A Hotelling spatial scan statistic for functional data: application to economic and climate data

Zaineb Smida, Thibault Laurent, Lionel Cucala

WORKING PAPERS

N° 1583

October 2024

"A Hotelling spatial scan statistic for functional data: application to economic and climate data"

Zaineb SMIDA, Thibault Laurent
& Lionel CUCALA

Toulouse School of Economics

A Hotelling spatial scan statistic for functional data: application to economic and climate data

Zaineb SMIDA*

Univ Lyon, INSA Lyon, UJM, UCBL, ECL, CNRS UMR 5208, ICJ, F-69621, France

Thibault LAURENT

Toulouse School of Economics, CNRS, Toulouse, France

Lionel CUCALA

Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France

Abstract

A scan method for functional data indexed in space has been developed. The scan statistic is derived from the Hotelling test statistic for functional data, extending the univariate and multivariate Gaussian spatial scan statistics. This method consistently outperforms existing techniques in detecting and locating spatial clusters, as demonstrated through simulations. It has been applied to two types of real data: economic data in order to identify spatial clusters of abnormal unemployment rates in Spain and climatic data in order to detect unusual climate change patterns in Great Britain, Nigeria, Pakistan, and Venezuela.

Keywords: Cluster detection, Functional data, Hotelling T^2 test, Spatial Scan statistic.
JEL Classification: C12, C21, E24, Q54

1. Introduction

Spatial cluster detection has become a prominent area of statistical research, with notable advancements in recent decades. It aims to identify concentrations of events within specific areas; for a comprehensive overview, see Lawson and Denison (2002).

One of the most widely used techniques for cluster detection is the scan statistic, first introduced by Naus (1963). This statistic was originally defined as the maximum number of events within a fixed-size window, known as the scanning window, as it continuously moves across the study area. Understanding the distributions of these scan statistics (Alm, 1997) is key to decide whether the occurrence of a cluster of events is statistically significant or not.

The field of spatial scan statistics was significantly advanced by the work of Kulldorff (1997): he introduced the use of circular windows with varying sizes to scan the study area and identified the most probable cluster using a likelihood ratio test. He used either the Bernoulli model

*Corresponding author

Email addresses: zaineb.smida@insa-lyon.fr (Zaineb SMIDA), thibault.laurent@tse-eu.fr (Thibault LAURENT), lionel.cucala@umontpellier.fr (Lionel CUCALA)

(Kulldorff and Nagarwalla, 1995) or the Poisson model (Kulldorff, 1997) and assessed the statistical significance of the clusters using a Monte Carlo procedure. These advancements led to numerous studies where researchers adapted spatial scan statistics for various data types, employing different distributions, such as exponential (Huang et al., 2007), normal (Kulldorff et al., 2009), Weibull (Bhatt and Tiwari, 2014) and Poisson with overdispersion (Lima et al., 2014).

Sometimes, numerous continuous variables have to be analysed in the same time. A multivariate scan statistic combining different univariate scan statistics has been proposed by Kulldorff et al. (2007). However, this scan statistic does not take into account the correlation structure of the observed variables. This problem was recently tackled by Cucala et al. (2017) who developed a spatial scan statistic based on a likelihood ratio and a multivariate normal probability.

Although previous scan statistics are parametric due to their reliance on likelihood ratios, alternative nonparametric methods have also been introduced such as Cucala (2014), Cucala (2016) in the univariate case and Cucala et al. (2019) in the multivariate case which are constructed using two-sample test statistics such as the Wilcoxon-Mann-Whitney ones (Wilcoxon, 1945; Mann and Whitney, 1947).

Advances in sensing and data storage have enabled continuous measurements over time, leading to the development of functional data analysis (FDA) by Ramsay and Silverman (2005). Much effort has been devoted to adapting classical statistical methods to functional data. Horváth and Kokoszka (2012) reviewed recent developments in functional data inference. Various techniques, such as principal component analysis (Boente and Fraiman, 2000), hypothesis testing (Cuevas et al., 2004; Horváth et al., 2013; Chakraborty and Chaudhuri, 2015; Joseph et al., 2015; Smida et al., 2022*a*), and regression (Ferraty and Vieu, 2002), have been developed. Recent work by Aneiros et al. (2019) and Aneiros et al. (2022) offers comprehensive reviews on these topics. Dabo-Niang and Frévent (2024) highlighted the rise of FDA across various fields, including medicine, biology, chemistry, economics, and environmental science.

