Detecting the stationarity of spatial dependence structure using spectral clustering

Véronique Maume-Deschamps, Pierre Ribereau, Manal Zeidan

Detecting the stationarity of spatial dependence structure using spectral clustering

Véronique MAUME-DESCHAMPS[1], Pierre RIBEREAU[1] and Manal ZEIDAN[1,2]

[1]Institut Camille Jordan, Université Claude Bernard Lyon 1,Lyon, France.
[2]Department of Operation Research and Intelligent techniques, University of Mosul, Mosul,Iraq.

Contributing authors: veronique.maume@univ-lyon1.fr; pierre.ribereau@univ-lyon1.fr; manal.zeidan@univ-lyon1.fr;

Abstract

Modeling extreme events require some knowledge on the spatial stationary of dependence structures in order to construct reliable statistical models. For spatial processes, assuming stationarity of the dependence structure may not be reasonable due to topology of the region under study for example. In this study, we present an adapted spectral clustering algorithm for spatial extremes by considering the extremal concurrence probability as a similarity metric of the dependence structure among the stations. This algorithm involves a heuristic method able to determine whether the dependence structure of the spatial process is stationary or not. It is furthermore able to detect the number of clusters $\boldsymbol{k}$ with high accuracy. In non stationary dependence structure case, the algorithm clusters the stations into $\boldsymbol{k}$ regional clusters with similar dependence structure. In order to validate our proposed methodology, we tested it on different simulation cases based on one or more max-stable models. The accuracy of the results encouraged us to apply it on two real data set: rainfall data in the east coast of Australia and rainfall over France.

Keywords: Max-stable processes, Non-stationary dependence structures, Extremal concurrence probability, Spectral clustering

1 Introduction

Constructing a reliable statistical model for environmental extreme events like rainfall, temperature and so on, is very important to understand the behavior of this events and thus to predict its occurrence accurately. Max-stable processes are natural models for spatial extremes since they are natural extensions of the Extreme Value Theory (EVT) to spatial domains. They are powerful statistical models for extreme events in a continuous space and thus assess the risk in area that do not contain stations. One basic assumption used in modeling is the stationarity of the dependence structure. This assumption may be incorrect and thus may lead to construct meaningless models. In particular, if the data sets are taken from a large region or from regions with complex spatial features, it is plausible that the dependence structure will appear non stationary (Richards and Wadsworth (2021)). So, it is necessary to check the stationarity of the dependence structure of the spatial process before the modeling.

In fact, dealing with non stationary spatial dependence structures is difficult in practice. There are few approaches presented for modeling non stationary dependence structures like in Huser and Genton (2016), Castro-Camilo and Huser (2020) and Richards and Wadsworth (2021). These methods are mathematically complex.

The challenge is to find simple and fast in implementation methods capable to detect changes in spatial dependence. To this aim, clustering was used recently to create regionalisations of the extreme events. Clustering is an unsupervised machine learning tool that is widely used in data analysis to discover sub-groups with similar interesting features. It has applications in computer science, statistics, biology and also in climate sciences.

In the context of spatial extremes, few studies used the clustering to partition the whole region into homogeneous sub-regions depending on similarity in dependence structure. For instance, Bernard et al (2013) presented a novel clustering algorithm for maxima. In this algorithm, The similarity measure used is the F- madogram introduced by Cooley et al (2006). Combined F- madogram with a partitioning around medoids (PAM) algorithm lead to clustering depending on the dependence strength among the maxima. This algorithm was applied to analysis the rainfall over France. After that, this algorithm was applied by Bador et al (2015) on large regions and with different variables. They analyzed the maxima of summer temperature over Europe. Saunders et al (2021) expanded the work of Bernard et al (2013): they showed that the PAM algorithm is sensitive to stations density. For that reason, they proposed to use hierarchical clustering with F-madogram. The proposed algorithm was applied on rainfall stations in Australia and the resulting clusters were compared with ones obtained by the PAM algorithm.

The main goal of our work is to investigate if a spatial process under study has a stationary dependence structure or not, and if so, the spatial process is clustered into k regional clusters, each with a stationary dependence structure. To achieve our goal, we adapt the spectral clustering for spatial extreme by

combined it with the extremal concurrence probability introduced by Dombry et al (2018). Also, we propose a heuristic method to determine the number of clusters. This combination of tools makes the proposed algorithm efficient to determine automatically the number of clusters and also clustering accurately each station to its own group. The validation of our method is done in a simulation study. We apply our method on two sets of real data. The first one is the rainfall data in the east coast of Australia. The second is the rainfall over France provided by Météo-France.
The outline of the paper is as follows. Section 2 is dedicated to a presentation of Max-stable processes. An overview of spectral clustering is exposed in Section 3. Section 4 is dedicated to describe the adapted spectral clustering for spatial extremes. We present a simulation study in Section 5. Application on real data: rainfall over east coast of Australia and rainfall over France are presented in Section 6. Finally, discussion and conclusions of our study is given by Section 7.

2 Max-stable processes

In this section, we give a brief overview of max-stable processes and provide a definition of the extremal concurrence probability which will be a central tool in our study.

2.1 Definition of Max-stable processes

Let $Z_1(s), Z_2(s) \cdots$ be a sequence of independent replication of a spatial process $\{Z(s), s \in \mathcal{S}\}, \mathcal{S} \subset \mathbb{R}^d, d \geq 1$. If there exists continuous functions $A_n(s) > 0$ and $B_n(s) \in \mathbb{R}$ such that

$$\frac{\max_{i=1,\cdots,n} Z_i(s) - B_n(s)}{A_n(s)} \overset{d}{=} X(s), s \in \mathcal{S}, n \to \infty, \tag{1}$$

is non-degenerate, then $\{X(s), s \in \mathcal{S}\}$ is a max stable process (see De Haan and Pereira (2006)). The univariate maxima $X(s)$ at location s, follows a Generalized Extreme Value distribution (GEV), i.e for all $x \in \mathbb{R}$,

$$\mathbb{P}(X(s) \leq x) = \exp[-(1 + \xi(s)\frac{x - \mu(s)}{\sigma(s)})^{-1/\xi(s)}], \tag{2}$$

where $\mu(s) \in \mathbb{R}$ is the location parameter, $\sigma(s) > 0$ is the scale parameter and $\xi(s) \in \mathbb{R}$ is the shape parameter. These parameters are possibly different from one location to another. Setting $\mu(s) = \sigma(s) = \xi(s) = 1$, lead to consider unit Fréchet distributions, i.e $\mathbb{P}(X(s) \leq x) = \exp[-1/x], x > 0$, and $\{X(s), s \in \mathcal{S}\}$ is called a simple max-stable process (see Ribatet (2017) and Ribatet et al

