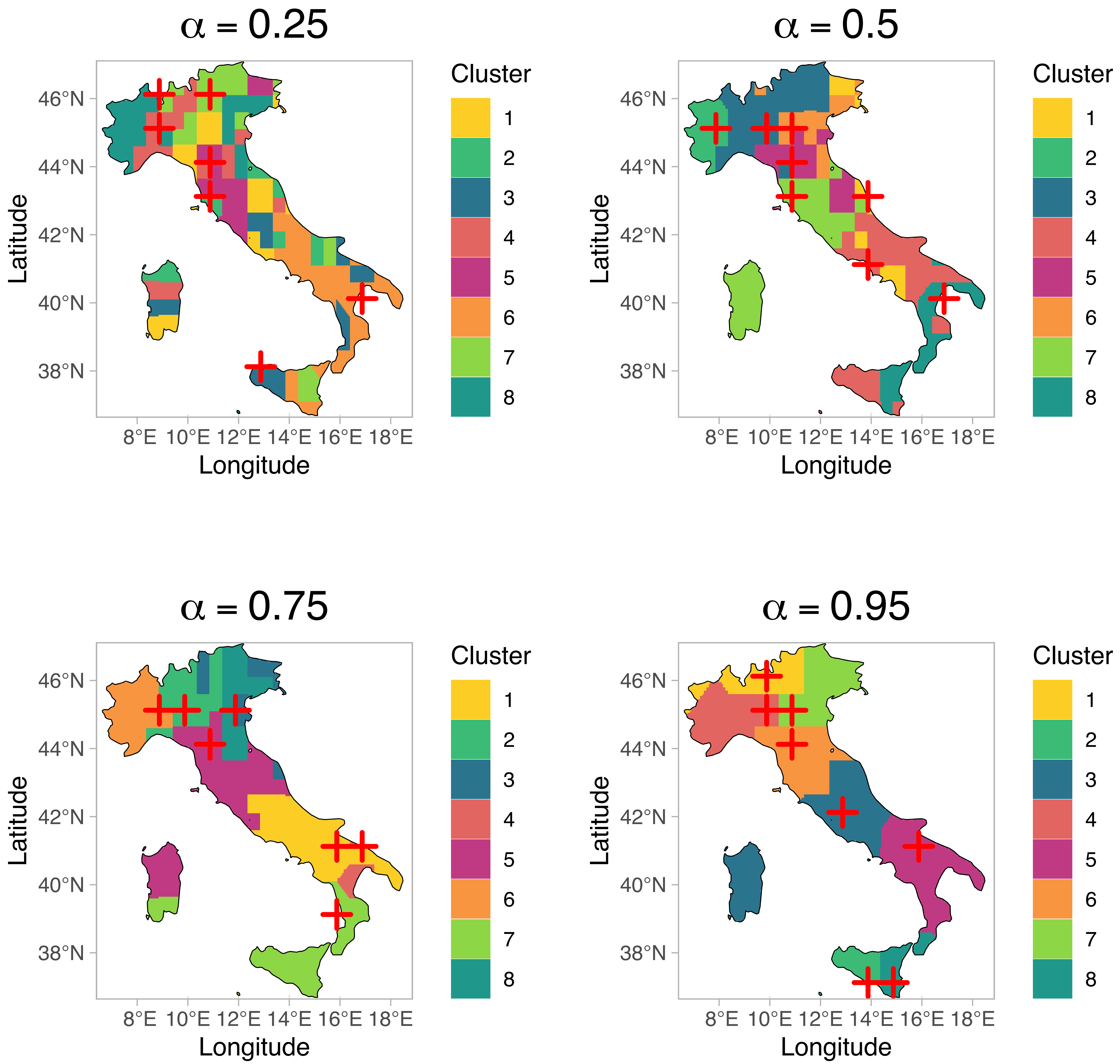
$$\Delta^{\alpha} = \underbrace{(1-\alpha)}_{\text{Hyper-parameter}} \underbrace{\begin{pmatrix} \Delta^{ts}_{11} & \cdots & \Delta^{ts}_{1n} \\ \cdots & \cdots & \cdots \\ \Delta^{ts}_{n1} & \cdots & \Delta^{ts}_{nn} \end{pmatrix}}_{\text{Temporal dissimilarity}} \underbrace{\oplus}_{\text{Aggregation function}} \underbrace{\alpha}_{\text{Hyper-parameter}} \underbrace{\begin{pmatrix} \Delta^{sp}_{11} & \cdots & \Delta^{sp}_{1n} \\ \cdots & \cdots & \cdots \\ \Delta^{sp}_{n1} & \cdots & \Delta^{sp}_{nn} \end{pmatrix}}_{\text{Spatial dissimilarity}}$$

