Environmental and Ecological Statistics (2024) 31:801–817
https://doi.org/10.1007/s10651-024-00626-6

Tail-dependence clustering of time series with spatial constraints

Alessia Benevento[1] · **Fabrizio Durante**[2] · **Roberta Pappadà**[3]

Received: 26 February 2024 / Revised: 27 April 2024 / Accepted: 11 May 2024 /
Published online: 16 June 2024

Abstract
We introduce a clustering method for time series based on tail dependence. Such a method also considers spatial constraints by means of a suitable procedure merging temporal and spatial dependence via extreme-value copulas. The cluster composition depends on the choice of the hyper-parameter $\alpha \in (0, 1)$ used to calibrate the contribution of the spatial dependence to the overall dissimilarity. A novel heuristic approach to select α based on a suitable connectedness index associated to each cluster of the partition is proposed.

Keywords Copula · Hierarchical clustering · Spatial statistics · Tail dependence · Time series

1 Introduction

Clustering algorithms are routinely run to summarize and visualize important spatial and/or temporal patterns in the climate sciences (Straus 2019). From a risk perspective, such methods are particularly useful in identifying extreme weather events for a

Handling Editor: Luiz Duczmal.

✉ Roberta Pappadà
rpappada@units.it

Alessia Benevento
alessia.benevento@unisalento.it

Fabrizio Durante
fabrizio.durante@unisalento.it

[1] Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy

[2] Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento, Lecce, Italy

[3] Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "B. de Finetti", Università degli Studi di Trieste, Trieste, Italy

better understanding of the underlying processes that can help to mitigate potentially severe impacts on the society (Field 2012).

Cluster analysis for extreme climate events has been mainly used to partition an entire region into homogeneous sub-regions (i.e. regionalization) based on similarities in dependence structure. For instance, a novel clustering algorithm for heavy rainfall is presented in Bernard et al. (2013) (see also Cooley et al. 2006). A similar algorithm has been also used in Bador et al. (2015) to analyze the maxima of summer temperatures across Europe. The issue of regionalization techniques in an extreme-value context has been also discussed in Saunders et al. (2021), where a hierarchical clustering procedure is applied. Apart from the approach based on extreme-value theory, the use of copulas has been also beneficial in this respect (see, for instance, Di Lascio et al. 2017; Pappadà et al. 2018; Palacios-Rodriguez et al. 2023). As a matter of fact, cluster methods with spatial constraints can be of interest in economics and finance (see, e.g., Asgharian et al. 2013; Fernández-Avilés et al. 2012; Hüttner et al. 2020; Kopczewska 2022).

The aim of the present work is to develop a clustering algorithm for identifying the extreme joint behaviour among various time series. Specifically, we adopt agglomerative hierarchical clustering procedures, which are particularly convenient in the dependence framework (De Keyser and Gijbels 2023; De Luca and Zuccolotto 2021, 2023; Kojadinovic 2004; Fuchs et al. 2021; Fuchs and Wang 2023). As a major aspect, our methodology combines the information on both the cross-sectional dependence and the spatial proximity. In fact, there are sometimes well grounded reasons to require the clusters to be composed of contiguous sites in order to create, for instance, spatial risk maps of administrative units, economic areas, etc. (D'Urso and Vitale 2020; Fouedjio 2020; Guénard and Legendre 2022). In addition, in some cases, the geographical contiguity may reflect the inherent structure of the phenomena and may serve as a proxy of the true dependence in the case of scarce/missing data.

To guarantee a clustering solution that can aid (without strictly enforce) the identification of sub-groups of time series that are also spatially related, the clustering algorithms are usually modified so that the input dissimilarity matrix is a combination of geographical and non-geographical information (Bourgault et al. 1992; Chavent et al. 2018; Oliver and Webster 1989). In a copula-based framework, the COFUST algorithm has been introduced in Disegna et al. (2017) to group time series by joining temporal and spatial information by means of a partitioning-around-medoids algorithm. This latter algorithm has been modified in Benevento and Durante (2024) by using the Wasserstein distance. In a hierarchical framework, Benevento and Durante (2023) propose to merge spatial and temporal information into suitable correlation matrices that take into account the underlying geometric structure of correlation matrix space. Moreover, in Di Lascio et al. (2023), spatial weighting is added to the dissimilarity matrix based on copula parameters. In the framework of tail dependence, preliminary results have been also developed in Benevento et al. (2023) and Zuccolotto et al. (2023).

Here, we develop a hierarchical clustering algorithm to merge temporal and spatial dependence in the framework of joint tail dependence coefficients (Durante et al. 2015a). Such coefficients quantify the probability that one time series is taking on very

large values given that another one is taking on large values. Specifically, we proceed as follows: (a) we assume that the temporal dependence can be conveniently represented by an *extreme-value copula* C, which can provide the upper tail dependence coefficient associated with C (Gijbels et al. 2020; Gudendorf and Segers 2012; Zhang et al. 2008); (b) then we propose a dissimilarity index based on a modification of the tail dependence coefficient of C that takes into account spatial information; (c) finally, we use the obtained dissimilarity matrix as an input for the hierarchical clustering procedure proposed in Bien and Tibshirani (2011). The latter algorithm has a number of advantages such as the interpretability of the obtained clustering in terms of the prototypes.

The proposed methodology is fully described in Sect. 2. Section 3 illustrates the methodology by means of an empirical analysis of geo-referenced data with observed maxima and highlights the main advantages of the proposed methodology. Section 4 presents some final remarks.

2 The methodology

We aim to cluster n different units that are represented by a univariate time series (the temporal feature) and a vector of (spatial) features embedding the information about the geographic location. For instance, the units are sites in which the temperature is recorded, while the spatial information is the geographic position of each weather station where the measurements are collected.

Thus, the starting point is represented by:

- a $(T \times n)$ (temporal) matrix, $\mathbf{X} = (x_{ti})$, whose element x_{ti} represents the value at time t $(t = 1, \ldots, T)$ of the temporal feature for the i-th unit $(i = 1, \ldots, n)$; each column of $\mathbf{X}$ is a time series;
- the p–dimensional vectors $\mathbf{s}_1^\top, \ldots, \mathbf{s}_n^\top$ associated with each time series in $\mathbf{X}$ representing the geographic information.