(2016)). De Haan (1984) provided the spectral representation for simple max-stable processes $\{X(s), s \in \mathcal{S}\}$ as follows,

$$X(s) = \max_{i \geq 1} \zeta_i Y_i(s), s \in \mathcal{S}, \mathcal{S} \subset \mathbb{R}^d, d \geq 1 \tag{3}$$

where $\{\zeta_i, i \geq 1\}$ is a Poisson point process on $(0, \infty)$, with intensity $\zeta^{-2} d\zeta$ and $Y_1(s), Y_2(s), \cdots$ denotes a sequence of independent replication of a positive process $\{Y(s), s \in \mathcal{S}\}$ with $\mathbb{E}[Y(s)] = 1$ for all $s \in \mathcal{S}$.
It may be more suitable to re-write Equation (3) as:

$$X(s) = \max_{\varphi \in \Phi} \varphi(s), s \in \mathcal{S} \tag{4}$$

where $\Phi = \{\varphi_i(s) = \zeta_i Y_i(s) : s \in \mathcal{S}, i \geq 1\}$ is a Poisson point process on $\mathbb{C}_0$, the space of non-negative continuous functions on $\mathcal{S}$ (see Ribatet (2017)).
Let $\mathbf{S}$ be a set of m spatial locations : $\mathbf{S} = \{s_1, \cdots, s_m\} \subset \mathcal{S}$, then the multivariate maxima distribution is given by

$$\mathbb{P}\{X(s_1) \leq x_1, \cdots, X(s_m) \leq x_m\} = \exp\{-V_{\mathbf{S}}(x_1, \cdots, x_m)\}, \tag{5}$$

where

$$V_{\mathbf{S}}(x_1, \cdots, x_m) = \mathbb{E}\left\{\max_{j=1,\cdots,m} \frac{Y(s_j)}{x_j}\right\}, \tag{6}$$

is called the exponent measure, which characterizes the dependence structure of $X(s_1), \cdots, X(s_m)$. Since the exponent measure is homogeneous of order -1, we can get a useful relation by setting $x_j = x$ for all $j = 1, \cdots, m$ and we get $V_{\mathbf{S}}(1, \cdots, 1) = \theta_{\mathbf{S}}$ where $\theta_{\mathbf{S}}$ is the extremal coefficient which gives us a summary of the dependence structure (see Schlather and Tawn (2003) and Smith (1990)). In particularly, when $\mathbf{S} = \{s_1, s_2\}$ the extremal coefficient satisfies $\theta_{\mathbf{S}} = V_{\mathbf{S}}(1, 1) \in [1, 2]$, the lower bound refers to variables $X(s_1)$ and $X(s_2)$ are completely dependent and the upper bound corresponds to independent.
Many models for max-stable process have been presented based on this spectral representation, such as Brown-Resnick (see Brown and Resnick (1977)), Smith (see Smith (1990)), Schlather (see Schlather (2002)) and Extremal-t (see Opitz (2013)).

2.2 Extremal concurrence probability

Other indices in order to measure the dependence between extremes exist in the literature. Dombry et al (2018) introduced the extremal concurrence probability for the analysis of extremal dependence, which is especially designed for max-stable processes. It has properties similar to the pairwise extremal coefficient but it has the advantage of being a probability measure, and this make it more interpretable and axiomatic. The idea of this metric can be explained as follows.

Recall the spectral representation in Equation (4), we say that the extremes are concurrent at locations $s_1, \cdots, s_m \in \mathcal{S}$ if

$$X(s_j) = \varphi_\ell(s_j), j = 1, \cdots, m \tag{7}$$

for some $\ell \geq 1$. This means that the values of the process $\{X(s), s \in \mathcal{S}\}$ at locations $s_1, \cdots, s_m$ come from the same spectral function φ_ℓ.
The extremal concurrence probability is defined as

$$p_r(s_1, \cdots, s_m) = \mathbb{P}\{\text{for some } \ell \geq 1 : X(s_j) = \varphi_\ell(s_j),\ j = 1, \cdots, m\} \tag{8}$$

Remark that, $p_r(s_i, s_j) = 0$ iff $X(s_i)$ and $X(s_j)$ are independent, and $p_r(s_i, s_j) = 1$ iff $X(s_i)$ and $X(s_j)$ are completely dependent.
Dombry et al (2018) present a simple estimator for this dependence measure in the bivariate case. This estimator is unbiased and coincides with Kendall's τ statistic:

$$\hat{p}_r(s_1, s_2) \equiv \hat{\tau} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}\{X_i(s_1) - X_j(s_1)\}\text{sign}\{X_i(s_2) - X_j(s_2)\}, \tag{9}$$

where $\{X_i(s), s \in \mathcal{S}, i = 1, \cdots, n\}$ are n independent copies of $\{X(s), s \in \mathcal{S}\}$.

3 Spectral clustering : an overview

Spectral clustering became one of the most popular modern technic in clustering analysis due to the simplicity of implementation, and the efficiency in addressing graph based clustering problems. This method is also flexible and does not have any assumption on the shape/size of the clusters.
It deals with the dataset as a graph. Each data point $x_i, i = 1, \cdots, n$ represents a vertex in an undirected weighted graph. Generally, an undirected graph $G = (V, E, S)$ is defined by $V = \{v_1, v_2, \cdots v_n\}$ a set of vertices, $E = \{(v_i, v_j) | v_i, v_j \in V\}$ a set of edges between these vertices and S the similarity matrix: $s_{ij} \in S$ is the amount of similarity between the vertices v_i, v_j. It represents the weight that will be assign to each edge. Since the graph is undirected, the similarity matrix should be symmetric. Note that $s_{ij} = 0$, means no edge between the vertices v_i, v_j. Each vertex v_i in the graph has degree d_i :

$$d_i = \sum_{j=1}^{n} s_{ij}. \tag{10}$$

The degrees $d_1, \cdots, d_n$ represent the elements in the diagonal matrix D called degree matrix of the graph.
The aim of clustering methods is to separate the main graph G into sub-graphs in such a way that the weights of the edges between these sub-graphs are small (this means that the clusters are dissimilar to each other), while the weights

of the edges within the sub-graphs are large (this means that there is a large similarity within the clusters).
In general spectral clustering algorithms follow three steps described below.