In spatial statistics, using a univariate approach with time-averaged data results in significant information loss, while a multivariate approach faces issues with high dimensionality and correlation, leading to reduced power. Smida et al. (2022*b*) highlight these challenges through simulation studies. To address this, they first developed a nonparametric scan statistic for functional data, based on the Wilcoxon-Mann-Whitney test (Chakraborty and Chaudhuri, 2015). Frévent et al. (2021) later introduced two additional spatial scan statistics: one using functional ANOVA (Cuevas et al., 2004) and another combining the distribution-free spatial scan statistic (Cucala, 2014) with the max statistic (Zhenhua Lin and Müller, 2021).

In this study, we introduce a scan statistic for functional data indexed in space. Since there is no associated likelihood for functional random variables (Ferraty et al., 2011), applying a traditional likelihood ratio test is not possible. However, in the Gaussian context, maximizing the likelihood ratio is equivalent to the Hotelling test statistic (Hotelling, 1931). Therefore, our proposed scan statistic is based on the Hotelling test for functional data (Joseph et al., 2015; Horváth et al., 2013). This approach extends the Gaussian spatial scan statistics previously developed for univariate and multivariate Gaussian data (Kulldorff et al., 2009; Cucala et al., 2017), providing a robust tool for identifying clusters in functional data.

The rest of this paper is organized as follows. In Section 2, we outline the construction of the proposed spatial scan statistic for functional data based on the Hotelling test statistic, present

its computational method, and evaluate its statistical significance using random permutations. In Section 3, we conduct a simulation study to assess the performance of the new approach, comparing it with the methods of Smida et al. (2022*b*) and Frévent et al. (2021). In Section 4, we apply our method to real economic and climate datasets. Finally, Section 5 provides a discussion and outlines potential directions for future research.

2. A Hotelling spatial scan statistic for functional data

2.1. Introducing the statistic

Consider a random variable X taking values in the infinite dimensional space $\chi = L^2(T, \mathbb{R})$ where $T = [a, b]$ is a closed interval. Let $X_1, \ldots, X_n$ be observations of X at n different spatial locations $s_1, \ldots, s_n$ included in $D \subset \mathbb{R}^2$. Following the terminology of point process theory, D is the observation domain and X_i is the mark associated with location s_i, for all $i = 1, \ldots, n$. Our goal is to detect a cluster of unusual marks, i.e. a spatial zone $Z \subset D$ in which the functional marks exhibit a different behaviour than elsewhere. In order to do that, we aim to set up a scan statistic, which is usually defined as the maximum of a concentration index observed in a collection of variable size potential clusters (Nagarwalla, 1996). Concerning the potential clusters, two main possibilities have been proposed in the literature. In the first one, the windows have known geometric shapes: rectangular (Loader, 1991; Chen and Glaz, 2009), circular (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997), elliptic (Kulldorff, 2006), or any other shape. In the second one, the windows have irregular shapes and the procedure to identify them is based only on pairwise distances (Dematteï et al., 2007; Assunção et al., 2006; Duczmal and Assunção, 2004). In this work, without loss of generality, we consider the circular clusters introduced by Kulldorff (1997). Hence, the set of potential clusters $\mathcal{S}$ is defined as follows:

$$\mathcal{S} = \{D_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n\},$$

where $D_{i,j}$ is the disc centred on s_i and passing through s_j. Notably, since i can equal j, the number of potential clusters is n^2. However, it is easy to notice that some disks may contain the same set of points. For instance, if s_1 and s_2 are each other's nearest neighbors, then the disks $D_{1,2}$ and $D_{2,1}$ will contain the same set of points. To optimize calculations, it may be useful to reduce the size of $\mathcal{S}$ by removing duplicates. We denote this reduced set as $\tilde{\mathcal{S}}$, with size $\tilde{N}$. For instance, in the simulation data presented in section 3, the initial set $\mathcal{S}$ has a size of $n^2 = 94^2 = 8836$, while the refined set $\tilde{\mathcal{S}}$, after eliminating duplicates, contains $\tilde{N} = 7044$ clusters.