The clustering procedure consists of the following steps: (1) modeling of the temporal dependence within each univariate series; (2) construction of the cross-sectional temporal (extremal) dependence model; (3) construction of the spatial dependence; (4) gluing of temporal and spatial dependence into a global model; (5) extraction of the dissimilarity matrix and selection of a dissimilarity-based clustering algorithm yielding a final partition. These steps are detailed below.

2.1 Modeling univariate time series

Following a classical copula-based time series model (Patton 2012; Neumeyer et al. 2019), we assume that, for $i = 1, \ldots, n$, the i–th time series $(x_{ti})_{t=1,\ldots,T}$ is generated by the stochastic process

$$X_{ti} = \mu_i(\mathbf{Z}_t) + \sigma_i(\mathbf{Z}_t)\varepsilon_{ti}, \tag{1}$$

where $\mu_i(\cdot)$ and $\sigma_i(\cdot)$ represent the conditional mean and standard deviation of the i-th time series, and the covariate $\mathbf{Z}_t$ may include past values of the process X_{ti} at different lags or other exogenous variables. For every i, the innovations ε_{ti} are distributed according to a marginal law $F_{ti} = F_i$ for every t (having mean zero and variance one, for identification) such that, for every t, the joint distribution function of $(\varepsilon_{t1}, \dots, \varepsilon_{tn})$ can be expressed in the form $C(F_1, \dots, F_n)$ for some copula C.

Notice that C cannot be directly estimated from the original time series, since it is necessary to disentangle the dependence from the marginal effects. To this end, we proceed as follows:

(i) First, we estimate the marginal serial dependence (including possible trend and seasonal cycles) from each time series with a model of type (1) by obtaining the fitted $\hat{\mu}_i$ and $\hat{\sigma}_i$. To validate the model, tests of homoscedasticity and uncorrelatedness are carried out to ensure that the residuals can be considered (approximately) as a sample of independent and identically distributed random variables.

(ii) The inference about the copula C is thus based on the estimated residuals extracted from the previous step, given by

$$\widehat{\varepsilon}_{ti} = (X_{ti} - \widehat{\mu}_i(\mathbf{Z}_t))/\widehat{\sigma}_i(\mathbf{Z}_t), \tag{2}$$

that are transformed into the *pseudo-observations*, $z_{ti} = F_i(\widehat{\varepsilon}_{ti})$, where F_i may be estimated from a parametric model (Gaussian, Student t, etc.) or by using the empirical distribution function.

As a result, $(z_{t1}, \dots, z_{tn})_{t=1,\dots,T}$ contains the information about the link (i.e. the copula) among the time series under consideration (see, e.g., Rémillard 2017) and can be used for estimating the associated tail dependence.

2.2 Construction of the temporal copula

Once the time series have been converted into pseudo-observations, a suitable copula C^{ts} is estimated to describe the dependence among them. Here, focusing on the extremal dependence, we assume that C^{ts} is an n–dimensional extreme-value (shortly, EV) copula (Gudendorf and Segers 2010). In particular, when extremes are of interest, it is natural to study the coefficients of tail dependence, which are conveniently expressed for bivariate EV copulas (Durante et al. 2015b; Frahm et al. 2005).

Any EV copula C can be written, for every $\mathbf{u} \in [0,1]^n$, as

$$C(\mathbf{u}) = \exp\left(\sum_{i=1}^{n} \log(u_i) A\left(\frac{\log(u_1)}{\sum_{i=1}^{n}\log(u_i)}, \dots, \frac{\log(u_n)}{\sum_{i=1}^{n}\log(u_i)}\right)\right) \tag{3}$$

for a convex function A defined on the n–dimensional unit simplex, called *Pickands dependence function* (see, e.g., Pickands 1981). Now, the dependence function A associated with C conveys the information about the extremal dependence and, as

such, it allows us to obtain all the pairwise tail dependence coefficients. In fact, for each $i, i' \in \{1, \dots, n\}$, the upper tail dependence coefficient λ_U associated with $C_{ii'}$, the copula of the i–th and i'–th component of C is given by

$$\lambda_U(C_{ii'}) = 2(1 - A_{ii'}(1/2, 1/2)), \tag{4}$$

where $A_{ii'}$ is the bivariate function obtained from A when we set all the coordinates equal to 0 up to the i-th and i'-th coordinates that are equal to 1/2.

For EV copulas, the estimation of the pairwise tail dependence coefficient may be based on the estimation of the Pickands dependence function. To this end, we consider the madogram estimator $\widehat{A}^{\mathrm{MD}}$ discussed in Marcon et al. (2017) (see also Gijbels et al. 2020, section 7) and implemented in Beranger et al. (2023). Specifically, the nonparametric estimator $\widehat{A}^{\mathrm{MD}}$ of the multivariate Pickands dependence function is given, for every $\mathbf{w}$ in the n–dimensional simplex, by

$$\widehat{A}^{\mathrm{MD}}(\mathbf{w}) = \frac{\widehat{\nu}(\mathbf{w}) + c(\mathbf{w})}{1 - \widehat{\nu}(\mathbf{w}) - c(\mathbf{w})}, \tag{5}$$

where

$$\widehat{\nu}(\mathbf{w}) = \frac{1}{T} \sum_{t=1}^{T} \left(\bigvee_{i=1}^{n} z_{ti}^{1/w_i} - \frac{1}{n} \sum_{i=1}^{n} z_{ti}^{1/w_i} \right),$$

with z_{ti} pseudo-observations and $c(\mathbf{w}) = n^{-1} \sum_{i=1}^{n} w_i/(1 + w_i)$. Notice that, in presence of missing data, the procedure can be adapted as explained in Boulin et al. (2022).

Thus, the resulting estimation of the upper tail dependence coefficient based on the temporal information for the i–th and i'-th time series is given by

$$\widehat{\lambda}^{\mathrm{ts}}_{ii'} = 2(1 - \widehat{A}^{\mathrm{MD}}_{ii'}(1/2, 1/2)), \tag{6}$$

where $\widehat{A}^{\mathrm{MD}}_{ii'}$ is obtained by calculating $\widehat{A}^{\mathrm{MD}}$ on the vector with all elements equal to 0 up to the i-th and i'-th coordinates that are equal to 1/2.