1. **Pre-processing**
   Evaluate the similarity matrix S from the dataset by a measure taking into account the aim of clustering, then construct the similarity graph. There are different ways to do that depending on pairwise similarity s_{ij}. The common aim is to model the neighborhood relation among the data points $x_1, \cdots, x_n$. These ways are summarized as follows:

   - ε **-neighborhood graph:** The vertices v_i, v_j will connected by an edge if they are similar enough, i.e if $s_{ij} > \varepsilon$, ε is a pre-defined non negative real number. Usually this graph is considered as an unweighted graph.
   - k**-nearest neighbor graphs:** In this graph, the distance between each pair of vertices is computed. The most popular distance metric used is the Euclidean distance. Then, the vertices v_i, v_j is connected by an edge if v_j is among the k nearest neighbors of v_i or vice versa, the edge is weighted by the similarity s_{ij}. The neighborhood relationship of data points is controlled by k, k is pre-defined integer number.
   - **The fully connected graph:** Each vertex will connected with all other vertices by edges, these edges are weighted by the similarities s_{ij}. This type of graph is useful only if the similarity function can models the neighborhood relation among the data points. The similarity function commonly used is the Gaussian similarity function $s_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$, where the neighborhood relation is controlled by σ.

   For more details about the similarity graphs see Von Luxburg (2007) and Parodi (2012).
2. **Spectral representation**
   This step uses the graph Laplacian matrix, which is an essential tool of the spectral clustering. There are two different definitions for this matrix as follows.
   (a) Unnormalized graph Laplacian matrix L: $L = D - S$.
   (b) Normalized graph Laplacian matrix L^{sym}: $L^{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$.
   The graph Laplacian matrix is used in the approximation of graph clustering problems. Let C_i be a subset of vertices i.e $C_i \subset V, i = 1, \cdots, k$ and its complement $\overline{C}_i := V \backslash C_i$, two common objective functions are considered in graph clustering problems. The first one is RatioCut (Hagen and Kahng (1992)) defined as:

$$\text{RatioCut}(C_1, \cdots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C}_i)}{|C_i|}. \tag{11}$$

Where

$$\text{cut}(C, \overline{C}) := \sum_{i \in C, j \in \overline{C}} s_{ij}$$

$$|C_i| := \text{number of vertices in } C_i$$

In this function, the size of a subset C_i is measured by its number of vertices. The approximate minimizer of RatioCut may is obtained by using spectral clustering with the unnormalized graph Laplacian matrix L. The second objective function is Normalized cut(Ncut)(Shi and Malik (2000)) which is defined as:

$$\text{Ncut}(C_1, \cdots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C}_i)}{\text{vol}(C_i)}. \tag{12}$$

where

$$\text{vol}(C) := \sum_{i \in C} d_i$$

Here the size of a subset C_i is measured by the weights of its edges. Using the normalized graph Laplacian matrix L^{sym} in spectral clustering gives approximate minimizer of Normalized cut. For more details see Von Luxburg (2007).

The matrices L and L^{sym} have some important properties: they are symmetric and positive semi-definite matrices; the n eigenvalues $\lambda_1, \cdots, \lambda_n$ of these matrices are non negative real-valued, so $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$; the multiplicity k of the value 0 as an eigenvalue of these matrices is equal to the number of connected components $C_1, \cdots, C_k$ in the graph. (for more details, see Mohar et al (1991), Mohar (1997) and Chung (1997)).

The spectrum λ (i.e. eigenvalue) of the graph Laplacian matrix and its associated eigenvectors are computed. Then the eigenvectors are used to constitute a low-dimensional representation of the data, typically k-dimensional representation is used, where k equals the number of clusters. (see Wierzchoń and Kłopotek (2018)).

3. **clustering**
   Apply a k-means clustering algorithm on the low-dimensional representation in order to assign the data points to a number of clusters.

4 Adapting spectral clustering for spatial extremes

Let $\{X(s), s \in \mathcal{S}\}, \mathcal{S} \subset \mathbb{R}^d, d \geq 1$ be a max-stable process. In order to apply spectral clustering in a spatial extreme context, the locations $s \in \mathcal{S}$ have to be considered as the vertices in a fully connected graph. These vertices are connected with each other by edges. The weights (similarity values) of such a graph will be the dependence strength among the locations.

Selecting the appropriate metric in order to construct the similarity matrix is an essential point in the clustering algorithm. In this study the spatial dependence measure used is the extremal concurrence probability introduced

by Dombry et al (2018) (see Section 2.1). After many attempts, this metric showed its ability to represent the spatial dependence for our purpose, compared with other extremal measures.
The similarity matrix represents the pairwise extremal concurrence probability matrix, we denote it by $CP \in \mathbb{R}^{m\times m}$. For a pair $(s, s') \in \mathcal{S}\times\mathcal{S}$ the element of the matrix CP is given by

$$cp(s,s') = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \text{sign}\{X_i(s) - X_j(s)\}\text{sign}\{X_i(s') - X_j(s')\} \quad (13)$$