Aggregation function

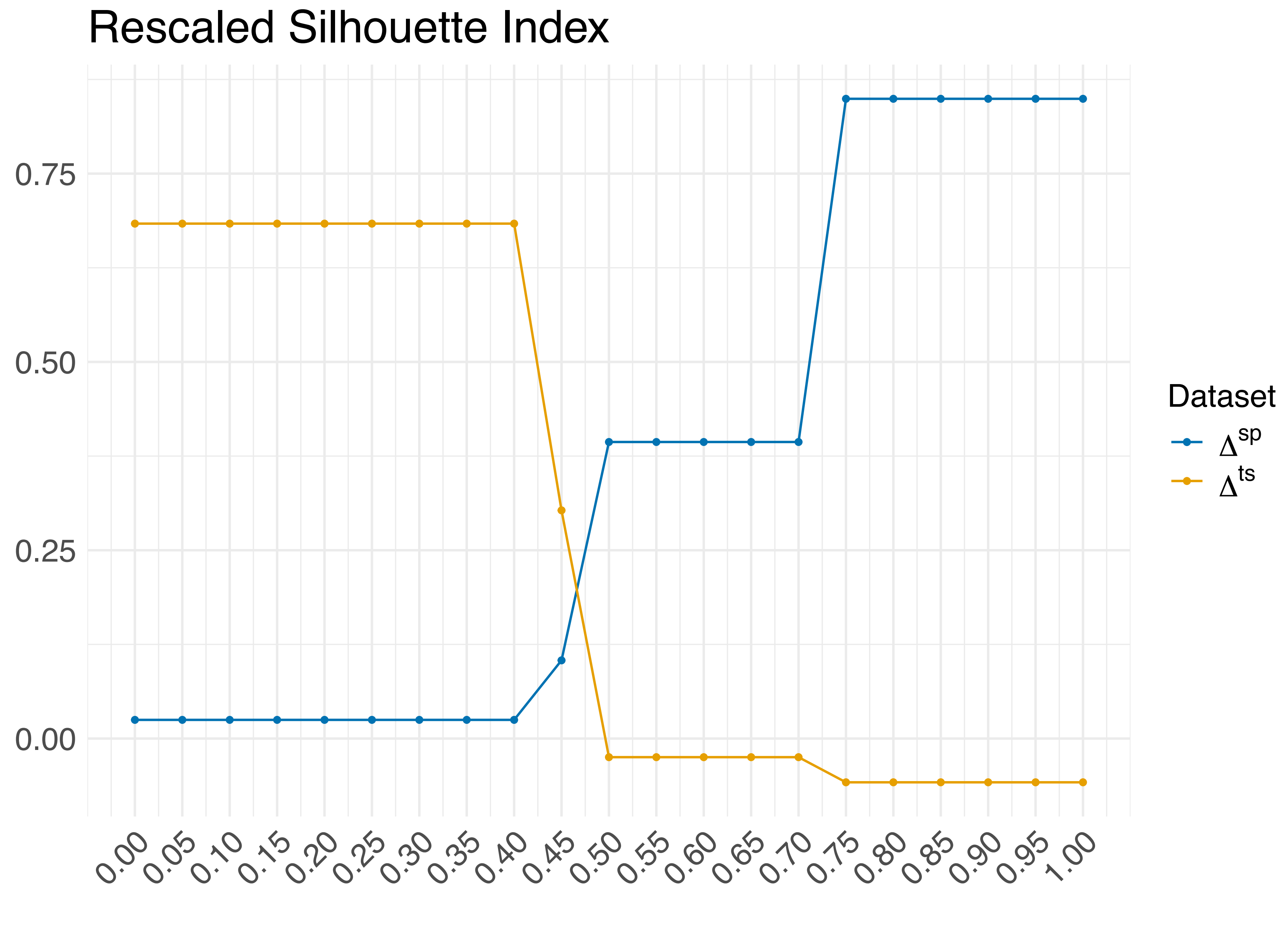
- Convex combination
- Hadamard product
- Geodesic in the Riemannian manifold Corr(n)