Remark 1 In the literature, both parametric (see, e.g., De Luca and Zuccolotto 2011) and non-parametric estimators (see, e.g., Durante et al. 2015b) have been considered for the pairwise tail dependence coefficients in the clustering framework. The advantage of considering a global (i.e. n–dimensional) estimator for the Pickands function A rather than pairwise functions, is that the resulting matrix $(\widehat{\lambda}^{\mathrm{ts}}_{ii'})$ may naturally preserve the geometric properties of the upper tail dependence matrices (Embrechts et al. 2016; Fiebig et al. 2017).

2.3 Construction of the spatial copula

In order to include the spatial information, the main idea of the proposed methodology is to model the spatial dependence using another n–dimensional EV copula C^{sp} that can embed the spatial information contained in $(\mathbf{s}_1^\top, \dots, \mathbf{s}_n^\top)$. To this end, we

consider the copula associated with the classical max-stable model by Smith (1990). This is the so-called Hüsler-Reiss copula (Hüsler and Reiss 1989), whose expression was recently presented in a convenient form in Nikoloulopoulos et al. (2009), section 2.3.

For such copulas, the (pairwise) upper tail dependence coefficient associated with the i–th and i'–th component is given by

$$\lambda_U(C^{\text{sp}}_{ii'}) = 2 - 2\Phi\left(\frac{((\mathbf{s}_i - \mathbf{s}_{i'})^\top \Sigma^{-1}(\mathbf{s}_i - \mathbf{s}_{i'}))^{1/2}}{2}\right), \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution (Schlather and Tawn (2003), page 147); Σ is the covariance matrix of the bivariate standard normal distribution associated with the observations at sites i and i'. In the present context, since we focus on the mere spatial information, we set $\Sigma = I_2$, the identity matrix. In such a way, $\lambda_{ii'}$, which corresponds to λ_U in Eq. (7) calculated for the pair (i, i'), is only expressed as a function of the distance $h_{ii'} = \|\mathbf{s}_i - \mathbf{s}_{i'}\|_2$. Hereinafter, we estimate it via the formula

$$\widehat{\lambda}^{\text{sp}}_{ii'} = 2 - 2\Phi\left(\frac{h_{ii'}}{2}\right). \tag{8}$$

Clearly, as $h_{ii'} \to +\infty$, $\lambda_{ii'} \to 0$.

Remark 2 Notice that, in practice, in order to determine a convenient way to use the distance information, it may be useful to standardize the geographic coordinates.

2.4 Construction of the spatio-temporal copula

Given the tail dependence matrices $(\widehat{\lambda}^{\text{ts}}_{ii'})$ and $(\widehat{\lambda}^{\text{sp}}_{ii'})$ and a fixed $\alpha \in [0, 1]$, we propose to model the spatio-temporal dependence via a suitable n–dimensional copula $C^{(\alpha)}$ that merges the EV copulas C^{ts} and C^{sp}.

In the general case, copulas can be combined via a convex combination (Durante and Sempi 2016). However, since the class $\mathcal{C}_{EV}$ of EV copulas is not closed under convex combinations, this latter approach cannot be replicated here. Therefore, we adopt a different operation, known as Khoudraji's device (Khoudraji 1995) (see also Durante 2009; Liebscher 2008) given by

$$\varphi_\alpha : \mathcal{C}_{EV} \times \mathcal{C}_{EV} \to \mathcal{C}_{EV}, \quad \varphi_\alpha(C, D)(\mathbf{u}) = C(\mathbf{u}^{1-\alpha})D(\mathbf{u}^\alpha). \tag{9}$$

Now, if both copulas represent the comonotonic (respectively, independent) case, i.e. $C(\mathbf{u}) = D(\mathbf{u}) = \min_{i=1,\ldots,n} u_i$ (respectively, $C(\mathbf{u}) = D(\mathbf{u}) = \prod_{i=1,\ldots,n} u_i$), then the resulting copula is the comonotonicity (respectively, independent) copula.

Thus, $C^{(\alpha)} = \varphi_\alpha(C^{\text{ts}}, C^{\text{sp}})$ represents an n–dimensional copula that merges both the temporal and the spatial information. Its parameter $\alpha \in [0, 1]$ represents the weight assigned to the spatial component; in particular, when $\alpha = 0$ the spatial information is not relevant.

As a matter of fact, the bivariate dependence function $A_{ii'}^{(\alpha)}$ associated with $C^{(\alpha)}$ is given by

$$A_{ii'}^{(\alpha)}(w_1, w_2) = (1-\alpha)A_{ii'}^{\text{ts}}(w_1, w_2) + \alpha A_{ii'}^{\text{sp}}(w_1, w_2),$$

i.e., it is the convex combination of the dependence functions associated with $C_{ii'}^{\text{ts}}$ and $C_{ii'}^{\text{sp}}$ (see Genest et al. 1998). Therefore, the upper tail dependence coefficient of $C_{ii'}^{(\alpha)}$ is given by

$$\lambda_U(C_{ii'}^{(\alpha)}) = (1-\alpha)\lambda_U(C_{ii'}^{\text{ts}}) + \alpha\lambda_U(C_{ii'}^{\text{sp}}). \tag{10}$$

It turns out that the tail dependence coefficient of the spatio-temporal model for the i–th and i'-th time series can be estimated via Eq. (6) and Eq. (8) as

$$\widehat{\lambda}_{ii'}^{(\alpha)} = (1-\alpha)\widehat{\lambda}_{ii'}^{\text{ts}} + \alpha\widehat{\lambda}_{ii'}^{\text{sp}}. \tag{11}$$

2.5 Application of the clustering algorithm

To cluster the time series under (soft) spatial constraints, we propose to use a cluster algorithm over the tail-dependence dissimilarity matrix $(\widehat{\lambda}_{ii'}^{(\alpha)})$ derived from the spatio-temporal copula $C^{(\alpha)}$, in analogy with some previous works (see, e.g., D'Urso et al. 2023; De Luca and Zuccolotto 2011, 2017; Durante et al. 2015b).