4.1 A heuristic method to determine the number of clusters

In spectral clustering, a specific heuristic method for choosing the number of clusters k was proposed. It depends on the gap between two consecutive eigenvalues. The number of clusters is given by the value of k that maximize the eigengap δ_k: $\delta_k =\mid \lambda_{k+1} - \lambda_k \mid$, $k \geq 2$ (see Von Luxburg (2007)). Determining k to be larger than or equal to 2 leads to surely cluster the dataset to at least two groups.
This method determines the number of clusters successfully only if the dataset is well separated, and this is not always achieved. Furthermore, it is not suitable for our goal since we want to check if we can consider the dataset as one group (i.e the dependence structure can be considered as stationary), and in that case, it makes no sense to cluster the data. For that, we propose another heuristic methodology.
The idea of this heuristic methodology came from the fact that the second smallest eigenvalue λ_2 of the Laplacian matrix corresponds to the algebraic connectivity or simply to Fiedler value. It reflects how well the overall graph is connected (see Fiedler (1973)). It informs about the intensity of the connections between the nodes of the graph. Thus a low λ_2 value suggest the existence of well separated sub-graphs (clusters) and vice versa (see Wierzchoń and Kłopotek (2018)). So, when the graph is well connected, λ_2 will be far from the first eigenvalue λ_1 and if we check the outlier values in the set of first ten eigenvalue set we will see that λ_1 is the only outlier value. This case indicates that the graph is well connected and the data has a stationary dependence structure. On the contrary, the graph can be clustered into sub-graphs and since we are interested in the smallest eigenvalues, a relative eigengap RE_k is used: $RE_k = \frac{\lambda_{k+1}-\lambda_k}{\lambda_k}$, $k \geq 2$.
If we do not have additional informations, we will use a highest value of RE_k to determine the number of clusters. A simulation study in Section 5 validates this method based on the highest value of RE_k. The steps of this heuristic methodology is described in Algorithm 1

Algorithm 1 Heuristic method to determine the number of clusters k

Require: Vector of eigenvalues λ.
Ensure: Number of clusters k.

1: Find the outliers value in the eigenvalues set $(\lambda_1, \cdots, \lambda_{10})$.
2: If λ_1 is the only outlier value, then $k = 1$. Else, go to step 3.
3: Calculate the relative eigengaps $RE_k = \frac{\lambda_{k+1} - \lambda_k}{\lambda_k}$,$k \geq 2$.
4: k is correspond to the highest value of RE_k.

4.2 Description of the proposed spectral clustering algorithm

Let $X_i(s_j), s_j \in \mathcal{S}$, $\mathcal{S} \subset \mathbb{R}^d$, $d = 2$, $i = 1, \cdots, n$ be a sequence of n independent and identically distributed max stable processes at different locations $s_j, j = 1, 2, \cdots, m$. Firstly, we have to construct the similarity matrix CP according to Equation (13) then use it to compute the graph Laplacian matrix. We use the normalized graph Laplacian matrix L^{sym} because we want the size of the resulting clusters to depend on the strength of the dependence structure (i.e the weights of the graph edges). The spectrum λ of L^{sym} is then computed and will be the input in Algorithm 1 in order to determine the number of clusters k. If $k \geq 2$, the eigenvectors $q_1, \cdots, q_k$ are used to constitute k-dimensional representation for the data. This is done by representing these eigenvectors as columns of an $m \times k$ matrix denoted Q. Each row in Q represents a location s_j: $s_j \rightarrow (q_{j,1}, \cdots, q_{j,k})$ $j = 1, \cdots, m$, this is called spectral mapping (see Wierzchoń and Kłopotek (2018)). Normalizing each row of Q to norm 1, results in the matrix denoted $Y \in \mathbb{R}^{m \times k}$. According to Ng et al (2001) this last step improves the performance of clustering algorithm. We used Gaussian Mixture Model (GMM) to cluster the rows of Y instead of k-means that is usually used at this step. GMM clusters the datapoints based on probability distribution. It considers that the datapoints come from a Gaussian mixture. Each cluster has a Gaussian distribution model with parameters mean and covariance. Taking the covariance into account make GMM more robust than k-means which depends only on the cluster mean. For more details about GMM see for example Bouveyron et al (2019). We sum up these steps in Algorithm 2.

5 Simulation study

In order to assess the accuracy of our algorithm, we test it on three simulation cases. In each cases, we implemented the algorithm on several spatial processes simulated from one or more max-stable models: Smith, Schather, Brown-Resnick and Extremal-t with parameters chosen randomly. Furthermore, the correlation functions for Schlather and Extremal-t models are chosen randomly from one of the following correlation functions: Cauchy, powered exponential and Whittle-Matérn. The number of observations is fixed to 1000 at each location $s_j, j = 1, 2, \cdots, m$. To avoid any effect of the number of locations m on the algorithm's work, m is chosen randomly at each simulation. We

Algorithm 2 Proposed spectral clustering

Require: The similarity matrix $CP \in \mathbb{R}^{m\times m}$, constructed according to Equation (13) .

Ensure: Clusters $\{C_1, \cdots, C_k\}$.

1: Compute the normalized Laplacian matrix $L^{sym} = D^{-\frac{1}{2}}(D - CP)D^{-\frac{1}{2}}$.
2: Compute the spectrum of L^{sym} and use Algorithm 1 in order to determine k. If $k = 1$, the algorithm stops. Else, go to step 3.
3: Compute the k smallest eigenvectors $q_1, q_2, \cdots, q_k$ of L^{sym}, and arrange these vectors in columns to be a matrix Q, where $Q \in \mathbb{R}^{m\times k}$.
4: Normalize the rows of Q to norm 1, resulting the matrix $Y \in \mathbb{R}^{m\times k}$: $Y_{jl} = Q_{jl}/(\sum_l Q_{jl}^2)^{\frac{1}{2}}$, $j = 1, \cdots, m$, $l = 1, \cdots, k$.
5: Consider each row of Y as a point in $\mathbb{R}^k$ and implement Gaussian Mixture Model (GMM) to cluster them into k clusters.
6: Assign the location s_j to cluster l if and only if row j of the matrix Y is assigned to cluster l.

simulate it uniformly between 30 and 100 locations. The locations are generated randomly and uniformly in $[0, 1]^2$. The three simulation cases are detailed as follows.

- Case 1: stationary dependence structure
  In this case, we simulate 100 spatial processes from one of the max-stable models chosen randomly.
- Case 2: non-stationary dependence structure / different models
  In this case, we considered three sub-cases where the number of clusters equals 2, 3 or 4. For each of these sub-cases, 100 spatial processes are simulated from two (three and four respectively) different max-stable models chosen randomly.
- Case 3: non-stationary dependence structure / one model with different parameters
  In this case, we considered three sub-cases where the number of clusters equals 2, 3 or 4. For each of these sub-cases, 100 spatial processes with two (three and four respectively) clusters are simulated from the same max-stable model chosen randomly. The parameters of each cluster are chosen randomly and are different from the parameters of other clusters.