Following the seminal work of Kulldorff (1997), spatial scan statistics for univariate ($X \in \mathbb{R}$) or multivariate ($X \in \mathbb{R}^d, d \geq 2$) marks are generally constructed using a concentration index derived from a likelihood ratio. This likelihood ratio is based on assuming a particular probability distribution for the marks and testing the null hypothesis H_0 (no cluster) against the alternative hypothesis $H_{1,Z}$ (presence of a cluster in Z) for each potential cluster $Z \in \mathcal{S}$.

As stated in section 1, Kulldorff et al. (2009) proposed a Gaussian-based scan statistic for detecting clusters in univariate continuous data. This approach relies on the likelihood ratio between two hypotheses: the null hypothesis assumes that the marks are normally distributed and independent with equal means and variances, while the alternative hypothesis considers

equal variances but different means inside and outside the potential cluster. Notably, the concentration index in this case corresponds to the two-sample Student's T^2 test statistic, as shown in the following Lemma 1.

Lemma 1. *For a univariate variable ($X \in \mathbb{R}$), maximizing the concentration index $I_G(Z)$ for any potential cluster Z using the likelihood Gaussian ratio $LR_G(Z)$ from Kulldorff et al. (2009) is equivalent to maximizing the Student's T^2 test (Student, 1908). This equivalence is captured by the following relationship:*

$$(LR_G(Z))^{\frac{2}{n}} = 1 + \frac{T^2}{n-2}.$$

The proof is provided in online supplementary material.

An extension of the Gaussian-based scan statistic from Kulldorff et al. (2009) to the multivariate case was proposed by Cucala et al. (2017). In this approach, all multivariate marks are assumed to be normally distributed and independent. The null hypothesis considers equal mean vectors and covariance matrices for all marks, while the alternative hypothesis assumes equal covariance matrices but different means inside and outside the cluster. In this multivariate case, it is notable that the concentration index is equivalent to the two-sample Hotelling T_H^2 test statistic of Hotelling (1931) as established by the following Lemma 2.

Lemma 2. *In the multivariate case ($X \in \mathbb{R}^d$, with $d \geq 2$), maximizing the concentration index $I_{MG}(Z)$ for any potential cluster Z using the multivariate Likelihood Gaussian ratio $LR_{MG}(Z)$ from Cucala et al. (2017) is equivalent to maximizing Hotelling's T_H^2 test (Hotelling, 1931). This equivalence is described by the following relationship:*

$$(LR_{MG}(Z))^{\frac{2}{n}} = 1 + \frac{T_H^2}{n-2}.$$

The proof can be found in online supplementary material.

In the context of functional random variables, the concept of probability density generally does not exist, despite various suggested approximations (Jacques and Preda, 2013; Liu and Houwing-Duistermaat, 2024). As a result, researchers have chosen clustering indices based on nonparametric tests for equality of distributions, with notable contributions from Jung and Cho (2015); Cucala (2016), and Cucala et al. (2019) for univariate and multivariate data, respectively, and from Smida et al. (2022*b*) and Frévent et al. (2021) for functional data.

Similarly, as there is an equivalence between the Gaussian likelihood ratio and the T^2 test statistic in the univariate case, and T_H^2 in the multivariate framework (as demonstrated in the online supplementary material), our approach adopts the functional extension of the Hotelling T_{FH}^2 test statistic, as proposed by Joseph et al. (2015) (or equivalently Horváth et al., 2013), to construct a concentration index.

Hereinafter, we suppose that $X_1, \ldots, X_n$ are independent observations of the functional random variable X (this is a classical assumption in scan statistics). Let $Z \in \mathcal{S}$ be any potential cluster of size n_Z, where $n_Z = \sum_{i=1}^{n} \mathbb{1}(s_i \in Z)$ and Z^c its complement of size $n_{Z^c} = n - n_Z$.