Thus, for every i, i' and every $\alpha \in [0,1]$, the dissimilarity $\Delta_{ii'}^{(\alpha)}$ between the i-th and i'-th time series can be defined by means of a decreasing function $f : [0,1] \to [0, +\infty]$ with $f(1) = 0$, so that

$$\Delta_{ii'}^{(\alpha)} = f(\widehat{\lambda}_{ii'}^{(\alpha)}). \tag{12}$$

In particular, maximal tail dependence (i.e. $\widehat{\lambda}_{ii'}^{(\alpha)} = 1$) corresponds to minimal dissimilarity.

Following De Luca and Zuccolotto (2011), a convenient choice is $f(x) = -\ln(x)$, which we adopt here and corresponds to

$$\Delta_{ii'}^{(\alpha)} = -\ln((1-\alpha)\widehat{\lambda}_{ii'}^{\text{ts}} + \alpha\widehat{\lambda}_{ii'}^{\text{sp}}). \tag{13}$$

Notice that $\Delta_{ii'}^{(\alpha)} = +\infty$ when both temporal and spatial copula are tail-independent (i.e. the tail dependence coefficient is equal to 0).

The dissimilarity matrix $\mathbf{\Delta}^{(\alpha)} = (\Delta_{ii'}^{(\alpha)})_{i,i'=1,\ldots,n}$ can be used as an input for a dissimilarity-based hierarchical clustering algorithm (Murtagh and Contreras 2017). As is known, such a method provides a representation in a binary tree (one or two child nodes at each non-terminal node) commonly referred to as a *dendrogram*. From the tree, we can recover n possible clusterings, where n is the number of objects in the clustering.

Among various hierarchical clustering algorithms, we suggest applying the method based on minimax linkage (Bien and Tibshirani 2011). This algorithm has a few advantages with respect to similar methods for the literature. In particular,

(a) monotone transformation of the dissimilarity matrix leaves the clustering unchanged (Bien and Tibshirani 2011, Property 4), so that the choice of the function f in Eq. (12) is not relevant;
(b) the resulting dendrogram does not admit inversion (Bien and Tibshirani 2011, Property 2), so it can be easily interpreted;
(c) with each node of the dendrogram tree, we have an associated *prototype*, namely the most central data point of its cluster, which eases the cluster interpretation.

Compared to other popular algorithms, agglomerative hierarchical methods based on single and complete linkage also satisfy (a) and (b), while average linkage does not satisfy (a). All these three linkage methods do not admit, however, a natural centroid/prototype. In contrast, centroid linkage does not satisfy (a) and (b).

2.6 Hyper-parameter selection

As underlined in Fouedjio (2020), the cluster composition depends on the choice of the hyper-parameter $\alpha \in (0, 1)$ that is used in the computation of the dissimilarity between two data locations (see Eq. (13)). In particular, when α increases, the cluster composition tends to favor the geographic proximity of the considered stations, due to the impact of the spatial component. Thus, roughly speaking, one needs to make sure that this parameter is not too small and not too large. For instance, in Romary et al. (2015) it is suggested to put a 5–30% on the geographical coordinates and the remainder to the attributes.

The choice of α should depend on a trade-off between the preservation of the spatial contiguity and the need to identify the units whose behaviour is only partially driven by geographic proximity. To provide some heuristics for the selection of α we rely on a kind of connectedness index of the cluster partition.

To this end, from the spatial matrix $(\mathbf{s}_1^\top, \dots, \mathbf{s}_n^\top)$, we derive the $n \times n$ contiguity matrix that assigns value 1 to its (i, j) entry if the units i and j are contiguous, i.e. spatially adjacent, and value 0, otherwise. Such a matrix is symmetric and can be interpreted as an undirected graph whose vertices are the sites where data are collected. In particular, we can count the number of connected components associated with this graph (see, e.g., Bollobás 1998), which are composed of units belonging to the same macro-area.

Now, consider the set of units with indices $\{1, 2, \dots, n\}$ that are partitioned according to the algorithm in Sect. 2.5 that depends on the (hyper-)parameter $\alpha \in [0, 1]$. For every $\alpha \in [0, 1]$, let $\mathcal{P}^{(\alpha)} = \{G_1^{(\alpha)}, \dots, G_{K^{(\alpha)}}^{(\alpha)}\}$ be the resulting partition into $K^{(\alpha)}$ clusters, $K^{(\alpha)} < n$. We consider the function $\xi_\alpha : \{1, \dots, K^{(\alpha)}\} \to \{1, \dots, n\}$ that assigns to each $k \in \{1, \dots, K^{(\alpha)}\}$, the number of connected components of

cluster $G_k^{(\alpha)}$, n_c^k, with $1 \leq n_c^k \leq n_k$, where $n_k \in \{1, \ldots, n\}$ denotes the cardinality of $G_k^{(\alpha)}$. Finally, we give the following definition.

Definition 1 For $\alpha \in [0, 1]$, the *connectedness index* of $\mathcal{P}^{(\alpha)}$ is the average number of connected components of each element of the partition, i.e.

$$\text{Conn}(\mathcal{P}^{(\alpha)}) = \frac{1}{K^{(\alpha)}} \sum_{k=1}^{K^{(\alpha)}} \xi_\alpha(k).$$

Clearly, if each element of the partition $\mathcal{P}^{(\alpha)}$ is connected, then $\text{Conn}(\mathcal{P}^{(\alpha)}) = 1$. The theoretical upper bound of the index is given by the partition with one element (cluster) formed of n disconnected components.

Starting with the set of n units, the connectedness index is expected to decrease at the increase of α, reaching 1 as α goes to 1. As we will show in the illustration of Sect. 3.2, the graph of the function $\alpha \mapsto \text{Conn}(\mathcal{P}^{(\alpha)})$ can provide some insights about the selection of the value α. For instance, it may allow the identification of the value of α for which the connectedness index sharply decreases and/or the minimal value of α that guarantees a spatially contiguous partition.