We assess the accuracy of our proposed algorithm using two evaluation measures. The first one is the accuracy of our proposed heuristic method in order to determine the number of clusters. For simplicity we will denote it by $\mathcal{A}_k$:

$$\mathcal{A}_k = \frac{\#\hat{k}_t}{T} \times 100 \tag{14}$$

where T is the total number of spatial processes tested and $\hat{k}_t, t \subset T$ is the correctly estimation the number of clusters.
The second one is the accuracy of our proposed spectral clustering algorithm in

clustering the locations correctly. That means, its ability to group the locations simulated from the same model in the same cluster. At the same time we justify the reason for choosing GMM in step 5 of the algorithm instead of k-means used in traditional spectral clustering. We will compare the accuracy of our proposed spectral clustering algorithm with GMM and with k-means in step 5. For simplicity we will denote this evaluation measure by $\mathcal{A}_c$:

$$\mathcal{A}_c = \frac{\#\hat{c}_h}{H} \times 100 \tag{15}$$

where H is the total number of spatial processes with number of clusters determined correctly and $\hat{c}_h, h \subset H$ is the correctly clustering the locations. The two evaluation measures, $\mathcal{A}_k$ and $\mathcal{A}_c$ are computed for all simulation cases considered. The results are presented in Table 1.

Regarding the performance of our proposed heuristic method (first row of Table 1) it appears accurate in detecting whether the spatial processes has a stationary dependence structure or not. For non-stationary dependence structures, it is also able to determine the correct number of clusters $\hat{k}$ with high accuracy between 93% and 100%.

Table 1 The evaluation measures for the proposed spectral clustering within all simulation cases

	Stationary [1]	Non-stationary [2]			Non-stationary [3]		
Evaluation measures	-	$k=2$	$k=3$	$k=4$	$k=2$	$k=3$	$k=4$
$\mathcal{A}_k$	100%	100%	95%	93%	100%	93%	99%
$\mathcal{A}_c$(GMM)	-	100%	100%	100%	100%	100%	100%
$\mathcal{A}_c$(k-means)	-	100%	83.16%	75.26%	100%	91.39%	82.83%

[1]Simulation case 1

[2]Simulation case 2

[3]Simulation case 3

The second row of Table1 shows that our proposed spectral clustering algorithm is accurate at 100% in clustering the locations according to the model from which they were simulated. This accuracy is the same for the two simulation cases of non-stationary dependence structure and for all number of clusters that tested. This reinforces the reason for using GMM instead of k-means in step 5 of the algorithm. Indeed, k-means is accurate 100% only when the number of clusters equals 2 but its accuracy decreases as the number of clusters increases as we note in row three in Table 1.

Since the results on simulation appear satisfactory, we can use this technics in order to detect stationary extreme area for precipitation in East Australia and in France.

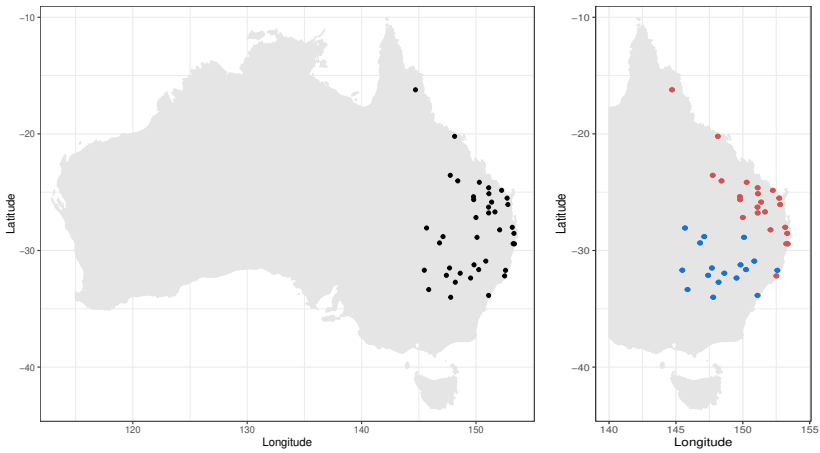

Fig. 1 Geographic locations of 40 stations on the east coast of Australia in the left panel, and proposed spectral clustering in the right panel.

6 Application on real data

This section is devoted to two real data application: one on rainfall in Australia east coast; the other one on rainfall in France.

6.1 Rainfall over east coast of Australia

We begin by a brief description of the data, then we apply our clustering method and comment the results.

6.1.1 Description of the data

This data represents daily rainfalls measured (in millimeters) at 40 stations on the east coast of Australia, recorded during the winter season (April − September) from 1972 to 2019. This leads to 183 * 48 = 8784 observations at each stations. The geographic locations of the 40 stations are illustrated in the left panel of Figure 1. The altitude of these stations was chosen from 2 to 540 m. For more details about this data see for example Ahmed et al (2022), Bacro et al (2016), Ahmed et al (2017) and Abu-Awwad et al (2020).These data are freely available on the website http://www.bom.gov.au.

6.1.2 Detecting the stationarity of Australia rainfall dependence structure

We shall apply our proposed spectral clustering algorithm described in Section 3 Algorithm 2 to detect the stationarity of rainfall dependence structure. In order to show the effect of the block maxima size on the detection of stationarity in the dependence structure we take different sizes of block maxima. We take them as in Ahmed et al (2022) to discuss the results. The block maxima sizes taken are 183 days, 30 days, 15 days, 10 days, 5 days, 3 days and 1 day.

Table 2 shows results in detecting the stationarity of the dependence structure for Australia rainfall data for each size of block maxima. We observe that the block maxima size affects the detection of stationarity in the dependence structure. The rainfall dependence structure is detected as stationary when the block maxima size is greater than or equal to 15 days. We can use only one model for the whole spatial data when modeling these data with these block sizes.

Ahmed et al (2022) used Convolutional Neural Networks (CNN) to classify

Table 2 Detecting the stationarity of the dependence structure for Australia rainfall data with different sizes of block maxima.