Thus, in the context of cluster detection, the null hypothesis H_0 (absence of a cluster) can be defined as follows: $H_0 : \forall Z \in \mathcal{S}, \mu_Z = \mu_{Z^c} = \mu_D$ (the absence of cluster), where μ_Z, μ_{Z^c}, and μ_D represent the mean functions within Z (i.e. $\mu_Z = \mathbb{E}(X_i(t)), i : s_i \in Z, t \in T$), outside Z

((i.e. $\mu_{Z^c} = \mathbb{E}(X_i(t)), i : s_i \in Z^c, t \in T$), and across the entire domain D, respectively. Thus, the aim is to test:

$$H_0 : \mu_Z = \mu_{Z^c} = \mu_D \qquad vs. \qquad H_{1,Z} : \mu_Z \neq \mu_{Z^c}.$$

Let $\hat{\mu}_Z = \frac{1}{n_Z} \sum_{i:s_i \in Z} X_i$ and $\hat{\mu}_{Z^c} = \frac{1}{n_{Z^c}} \sum_{i:s_i \in Z^c} X_i$ be the sample mean functions within Z and outside Z respectively. We consider Γ_{Z,Z^c} the covariance operator of $\hat{\mu}_Z - \hat{\mu}_{Z^c}$.

To test the equality of means, a functional Hotelling's T^2_{FH} statistic is defined as:

$$T^2_{FH} = \sum_{k=1}^{K} \frac{\hat{a}_k^2}{\hat{\lambda}_k},$$

where the choice of K will be discussed further below. Here, $\hat{a}_k = \langle \hat{\mu}_Z - \hat{\mu}_{Z^c}, \hat{\psi}_k \rangle$ represents the functional principal component scores for $k \geq 1$, with $\hat{\psi}_1, \ldots, \hat{\psi}_k$ and $\hat{\lambda}_1, \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_K$ denoting the eigenfunctions and their associated eigenvalues, respectively, of $\hat{\Gamma}_{Z,Z^c}$, which is an estimate of Γ_{Z,Z^c} to be provided below. The notation $\langle u, v \rangle = \int_T u(t)v(t)dt$ denotes the standard inner product in the functional space χ.

As in the univariate and multivariate contexts, following the Gaussian spatial scan statistics proposed by Kulldorff et al. (2009) and Cucala et al. (2017), we assume that the covariance matrices inside and outside the cluster are equal, i.e., $\Gamma_Z = \Gamma_{Z^c} = \Gamma_X$. Consequently, the covariance operator of $\hat{\mu}_Z - \hat{\mu}_{Z^c}$, is given by:

$$\Gamma_{Z,Z^c} = \frac{n_Z + n_{Z^c}}{n_Z n_{Z^c}} \Gamma_X$$

which can be estimated by:

$$\hat{\Gamma}_{Z,Z^c} = \frac{n_Z + n_{Z^c}}{n_Z n_{Z^c}} \hat{\Gamma}_X,$$

where $\hat{\Gamma}_X$ is the pooled covariance, defined as:

$$\hat{\Gamma}_X(\eta) = \frac{1}{n_Z + n_{Z^c} - 2} \Big((n_Z - 1)\hat{\Gamma}_Z(\eta) + (n_{Z^c} - 1)\hat{\Gamma}_{Z^c}(\eta) \Big)$$

for $\eta \in \chi$. Here, $\hat{\Gamma}_Z$ and $\hat{\Gamma}_{Z^c}$ are the sample covariance operators of Γ_Z and Γ_{Z^c}, respectively, given by:

$$\hat{\Gamma}_Z(\eta) = \frac{1}{n_Z - 1} \sum_{i=1}^{n_Z} \langle X_i - \hat{\mu}_Z, \eta \rangle (X_i - \hat{\mu}_Z)$$

and

$$\hat{\Gamma}_{Z^c}(\eta) = \frac{1}{n_{Z^c} - 1} \sum_{i=1}^{n_{Z^c}} \langle X_i - \hat{\mu}_{Z^c}, \eta \rangle (X_i - \hat{\mu}_{Z^c}).$$

As noted by Smida et al. (2022*b*) and recommended by Cucala (2017), when defining a concentration index for constructing a scan statistic, it is essential to check that its distribution

under the null hypothesis does not depend on n_Z, the size of the potential cluster Z. It is important to note that the Hotelling's T^2_{FH} statistic is equivalent to a normalized version of the two-sample test statistics proposed by Horváth and Kokoszka (2012), known as $T^{(1)}_{N,M}$, or as $U^2_{N,M}$ in Horváth et al. (2013). Consequently, under H_0, the limiting distribution of the Hotelling's T^2_{FH} statistic (the normalized statistic of Horváth and Kokoszka, 2012) follows a chi-squared distribution with K degrees of freedom, which is fixed and does not depend on the size of the potential cluster Z. Thus, we believe the concentration index

$$I(Z) := T^2_{FH}$$

is relevant for comparing potential clusters with different population sizes.