3 Illustration

In this section, we apply the proposed algorithm on monthly maxima precipitation data collected throughout the Italian territory. Specifically, the data have been downloaded from the Climate Data Store,[1] which collects global climate and weather data of the past 8 decades. They consist of monthly precipitations from January 2011 till November 2023. The considered data spread on $n = 527$ grid points.

Following an approach similar to the one in Bador et al. (2015), we de-trend the observed precipitations following a two-step procedure. First, we remove the multi-year climatological average from monthly precipitation maxima within the data-set. Then, from these residuals, we remove the monthly running average. After this de-trend process, we have a collection of time series of length $T = 155$.

Given the set of de-trended time series and the geographical locations, we proceed in two phases:

(a) we compute the temporal copula as described in Sect. 2 and the associated dissimilarity matrix. Once the dissimilarity matrix has been obtained, we apply four hierarchical clustering algorithms based, respectively, on single, complete, average, and minimax linkage. As known, this latter approach will also provide the prototypes associated with each cluster.
(b) We compute the spatial copula as described in Sect. 2. Then, we compute the copula $C^{(\alpha)}$ merging the spatial and temporal information for different values $\alpha \in [0, 1]$ and use it to compute the dissimilarity matrix $\mathbf{\Delta}^{(\alpha)}$ as in (12).

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

3.1 Comparison among hierarchical methods on the pure temporal data

As mentioned, in phase (a) we use the time series to compare different hierarchical clustering algorithms. With the pure temporal dissimilarity matrix at hand, which is obtained from Eq. (13) for $\alpha = 0$, we apply the four algorithms based on the linkages mentioned above. In a hierarchical clustering procedure, the cluster composition is obtained by cutting the dendrogram tree at a given height h. Notice that, according to the linkage method, this will result in a different interpretation of the obtained groups. Indeed, cutting at height h gives a clustering such that: it is not possible to find two clusters having points closer than h from each other if the single linkage is used; all points of a cluster are within h of one another, if the complete linkage is used; all points in each cluster are within dissimilarity h from their prototype, when the prototype-based method is chosen.

To visualize the results on the Italian map, we need to select a suitable number of clusters (we let k vary between 5 and 15). To this end, we adopt the Dunn index, which is a scalar that formalizes the idea of a ratio between between-cluster separation and within-cluster compactness for general dissimilarity input data and a fixed number of clusters (Dunn 1974; Hennig et al. 2015). Specifically, we adopt an index belonging to the family of Dunn indexes implemented in the `fpc` R package (Hennig 2023) and defined as the ratio of minimum average dissimilarity between two clusters and maximum average within cluster dissimilarity. The number of clusters that maximizes the Dunn index is used to cut the dendrogram associated with each hierarchical clustering. The resulting groups are shown in Fig. 1 on the Italian map. As can be seen, the partitions produced by the single and complete linkages are formed of a larger number of clusters that appear more unbalanced compared to those obtained via the minimax and average linkages. The agreement between the different partitions is measured by the Adjusted Rand Index (ARI), where the maximum agreement is achieved when such index equals one (Hubert and Arabie 1985). The ARI between the minimax and the single, the average, and the complete linkage is 0.03, 0.67, and 0.51, respectively.

As explained in Sect. 2.5, we exploit the minimax linkage, whose main advantage is to provide cluster prototypes having the desirable property that cutting at height h will guarantee that no point will be farther than h from its prototype. The minimax linkage hierarchical clustering, implemented via the `R` package `protoclust` (Bien and Tibshirani 2022), produces the 5-cluster solution displayed in Fig. 1. When $\alpha = 0$, the resulting partition does not carry information concerning the geographic proximity of the units being clustered, even though some units (time series) belong to the same macro-area. By contrast, when $\alpha = 1$ the cluster composition is completely determined by the spatial dependence. Hence, a convenient choice of α can be a value yielding a suitable compromise between the two extreme cases, for which we propose a possible approach based on the representation of groups in terms of their network structure.

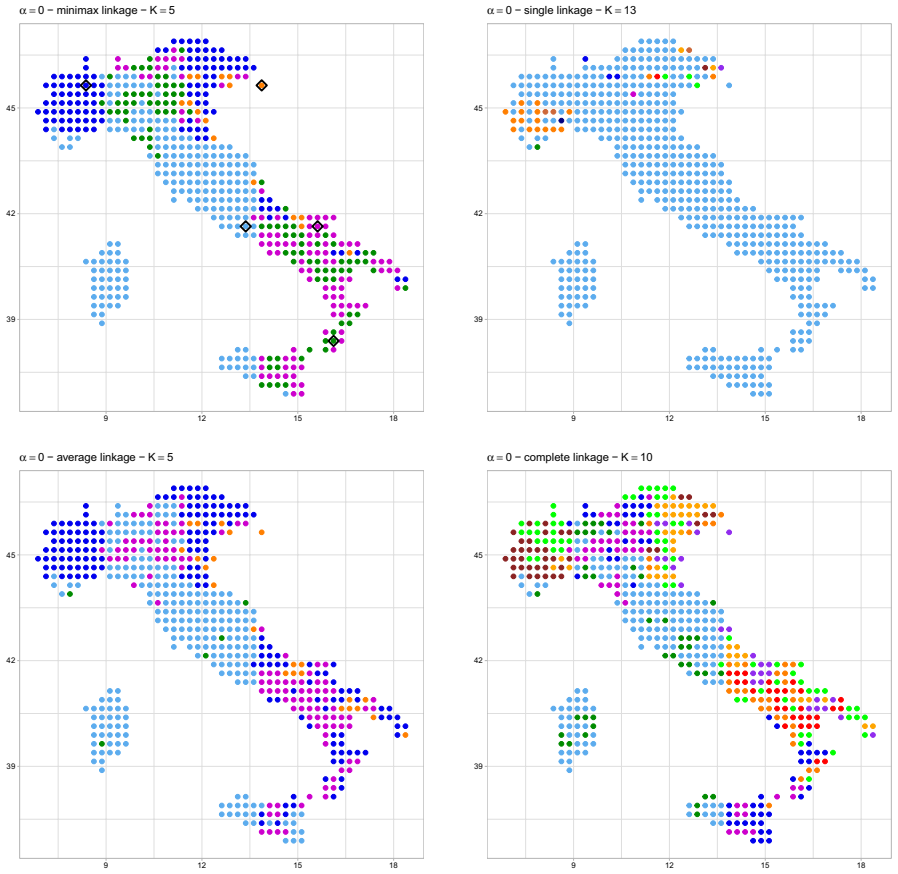

Fig. 1 Representation of clusters in the pure temporal case with minimax linkage ($K = 5$) along with cluster prototypes marked on the map (upper left), single linkage with $K = 13$ (upper right), average linkage with $K = 5$ (lower left), and complete linkage $K = 10$ (lower right)