Block size	Spectral clustering implementation outputs		
	Stationary dependence structure	Non-stationary dependence structure	No. of clusters
183 days	✓	-	-
30 days	✓	-	-
15 days	✓	-	-
10 days	-	✓	$k = 2$
5 days	-	✓	$k = 2$
3 days	-	✓	$k = 2$
1 day	-	✓	$k = 2$

the dependence structure between Asymptotic Dependence (AD) and Asymptotic Independence (AI). They trained their network on AD (resp. AI) by simulating the data from one of the max-stable models (or inverse max-stable models). They also considered max-mixture models constructed with AD and AI models chosen randomly. Their CNN classified the dependence structure of this data with these block maxima sizes as asymptotic independence. Since the CNN training for asymptotic independence was based on one model, this means the data has one type of dependence structure and this confirms the result we obtained which is the data with block maxima size greater than or equal to 15 days has stationary dependence structure.

The rainfall dependence structure is non-stationary when the block maxima size is less than or equal to 10 days. Our proposed heuristic method determines that the number of cluster is equal to 2 for all these block sizes as illustrated in Figure 2. The spectral clustering algorithm clustered the 40 locations to two clusters each with stationary dependence structure. These clusters are the same for each of these block sizes and illustrated in the right panel of Figure 1. So, it is suitable to use two models when modeling this data with these block maxima sizes.

Predicting the class of the dependence structure for block size 5 days is not conclusive in Ahmed et al (2022), while for 3 days and daily block size, their CNN classified the dependence structure as mixture between AD and AI. Depending on CNN training, this means that the data include a mixture of

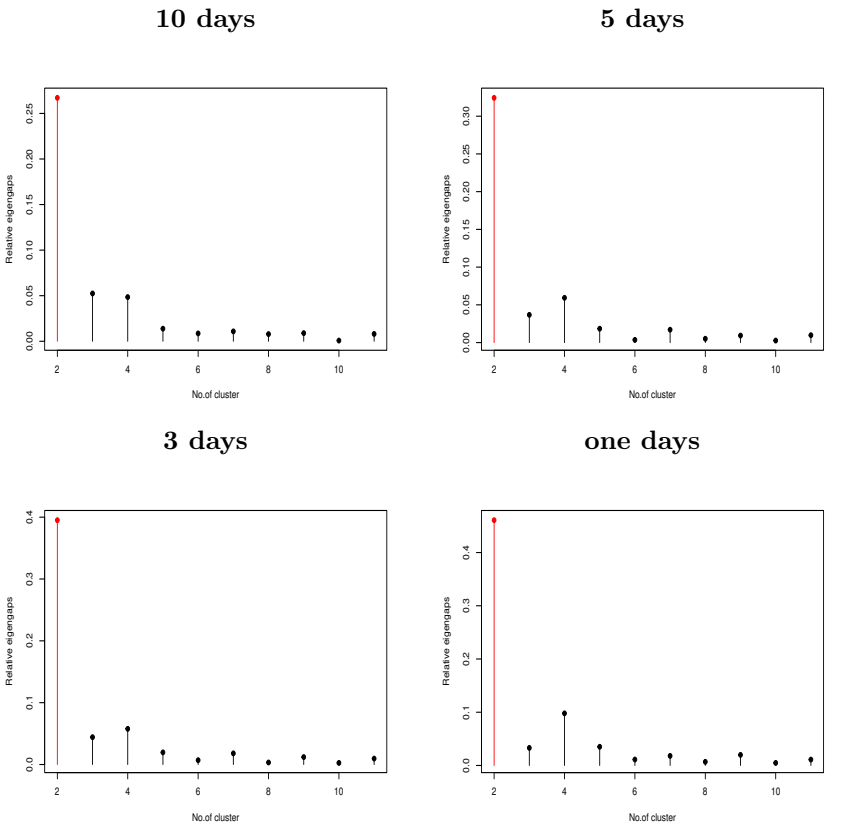

Fig. 2 The value of the relative eigengap associated with each number of clusters (red line refers to the number of clusters) for Australia data with four different block maxima sizes. The left panel in the first row for 10 days block maxima and the right one for 5 days block maxima. While the left panel in the second row for 3 days block maxima and the right one for one day block maxima.

two type of dependence structure. For that our results for 3 days and daily block size are in accordance with in Ahmed et al (2022). Furthermore, the daily rainfall has been studied by Bacro et al (2016), Ahmed et al (2017) and Abu-Awwad et al (2020). They showed that max-mixture models are suitable for modeling this data.

6.2 Rainfall over France

This subsection is devoted to the study of rainfall data in France.

6.2.1 Description of the data

This data are provided by Météo-France and represent the hourly precipitation recorded at 80 French monitoring stations. The data were measured during the period 1993 - 2021 over fall season (September,October and November). At

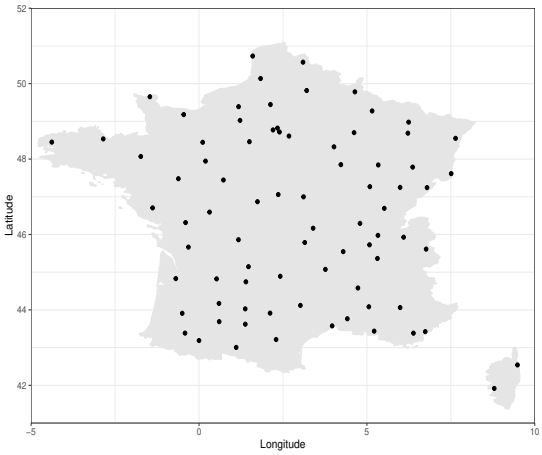

Fig. 3 Geographic locations of 80 stations over France region

each stations there are 91 * 29 = 2639 observations. The geographic locations of these stations were chosen in order to cover all the French metropolitan regions. Figure 3 illustrates the geographic locations of the 80 stations. This data was studied by Bernard et al (2013) on the period 1993 - 2011.

6.2.2 Detecting the stationarity of France rainfall dependence structure

To detect the stationarity of rainfall dependence structure, we implement our proposed spectral clustering on the data. Also we study the effect of the block maxima size on the stationary of the dependence structure. The block sizes we consider are : annual, monthly, 2 weeks, weekly and 5 days. The results obtained with our spectral clustering on this data for each block maxima size are showed in Table 3.