Thus, the scan statistic can be defined as the maximum of this concentration index on the set of potential clusters $\mathcal{S}$ which has been previously defined. The Hotelling functional scan statistic (HFSS) is

$$\Lambda_{\text{HFSS}} = \max_{Z \in \mathcal{S}} I(Z)$$

and the potential cluster detected, for which Λ_{HFSS} is obtained, is

$$\hat{\text{C}} = \arg\max_{Z \in \mathcal{S}} I(Z).$$

This latter is called the Most Likely Cluster (MLC).

2.2. Computing the scan statistic

2.2.1. The choice of the threshold parameter K

The computation of the scan statistic Λ_{HFSS} requires evaluating the concentration index $I(Z)$ for each potential cluster $Z \in \mathcal{S}$, which depends on the choice of the threshold parameter K. This parameter determines the number of eigenfunctions and eigenvalues of the sample covariance matrice Γ_{Z,Z^c} to use. In practice, since our concentration index is equal to the Hotelling's T^2_H statistic, it can be applied to solve the testing problem using various values of K. The results of these tests can then be compared. However, when K equals the number of measurements k_{max} (which is finite in practice), the test statistic tends to be abnormally large, often leading to a detected cluster that contains only one observation. This issue arises due to numerical precision errors associated with small eigenvalues. Therefore, it is advisable to establish a procedure for selecting an appropriate value of K to ensure a consistent decision when applying this hypothesis to real data. To this end, we rely on the cumulative percentage of total variance (CPV) (for more details, see, for example, Horváth and Kokoszka, 2012; Horváth et al., 2013; Joseph et al., 2015). This approach is the standard method for determining the number of sample principal components to retain. For each potential cluster Z, the CPV_Z function is defined as follows:

$$CPV_Z(k) = \frac{\sum_{j=1}^{k} \hat{\lambda}_j}{\sum_{j=1}^{k_{max}} \hat{\lambda}_j},$$

where the $\hat{\lambda}_j$'s are the eigenvalues of $\hat{\Gamma}_{Z,Z^c}$ and k_{max} represents the total number of estimated eigenvalues. CPV_Z is an increasing function that approaches 1. It is calculated for each potential

cluster Z, and by averaging over all Z, we obtain a function $C\bar{P}V$ that depends only on k. The next step is to determine the optimal value K, which can be done in at least two ways. Horváth and Kokoszka (2012) recommend choosing K such that the $C\bar{P}V$ function exceeds a desired threshold, with 85% being the recommended value. Alternatively, as in Joseph et al. (2015), K can be chosen where $C\bar{P}V$ shows a marginal increase toward 1. This procedure ensures that the same value K is selected for all potential cluster Z, so that the limiting distribution of T^2_{FH} is always the same whatever potential cluster Z.

2.2.2. Algorithm for the computation of the proposed scan statistic

The algorithm used to derive Λ_{HFSS} and its associated most likely cluster $\hat{\text{C}}$ is as follows:

Algorithm 1 Computing the HFSS and the Most Likely Cluster (MLC)

Data: $\{(s_1, X_1), \ldots, (s_n, X_n)\}$, each curve X_k is observed in a finite number of points k_{max} and each spatial location s_k is expressed in Cartesian coordinates.

For all $i, j \in \{1, \ldots, n\}$, compute the distance $d_{i,j}$ between locations s_i and s_j to determine the disc $D_{i,j}$ centered at s_i and passing through s_j.

Define $\tilde{\mathcal{S}}$ as the set of potential clusters without duplicates, with size $\tilde{N}$. To create $\tilde{\mathcal{S}}$, we add the discs $D_{i,j}$ for $i, j = 1, \ldots, n$ from $\mathcal{S}$ only if a disc with the same locations isn't already included. Define Z_p, for $p = 1, \ldots, \tilde{N}$, as the potential clusters in $\tilde{\mathcal{S}}$, of size n_{Z_p} and Z_p^c its complement of size $n_{Z_p^c}$.

function HFSS (computing the HFSS scan statistic)