3.2 Effects of the spatial constraints for different weights

As mentioned in Sect. 2.6, plotting the connectedness index $\alpha \mapsto \text{Conn}(\mathcal{P}^{(\alpha)})$ for different values of the hyper-parameter α can help identifying the values α for which the resulting partitions are (almost) spatially contiguous. Indeed, when α increases the geographic proximity strongly influences the associated clustering based on Eq. (13), thus producing clusters that are well-identified as geographic regions. The connectedness index for $\alpha \in \{0, 0.02, 0.04, \ldots, 1\}$ is reported in Fig. 2. The plot shows that the connected index stabilizes around 1 for α above a certain value, that is, all the clusters in the partition $\mathcal{P}^{(\alpha)}$ represent a connected component formed of spatially adjacent units. Note that two partitions obtained for two different values of α having connected index equal to 1 may differ in the cluster composition, even though the spatial contiguity is preserved.

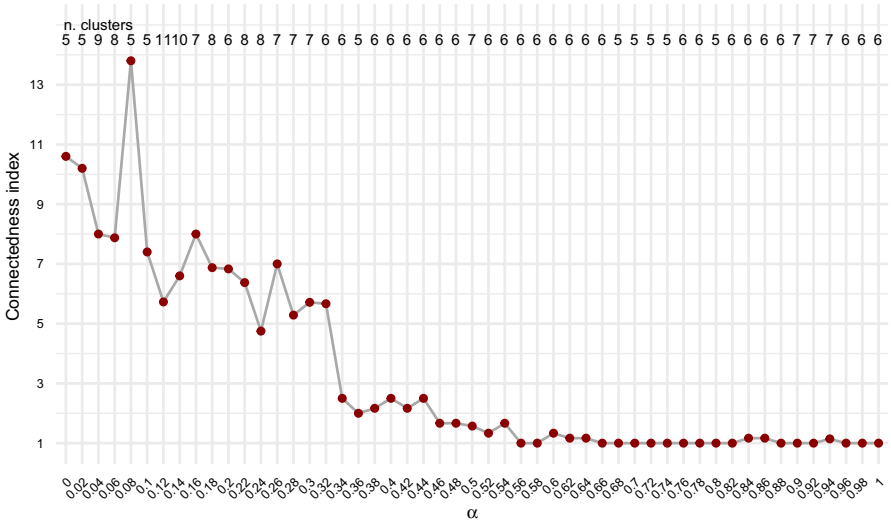

Fig. 2 Connectedness index for $\alpha \in \{0, 0.02, 0.04, \ldots, 1\}$. The number of clusters of each partition $\mathcal{P}^{(\alpha)}$ is reported on the graph

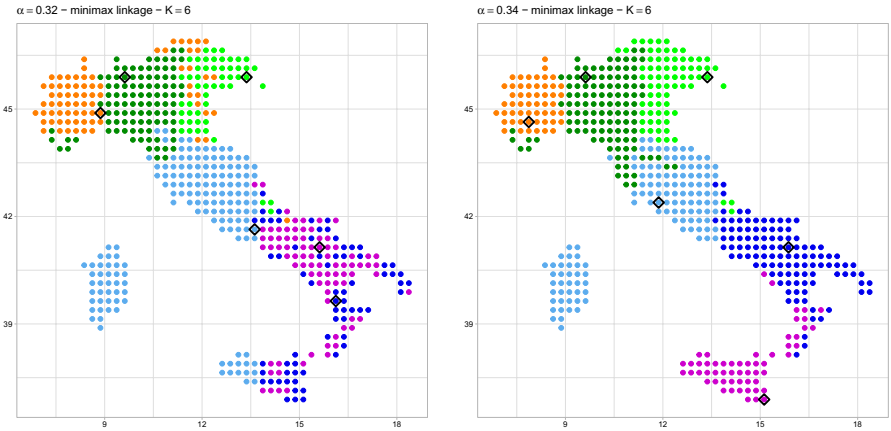

Fig. 3 Representation of clusters for $\alpha = 0.32$ (left plot), $\alpha = 0.34$ (right plot). The marked points are the prototypes of each group

To account for both the spatial and temporal component, the optimal α can be selected by choosing the value immediately before a sharp decrease in $\alpha \mapsto \text{Conn}(\mathcal{P}^{(\alpha)})$, that is, the last value for which temporal dependence has a non-negligible impact on the final clustering. Hence, from Fig. 2 we select $\alpha = 0.32$ and display the resulting partition into $k = 6$ clusters in Fig. 3 (for completeness, the clustering obtained for $\alpha = 0.34$ is also reported).

The results obtained by choosing the value $\alpha = 0.32$ based on the connectedness index suggest dividing the Italian territory into six areas that appear spatially consistent and take into account the temporal extreme dependence, although the spatial cohesion is only partially attained in the final clustering. A major advantage of our proposal is that it allows us to improve the configuration based on pure temporal clustering, by including relevant information on the spatial dependence among the time series. In addition, by looking at the cluster composition when α varies from 0 to 1, one can discover atypical data points within a dataset, i.e. time series that belong to a cluster that is different than that of the neighbors, regardless of any spatial constraint.

4 Conclusions

This paper proposes a new dissimilarity measure to cluster time series observed at different spatial locations. Our focus is on the extreme joint behavior that characterizes a set of climate time series and, specifically, on the identification of clusters of time series of precipitation maxima through extreme-value copulas, used to characterize the (extremal) dependence structure in the data.