It is clear that for both annual and monthly block sizes, the rainfall has station-

Table 3 Detecting the stationarity of the dependence structure for France rainfall data during the period (1993 - 2021) with different sizes of block maxima.

	Spectral clustering implementation outputs		
Block size	Stationary dependence structure	Non-stationary dependence structure	No. of clusters
Annual	✓	-	-
Monthly	✓	-	-
2 Weeks	-	✓	$k = 2$
Weekly	-	✓	$k = 2$
5 days	-	✓	$k = 2$

ary dependence structure, while non-stationarity in the dependence structure

appears when the block size is less than or equal to 2 week.
The number of clusters determined by the proposed heuristic method and the clustering of stations obtained by spectral clustering for each block size leading to non stationary dependence structure are illustrated in Figure 4, where the first (second and third) row is related to block maxima size equals to 2 weeks (weekly and 5 days respectively).
We will start our discussion with weekly block maxima since it was studied by Bernard et al (2013). Implemention of spectral clustering on the data with this block size shows that the rainfall has a non stationary dependence structure. Our spectral clustering divides France into north/south regional areas along Loire valley line, each of these regions has a different dependence structure as explained by left panel in the second row of Figure 4. This can be interpreted easily. The extreme rainfall in north of France is produced by disturbances from the Atlantic, this affects Brittany, Paris and also other areas in the north of France. While in the south of France, the extreme rainfall are caused by the Mediterranean sea, it affects the coastal area and especially on Cévennes and the Montagne Noire.
The results are similar to those obtained by Bernard et al (2013), where the selection criterion for the number of clusters in Bernard et al (2013) indicated that $k = 2$. The locations clustering result are illustrated in the right panel in the second row of Figure 4. It is relatively close to Bernard et al (2013), where France were divided into north and south regions. This indicates that the behavior of the data has not changed since 2011.
The right panel in the first row of Figure 4 shows the stations clustering result when the size of block maxima equals to 2 weeks. It is clear that the south region is smaller than ones in weekly block maxima. This change in the size of clusters is due to the fact that the size of block maxima had become larger compared with the weekly block maxima. This indicates that the larger size of block the nearest to stationary in dependence structures.
In case of 5 days block maxima, the stations clusters is quite similar to the weekly block maxima case. This can be seen on the right panel in the third row of Figure 4.

7 Discussion and Conclusion

Stationarity of dependence structure is an essential matter in modeling environmental extreme events. In most studies, it is assumed that the dependence structure is stationary. This may be incorrect especially in large regions and regions with complex geographical or climatic patterns. Finding a method able to detect regions with similar dependence structure is thus useful. In this study, we proposed to take advantage of spectral clustering in clustering regions which have similar dependence structure together. So, we combined spectral clustering with extremal concurrence probability to create a simple clustering method for max-stable processes. Also, we proposed a heuristic method able to determine if the dependence structure of the data are stationary or not, and

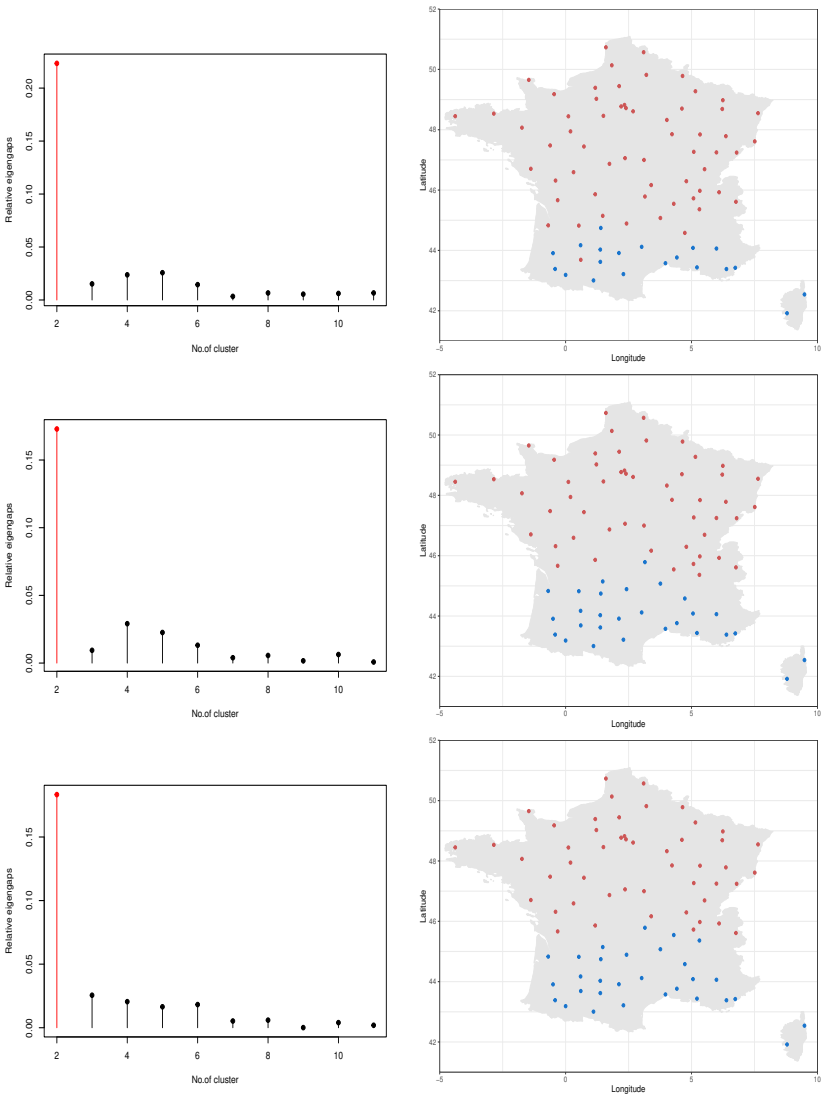

Fig. 4 The value of the relative eigengap associated with each number of clusters (red line refers to the number of clusters) in the left panel, the result of proposed spectral clustering for the France rainfall data along the period (1993 - 2021) with each block maxima size in the right panel. The first row relates to 2 week block maxima. The second row relates to weekly block maxima. The third row relates to 5 days block maxima.