    Input: X, $\tilde{\mathcal{S}}$

    Output: $\hat{\Lambda}_{\text{HFSS}}$, $\hat{\text{C}}$ (MLC)

    for $p = 1$ to $\tilde{N}$ **do**

        Compute: $\hat{\mu}_{Z_p} = \frac{1}{n_{Z_p}} \sum_{k:s_k \in Z_p} X_k$ and $\hat{\mu}_{Z_p^c} = \frac{1}{n_{Z_p^c}} \sum_{k:s_k \in Z_p^c} X_k$

        Compute: $\hat{\Gamma}_{Z_p}$, $\hat{\Gamma}_{Z_p^c}$, and $\hat{\Gamma}_X = \frac{((n_{Z_p}-1)\hat{\Gamma}_{Z_p} + (n_{Z_p^c}-1)\hat{\Gamma}_{Z_p{}^c})}{n_{Z_p}+n_{Z_p^c}-2}$

        Compute eigenvalues $\hat{\lambda}_j$ and eigenvectors $\hat{a}_j$, $j = 1, \ldots, k_{max}$, of $\hat{\Gamma}_{Z_p, Z_p^c} = \frac{n_{Z_p}+n_{Z_p^c}}{n_{Z_p} n_{Z_p^c}} \hat{\Gamma}_X$

        Compute $CPV_{Z_p}(k) = \frac{\sum_{j=1}^{k} \hat{\lambda}_j}{\sum_{j=1}^{k_{max}} \hat{\lambda}_j}$, for $k = 1, \ldots, k_{max}$

        Compute $I_k(Z_p) = \sum_{j=1}^{k} \frac{\hat{a}_j^2}{\hat{\lambda}_j}$, for $k = 1, \ldots, k_{max}$

    Compute $C\bar{P}V(k) = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} CPV_{Z_p}(k)$ and select K as the point at which the function $C\bar{P}V$ begins to increase very slowly towards 1.

    $\hat{\Lambda}_{\text{HFSS}} = \max_{Z \subset \tilde{\mathcal{S}}} I_K(Z)$

    $\hat{\text{C}} = \arg\max_{Z \subset \tilde{\mathcal{S}}} I_K(Z)$

2.3. Computing the statistical significance

After computing the scan statistic Λ_{HFSS} and identifying the most likely cluster $\hat{C}$, it is necessary to assess its significance. However, the distribution of a variable window scan statistic under H_0 does not have an analytical form. To address this issue, Dwass (1957) proposed a test

procedure based on Monte Carlo simulations to approximate the null distribution. This method was later extended by Barnard (1963) and Hope (1968). The approach involves comparing the observed scan statistic to scan statistics generated from datasets simulated under H_0. Since no assumptions are made about the distribution of the functional marks, the only way to obtain such datasets is by using a method called random labeling (Cucala, 2014). In this method, a simulated dataset is created by randomly associating the functional marks X_i with the spatial locations s_i. Based on T random permutations, we consider

$$\Lambda_{\text{HFSS}}^{(1)}, \ldots, \Lambda_{\text{HFSS}}^{(T)}$$

which represent the scan statistics obtained from the simulated datasets. Then, as outlined by Dwass (1957), the p-value of the observed scan statistic Λ_{HFSS} in the initial sample is calculated as follows:

$$p_{value} = \frac{1 + \sum_{i=1}^{T} \mathbb{1}_{\{\Lambda_{\text{HFSS}}^{(i)} > \Lambda_{\text{HFSS}}\}}}{T+1}.$$

Undoubtedly, a higher number of permutations T leads to a more accurate estimation of the scan statistic's p-value. However, due to the considerable computational cost, one needs to find a trade-off between these factors. Lastly, the MLC $\hat{C}$ is considered to be statistically significant if the associated p_{value} is less than the type I error.

3. Simulation study

In the simulation study, we evaluated the performance of the Hotelling functional scan statistic (HFSS) Λ_{HFSS} introduced in the previous section. We compared it with the distribution-free functional spatial scan statistic (DFFSS) Λ_{DFFSS}, the parametric functional spatial scan statistic (PFSS) Λ_{PFSS} proposed by Frévent et al. (2021), and the nonparametric functional spatial scan statistic (NPFSS) Λ_{NPFSS} developed by Smida et al. (2022*b*).