Being based on copulas, our approach is not affected by the choice of the marginal models for the univariate time series and, hence, it allows us to disentangle the dependence from the marginal effects. Both the spatial and the temporal dependence are taken into account to build a spatio-temporal copula and estimate the coefficients of (upper) tail dependence. In particular, we introduce a tail-dependence dissimilarity measure that combines the information on both the cross-sectional dependence and the spatial proximity, according to a hyper-parameter $\alpha \in [0, 1]$. This aspect is relevant when the interest is in finding clusters that may reflect a reasonable compromise between spatial contiguity and the cross-sectional temporal dependence among the time series that ignores the geographic information on the observed phenomena.

Concerning the value α yielding a final clustering, it should be mentioned that there exist a few approaches suggested in the literature to set such a parameter, which poses a trade-off between the spatial and temporal contribution to the overall dissimilarity. On the one hand, the value of α may be arbitrarily set by the user to assign a larger weight to the spatial (or temporal) aspect; on the other hand, one can be interested in finding groups that are only partially driven by geographic proximity. Following the latter approach, we provide some heuristics for the selection of α that rely on the notion of connected components from graph theory.

The real data example, concerning precipitation maxima observed throughout the Italian territory, illustrates the usefulness and effectiveness of the suggested dissimilarity measure, which is embedded into a classical hierarchical framework. Among the available linkages, we suggest the adoption of the minimax linkage, whose main advantage is to naturally provide prototypical units for each cluster in the final partition.

The main contribution of the proposed approach is twofold: on the one hand, the final cluster can be used to develop more effective risk-mitigation strategies, arising from a combined spatio-temporal dependence modeling; on the other hand,

the procedure can be used as a tool for detecting anomalies in spatially consistent regions. The latter issue represents a relevant research direction to be explored.

Acknowledgements AB acknowledges the support of Regione Puglia (Italy) via the Programma Regionale "RIPARTI (assegni di RIcerca per riPARTire con le Imprese)"—research project "FIRST: a Framework for Innovation in Risk management to support Territories" (code: c19a5daa). FD has 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 the partial financial support from ICSC—Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union—Next Generation EU (CUP F83C22000740001). FD is also member of the group GNAMPA of INdAM (Istituto Nazionale di Alta Matematica). 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.

Author contributions A. Benevento, F. Durante, and R. Pappadà conceptualized the research methodology; data collection, cleaning, statistical analysis, and data visualization were performed by A. Benevento. F. Durante and R. Pappadà contributed to software validation and formal analysis. All authors contributed to the paper's writing.

Funding Open access funding provided by Università degli Studi di Trieste within the CRUI-CARE Agreement.

Data availability The data used in the current study are available at https://cds.climate.copernicus.eu/.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Asgharian H, Hess W, Liu L (2013) A spatial analysis of international stock market linkages. J Bank Financ 37(12):4738–4754

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

Benevento A, Durante F (2023) Correlation-based hierarchical clustering of time series with spatial constraints. Spatial Stat 59:100797

Benevento A, Durante F (2024) Wasserstein dissimilarity for copula-based clustering of time series with spatial information. Mathematics 12(1):67

Benevento A, Durante F, Pappadà R (2023) An approach to cluster time series extremes with spatial constraints. In: Chelli F, Ciommi M, Ingrassia S et al (eds) Book of Short Papers SEAS IN 2023. Pearson, pp 679–684

Beranger B, Padoan S, Marcon G (2023) ExtremalDep: extremal dependence models. R package version 0.0.4-1

Bernard E, Naveau P, Vrac M et al (2013) Clustering of maxima: spatial dependencies among heavy rainfall in France. J Clim 26(20):7929–7937
Bien J, Tibshirani R (2011) Hierarchical clustering with prototypes via minimax linkage. J Am Stat Assoc 106(495):1075–1084
Bien J, Tibshirani R (2022) protoclust: hierarchical clustering with prototypes. https://CRAN.R-project.org/package=protoclust, R package version 1.6.4
Bollobás B (1998) Modern graph theory, Grad. Texts Math., vol 184. Springer, New York
Boulin A, Di Bernardino E, Laloë T et al (2022) Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework. J Multivar Anal 192:21
Bourgault G, Marcotte D, Legendre P (1992) The multivariate (co)variogram as a spatial weighting function in classification methods. Math Geol 24(5):463–478
Chavent M, Kuentz-Simonet V, Labenne A et al (2018) ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat 33(4):1799–1822
Cooley D, Naveau P, Poncet P (2006) Variograms for spatial max-stable random fields. In: Bertail P, Soulier P, Doukhan P (eds) Dependence in probability and statistics. Springer, New York, pp 373–390
De Keyser S, Gijbels I (2023) Hierarchical variable clustering via copula-based divergence measures between random vectors. Int J Approx Reason 165:109090
De Luca G, Zuccolotto P (2011) A tail dependence-based dissimilarity measure for financial time series clustering. Adv Data Anal Classif 5(4):323–340
De Luca G, Zuccolotto P (2017) Dynamic tail dependence clustering of financial time series. Stat Pap 58:641–657
De Luca G, Zuccolotto P (2021) Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach. Int J Approx Reason 139:88–103
De Luca G, Zuccolotto P (2023) Dynamic time series clustering with multivariate linkage and automatic dendrogram cutting using a recursive partitioning algorithm. Inf Sci 649:119605
Di Lascio FML, Durante F, Pappadà R (2017) Copula-based clustering methods. In: Úbeda Flores M, de Amo E, Durante F et al (eds) Copulas and dependence models with applications. Springer, New York, pp 49–67
Di Lascio FML, Menapace A, Pappadà R (2023) A spatially-weighted AMH copula-based dissimilarity measure for clustering variables: an application to urban thermal efficiency. Environmetrics 35:e2828
Disegna M, D'Urso P, Durante F (2017) Copula-based fuzzy clustering of spatial time series. Spatial Stat 21:209–225
Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. J Cybern 4(1):95–104
Durante F (2009) Construction of non-exchangeable bivariate distribution functions. Stat Pap 50(2):383–391
Durante F, Sempi C (2016) Principles of copula theory. CRC Press, Boca Raton
Durante F, Fernández-Sánchez J, Pappadà R (2015a) Copulas, diagonals and tail dependence. Fuzzy Sets Syst 264:22–41
Durante F, Pappadà R, Torelli N (2015b) Clustering of time series via non-parametric tail dependence estimation. Stat Pap 56(3):701–721
D'Urso P, Vitale V (2020) A robust hierarchical clustering for georeferenced data. Spatial Stat 35:100407
D'Urso P, De Luca G, Vitale V, et al (2023) Tail dependence-based fuzzy clustering of financial time series. Ann Oper Res
Embrechts P, Hofert M, Wang R (2016) Bernoulli and tail-dependence compatibility. Ann Appl Probab 26(3):1636–1658
Fernández-Avilés G, Montero JM, Orlov AG (2012) Spatial modeling of stock market comovements. Financ Res Lett 9(4):202–212
Fiebig UR, Strokorb K, Schlather M (2017) The realization problem for tail correlation functions. Extremes 20(1):121–168
Field C (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
Fouedjio F (2020) Clustering of multivariate geostatistical data. WIREs Comput Stat 12(5):e1510
Frahm G, Junker M, Schmidt R (2005) Estimating the tail-dependence coefficient: properties and pitfalls. Insurance Math Econ 37(1):80–100
Fuchs S, Wang Y (2024) Hierarchical variable clustering based on the predictive strength between random vectors. Int. J. Approx. Reason 170:109185