to determine the number of clusters with high accuracy. The validation of the proposed spectral clustering algorithm was done via a simulation study. Then two environmental dataset has been studied, the first one is the daily rainfall data over east coast of Australia. We found that this data has a stationary dependence structure when the size of block maxima is larger than or equals to 15 days. The data has the same two regional clusters when the block size

is smaller than or equals to 10 days. The second dataset is the hourly precipitation over France. we found that the stationarity appeared for large block maxima sizes (monthly and annual). While non-stationary dependence structures is plausible for block size less than or equal to 2 weeks. The regional clusters is not the same for all these block sizes. From the foregoing, we conclude that the size of block maxima affects the stationarity of the dependence structures and sometimes make changes in the regional clusters too. So, this must be taken into account in the modelization. Finally, despite the simplicity of our proposed algorithm, it is powerful. As a future direction of this study, one can study other variables like temperature. Another direction is to state the efficiency of this algorithm when applied it to very large region like the whole of Australia or the continent of Europe.

Acknowledgments We acknowledge the partially support by PAUSE program, which is operated by the Collège de France. Also, we would like to thank French meteorological service (Météo-France) for providing us the data.

References

Abu-Awwad A, Maume-Deschamps V, Ribereau P (2020) Fitting spatial max-mixture processes with unknown extremal dependence class: an exploratory analysis tool. Test 29(2):479–522. https://doi.org/10.1007/s11749-019-00663-5

Ahmed M, Maume-Deschamps V, Ribereau P, et al (2017) A semi-parametric estimation for max-mixture spatial processes. arXiv preprint arXiv:171008120

Ahmed M, Maume-Deschamps V, Ribereau P (2022) Recognizing a spatial extreme dependence structure: A deep learning approach. Environmetrics 33(4):e2714. https://doi.org/10.1002/env.2714

Bacro JN, Gaetan C, Toulemonde G (2016) A flexible dependence model for spatial extremes. Journal of Statistical Planning and Inference 172:36–52. https://doi.org/10.1016/j.jspi.2015.12.002

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 and climate extremes 9:17–24. https://doi.org/10.1016/j.wace.2015.05.003

Bernard E, Naveau P, Vrac M, et al (2013) Clustering of maxima: Spatial dependencies among heavy rainfall in france. Journal of climate 26(20):7929–7937. https://doi.org/10.1175/JCLI-D-12-00836.1

Bouveyron C, Celeux G, Murphy TB, et al (2019) Model-based clustering and classification for data science: with applications in R, vol 50. Cambridge

University Press

Brown BM, Resnick SI (1977) Extreme values of independent stochastic processes. Journal of Applied Probability 14(4):732–739. https://doi.org/10.2307/3213346

Castro-Camilo D, Huser R (2020) Local likelihood estimation of complex tail dependence structures, applied to us precipitation extremes. Journal of the American Statistical Association 115(531):1037–1054. https://doi.org/10.1080/01621459.2019.1647842

Chung FR (1997) Spectral graph theory, vol 92. American Mathematical Soc.

Cooley D, Naveau P, Poncet P (2006) Variograms for spatial max-stable random fields, Springer New York, New York, NY, pp 373–390. https://doi.org/10.1007/0-387-36062-X_17

De Haan L (1984) A spectral representation for max-stable processes. The annals of probability 12(4):1194–1204

De Haan L, Pereira TT (2006) Spatial extremes: Models for the stationary case. The annals of statistics 34(1):146–168. https://doi.org/10.1214/009053605000000886

Dombry C, Ribatet M, Stoev S (2018) Probabilities of concurrent extremes. Journal of the American Statistical Association 113(524):1565–1582. https://doi.org/10.1080/01621459.2017.1356318

Fiedler M (1973) Algebraic connectivity of graphs. Czechoslovak mathematical journal 23(2):298–305

Hagen L, Kahng AB (1992) New spectral methods for ratio cut partitioning and clustering. IEEE transactions on computer-aided design of integrated circuits and systems 11(9):1074–1085. https://doi.org/10.1109/43.159993

Huser R, Genton MG (2016) Non-stationary dependence structures for spatial extremes. Journal of agricultural, biological, and environmental statistics 21(3):470–491. https://doi.org/10.1007/s13253-016-0247-4

Mohar B (1997) Some applications of Laplace eigenvalues of graphs, Springer Netherlands, Dordrecht, pp 225–275. https://doi.org/10.1007/978-94-015-8937-6_6

Mohar B, Alavi Y, Chartrand G, et al (1991) The laplacian spectrum of graphs. Graph theory, combinatorics, and applications 2(871-898):12

Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems 14

Opitz T (2013) Extremal t processes: Elliptical domain of attraction and a spectral representation. Journal of Multivariate Analysis 122:409–413. https://doi.org/10.1016/j.jmva.2013.08.008

Parodi P (2012) Computational intelligence with applications to general insurance: a review: I–the role of statistical learning. Annals of Actuarial Science 6(2):307–343. https://doi.org/10.1017/S1748499512000036

Ribatet M (2017) Modelling spatial extremes using max-stable processes

Ribatet M, Dombry C, Oesting M (2016) Spatial extremes and max-stable processes. Extreme Value Modeling and Risk Analysis: Methods and Applications pp 179–194

Richards J, Wadsworth JL (2021) Spatial deformation for nonstationary extremal dependence. Environmetrics 32(5):e2671. https://doi.org/10.1002/env.2671

Saunders K, Stephenson A, Karoly D (2021) A regionalisation approach for rainfall based on extremal dependence. Extremes 24(2):215–240. https://doi.org/10.1007/s10687-020-00395-y

Schlather M (2002) Models for stationary max-stable random fields. Extremes 5(1):33–44. https://doi.org/10.1023/A:1020977924878

Schlather M, Tawn JA (2003) A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika 90(1):139–156. https://doi.org/10.1093/biomet/90.1.139

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence 22(8):888–905. https://doi.org/10.1109/34.868688

Smith RL (1990) Max-stable processes and spatial extremes. Unpublished manuscript 205:1–32

Von Luxburg U (2007) A tutorial on spectral clustering. Statistics and computing 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z

Wierzchoń ST, Kłopotek MA (2018) Modern algorithms of cluster analysis, vol 34. Springer