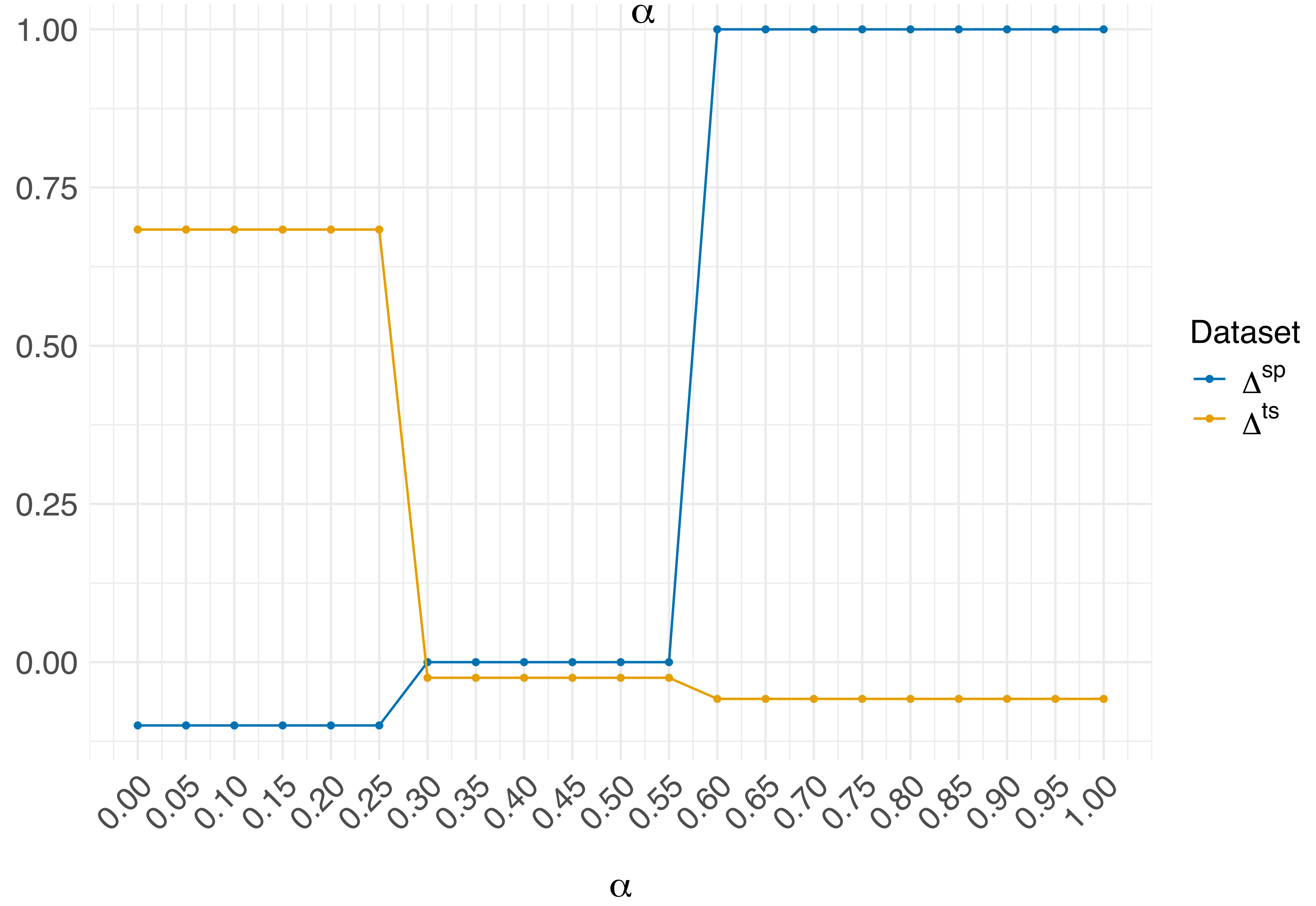
(De Carvalho et al, 2023; Legendre and Gauthier, 2014; B. and Durante, 2024)