Fuchs S, Di Lascio FML, Durante F (2021) Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. Comput Stat Data Anal 159:107201
Genest C, Ghoudi K, LP Rivest (1998) Understanding relationships using copulas, by Edward Frees and Emiliano Valdez. N Am Actuar J 2(3):143–149
Gijbels I, Kika V, Omelka M (2020) Multivariate tail coefficients: properties and estimation. Entropy 22(7):728
Gudendorf G, Segers J (2010) Extreme-value copulas. In: Jaworski P, Durante F, Härdle WK et al (eds) Copula theory and its applications, vol 198. Lecture Notes in Statistics—Proceedings. Springer, Berlin, pp 127–145
Gudendorf G, Segers J (2012) Nonparametric estimation of multivariate extreme-value copulas. J Stat Plann Inference 142(12):3073–3085
Guénard G, Legendre P (2022) Hierarchical clustering with contiguity constraint in R. J Stat Softw 103(1):1–26
Hennig C (2023) fpc: flexible procedures for clustering. https://CRAN.R-project.org/package=fpc, R package version 2.2-10
Hennig C, Meila M, Murtagh F et al (2015) Handbook of cluster analysis. CRC Press, Boca Raton
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Hüsler J, Reiss RD (1989) Maxima of normal random vectors: Between independence and complete dependence. Stat Probab Lett 7(4):283–286
Hüttner A, Scherer M, Gräler B (2020) Geostatistical modeling of dependent credit spreads: estimation of large covariance matrices and imputation of missing data. J Bank Financ 118:105897
Khoudraji A (1995) Contributions à l'étude des copules et à la modélisation des valeurs extrêmes bivariées. PhD thesis, Université de Laval, Québec (Canada)
Kojadinovic I (2004) Agglomerative hierarchical clustering of continuous variables based on mutual information. Comput Stat Data Anal 46(2):269–294
Kopczewska K (2022) Spatial machine learning: new opportunities for regional science. Ann Reg Sci 68(11):713–755
Liebscher E (2008) Construction of asymmetric multivariate copulas. J Multivar Anal 99(10):2234–2250
Marcon G, Padoan SA, Naveau P et al (2017) Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials. J Stat Plann Inference 183:1–17
Murtagh F, Contreras P (2017) Algorithms for hierarchical clustering: an overview, II. WIREs Data Min Knowl Discov 7(6):e1219
Neumeyer N, Omelka M, Hudecová Š (2019) A copula approach for dependence modeling in multivariate nonparametric time series. J Multivar Anal 171:139–162
Nikoloulopoulos AK, Joe H, Li H (2009) Extreme value properties of multivariate t copulas. Extremes 12(2):129–148
Oliver MA, Webster R (1989) A geostatistical basis for spatial weighting in multivariate classification. Math Geol 21(1):15–35
Palacios-Rodriguez F, Di Bernardino E, Mailhot M (2023) Smooth copula-based generalized extreme value model and spatial interpolation for sparse extreme rainfall in central eastern canada. Environmetrics 34(3):e2795
Pappadà R, Durante F, Salvadori G et al (2018) Clustering of concurrent flood risks via hazard scenarios. Spatial Stat 23:124–142
Patton A (2012) A review of copula models for economic time series. J Multivar Anal 110:4–18
Pickands J (1981) Multivariate extreme value distributions. In: Proceedings of the 43rd session of the international statistical institute, vol 2. Buenos Aires, pp 859–878, 894–902
Rémillard B (2017) Goodness-of-fit tests for copulas of multivariate time series. Econometrics 5(1):13
Romary T, Ors F, Rivoirard J et al (2015) Unsupervised classification of multivariate geostatistical data: two algorithms. Comput Geosci 85:96–103
Saunders KR, Stephenson AG, Karoly DJ (2021) A regionalisation approach for rainfall based on extremal dependence. Extremes 24(2):215–240
Schlather M, Tawn JA (2003) A dependence measure for multivariate and spatial extreme values: properties and inference. Biometrika 90(1):139–156
Smith R (1990) Max-stable processes and spatial extremes. Unpublished work
Straus DM (2019) Clustering techniques in climate analysis. In: Oxford Research Encyclopedia of Climate Science. Oxford University Press
Zhang D, Wells MT, Peng L (2008) Nonparametric estimation of the dependence function for a multivariate extreme value distribution. J Multivar Anal 99(4):577–588

Zuccolotto P, De Luca G, Metulini R, et al (2023) Modeling and clustering of traffic flows time series in a flood prone area. In: Cerchiello P, Agosto A, Osmetti S, et al (eds) Proceedings of the statistics and data science conference. Pavia University Press, Pavia, pp 113–118