We generated artificial datasets using the geographic locations of the administrative centers of the 94 French administrative areas, known as *départements*. The simulated true cluster, denoted by C, was defined as a group of *départements* within the Parisian region, based on two configurations: (i) 8 *départements* and (ii) 10 *départements*. Maps of the simulated clusters are available in Fig. S1 in online supplementary material. To calculate the distances between *départements*, we use the Cartesian coordinates of their centroids, expressed in meters. The size of $\mathcal{S}$ is $n^2 = 47^2 = 8836$, while after removing duplicates, the size of $\tilde{\mathcal{S}}$ is $\tilde{N} = 7044$.

At each location $s_i, i \in [1; 94]$, the functional marks X_i associated with these location take values in $\chi = L^2([0,1], \mathbb{R})$ and are generated according to the following model (see, Smida et al., 2022*b*, for more details):

$$\forall i = 1, \ldots, 94, \quad X_i(t) = \sum_{k=1}^{\infty} Z_{i,k} e_k(t) + \Delta(t) \mathbb{1}_{\{s_i \in C\}},$$

where, for all $k \geq 1$, $e_k(t) = \sqrt{2}\sin(t/\sigma_k)$ forms an orthonormal basis for χ, with $\sigma_k = ((k-0.5)\pi)^{-1}$, and the $Z_{i,k}$'s are independent random variables representing the projection of X_i onto the Karhunen-Loève basis (Karhunen, 1947; Lévy, 1965). This decomposition of the functional marks is based on the Karhunen-Loève expansion, a technique widely used in image processing and functional data analysis (Ahmed et al., 2018; Chakraborty and Chaudhuri, 2015). In our case, these marks are observed at 101 equispaced points in $[0, 1]$. We considered four scenarios:

(i) A standard Brownian motion (sBm), i.e. $Z_{i,k}/\sigma_k \backsim \mathcal{N}(0, 1)$.

(ii) A centered Student-t process on $[0, 1]$ with 4 degrees of freedom, i.e. $Z_{i,k}/\sigma_k \backsim t(4)$.

(iii) An exponential distribution with parameter 4, where $Z_{i,k}/\sigma_k \backsim \mathcal{E}xp(4)$.

(iv) A chi-squared distribution with parameter 4, where $Z_{i,k}/\sigma_k \backsim \chi^2(4)$.

Three types of clusters were simulated with an intensity controlled by a parameter $\alpha > 0$. The chosen shifts Δ, which are positive and vary over time, are given by: $\Delta_1(t) = \alpha t$, $\Delta_2(t) = \alpha t(1-t)$ and $\Delta_3(t) = \alpha \exp\left[-100(t-0.5)^2\right]/3$, for all $t \in [0, 1]$. According to Smida et al. (2022*b*), the level of spatial heterogeneity in the functional marks is entirely controlled by the parameter α, as the marks are independent. Different values of this parameter were considered for each Δ. Additionally, $\alpha = 0$ was also tested to check if the nominal type I error rate is preserved as shown in Smida et al. (2022*b*) and Frévent et al. (2021). An example of simulated data is available in Fig. S2 in online supplementary material.

To compare the four scan methods HFSS, DFFSS, NPFSS and PFSS, we created 200 simulated datasets for each distribution of the marks and each value of the cluster intensity α. For each method, we assessed three distinct criteria: the power to detect a significant cluster, the true positive rate (TP), and the false positive rate (FP). These three criteria were calculated as follows:

- The power of the test was defined as the proportion of datasets where a significant cluster was detected with a type I error rate of 0.05, using T = 199 random permutations.

- The TP was defined as the average proportion of true positive *départements* across all simulated datasets. It was computed as the ratio of the number of *départements* that are present in both the significant cluster $\hat{C}$ and the true cluster C to the total number of *départements* in C.

- The FP was calculated as the average proportion of false positive *départements*, which are the *départements* found in the most significant cluster $\hat{C}$ but not in the true cluster C, divided by the number of *départements* not included in C.