Clustering with spatial constraints



How to choose α ?





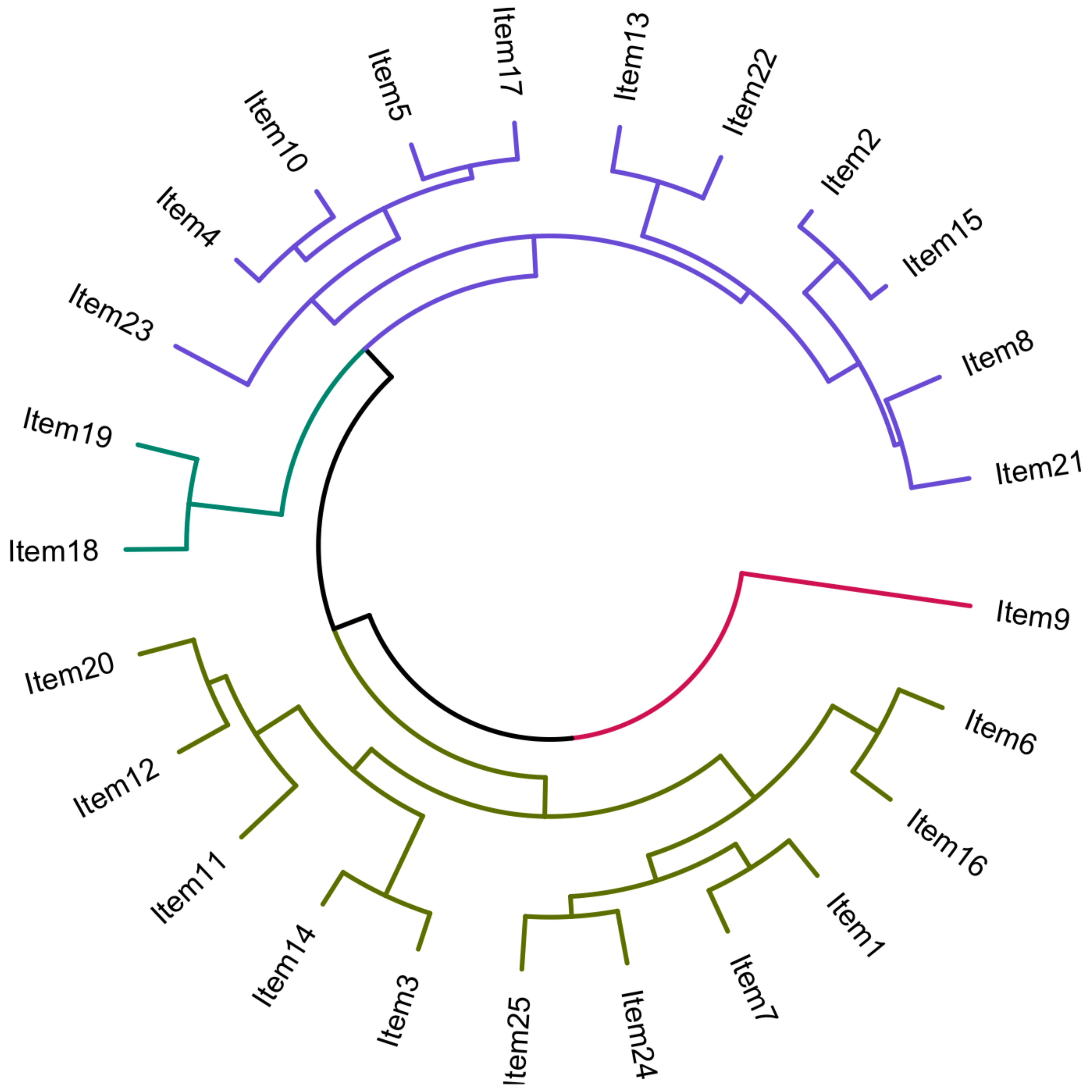
Evolution of the Silhouette Index for $\alpha \in [0,1]$.

Orange points: Silhouette Index computed with respect to the temporal matrix

Blue points: Silhouette Index computed with respect to the spatial matrix.

Silhouette index = ratio of the difference between minimal inter-cluster dissimilarity and intra-cluster dissimilarity to their maximum.

Conclusions



Conclusions

- Copula functions offer a natural way to describe joint comovements among time series, that are particularly useful to analyze joint extremes such as maxima of precipitations, temperature, or modeling flood risks.
- Using only temporal information can lead to heterogeneous clusters that are hard to interpret.
- Copula-based algorithms can be based on a “regularized” dissimilarity matrix taking into account the spatial information.

Open problems and future directions

- Choice of the hyper-parameter α.
- Introduce the ocean in the analysis
- Learning Joint Spatio-Temporal Patterns for *Multivariate Anomaly Detection*

References

1. A. Benevento and F. Durante. *Wasserstein dissimilarity for copula-based clustering of time series with spatial information*. Mathematics, 12:67, 2024. doi:10.3390/math12010067.
2. A. Benevento and F. Durante. *Correlation-based hierarchical clustering of time series with spatial constraints*. Spatial Statistics, 59:100797, 2024. doi:10.1016/j.spasta.2023.100797.
3. A. Benevento, F. Durante, and R. Pappadà. *Tail-dependence clustering of time series with spatial constraints*. Environ. Ecol. Stat., pages 1–17, 2024. doi:10.1007/s10651-024-00626-6.
4. A. Benevento, F. Durante, D. Gallo, and A. Gatto. *Hierarchical clustering of time series with Wasserstein distance*. In: M. Corazza, C. Perna, C. Pizzi, and M. Sibillo, editors, Mathematical and Statistical Methods for Actuarial Sciences and Finance, in press. Springer, Cham, 2024.
5. A. Benevento, F. Durante, and R. Pappadà. *Comonotonic-based Time Series Clustering with constraints: a review and a conceptual framework* Environmetrics, in press, 2025. doi:10.1002/env.70047

References

6. de Carvalho, M., Huser, R., & Rubio, R. (2023). Similarity-based clustering for patterns of extreme values. Stat, 12(1), e560.
7. M. Disegna, P. D'Urso and F. Durante. *Copula–based fuzzy clustering of spatial time series*. Spatial Statistics, 21:209–225, 2017. doi:10.1016/j.spasta.2017.07.002.
8. S. Fuchs, F. M. L. Di Lascio and F. Durante. *Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables*. Comput. Statist. Data Anal., 159:107201, 2021. doi:10.1016/j.csda.2021.107201.
9. Legendre, P., & Gauthier, O. (2014). Statistical methods for temporal and space–time analysis of community composition data. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778), 20132728.
10. Neumeyer, N., Omelka, M., & Hudecová, Š. (2019). A copula approach for dependence modeling in multivariate nonparametric time series. *Journal of Multivariate Analysis*, 171, 139-162.
11. Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110, 4-18.

Thanks for your attention

MUR PRIN 2017, Project "*Stochastic Models for Complex Systems*" (No. 2017JFFHSH)

MUR-PRIN 2022 PNRR, Project "*Stochastic Modeling of Compound Events*" (No. P2022KZJTZ)

ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing

Finanziato dall'Unione europea NextGenerationEU

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

MUR

UNIVERSITÀ DEL SALENTO