To implement the HFSS method, we selected the optimal value of K according to the CPV criterion. The aggregated $C\bar{P}V$ curves from several simulations, based on different values of the cluster intensity α, the three shifts, and the four probabilistic models,, are plotted in Fig. S3 in online supplementary material. It seems that keeping five eigenvalues prevents the $C\bar{P}V$ curves from increasing, regardless of the parameters used. Additionally, we chose $K = 5$ for each

simulation to ensure that the results are comparable across different scenarios. This choice of $K = 5$ is also supported by the theoretical simulation model, particularly within the Gaussian framework. In this context, we know the true eigenvalues of the covariance operators used in the simulation procedure (see the formula provided in the online supplementary material). The first ten cumulative percentages are 0.8122, 0.9025, 0.9350, 0.9515, 0.9616, 0.9683, 0.9731, 0.9767, 0.9795, and 0.9817, respectively. As shown, the cumulative percentages grow very slowly from the fifth eigenvalue.

The results of the simulation study are presented in Fig. 1, and Fig. 2 for scenarios where the true cluster consists of 8 *départements*. For scenarios where the true cluster consists of 10 *départements*, the results can be found in Fig. S4, Fig. S5, and Fig. S6, online supplementary material.

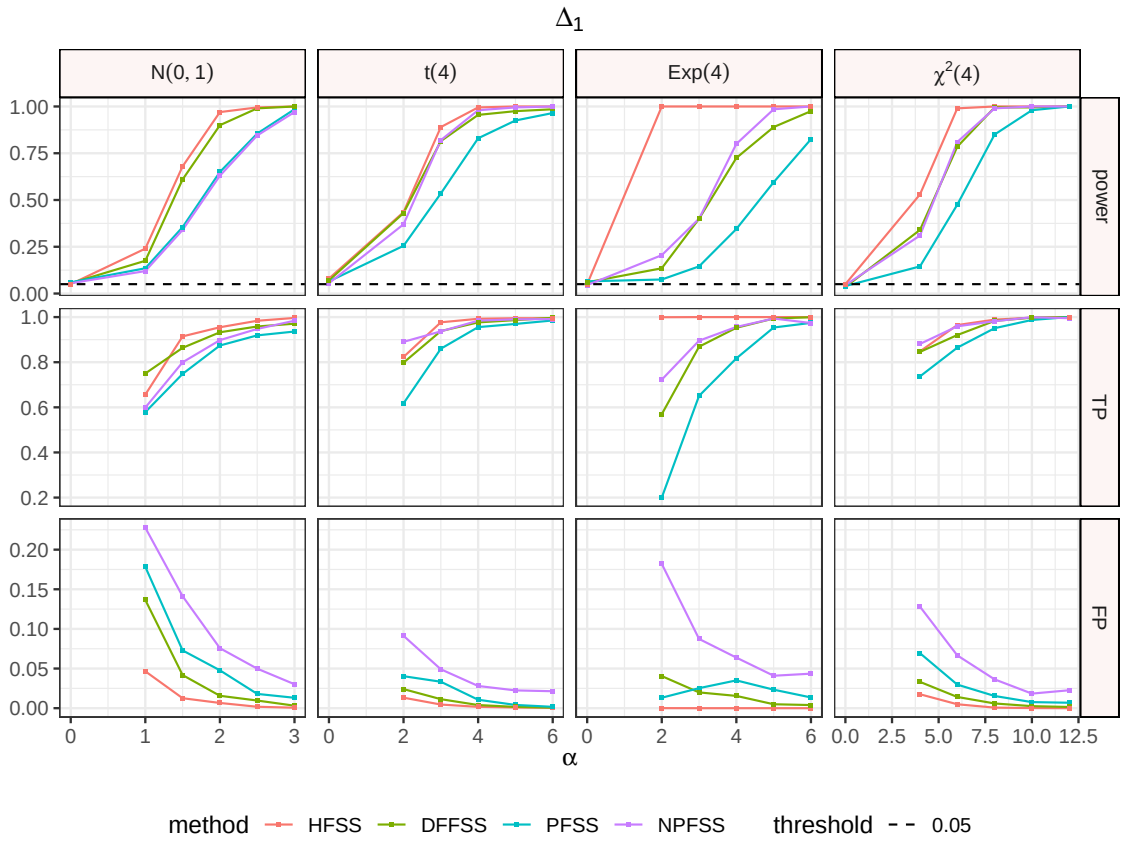

Figure 1: Simulation study- Power, TP and FP results of the HFSS, DFFSS, NPFSS and PFSS methods for the shift $\Delta_1(t) = \alpha t$ using four distributions: Normal, Student-t, Exponential and Chi-squared. The true cluster contains 8 *départements*.

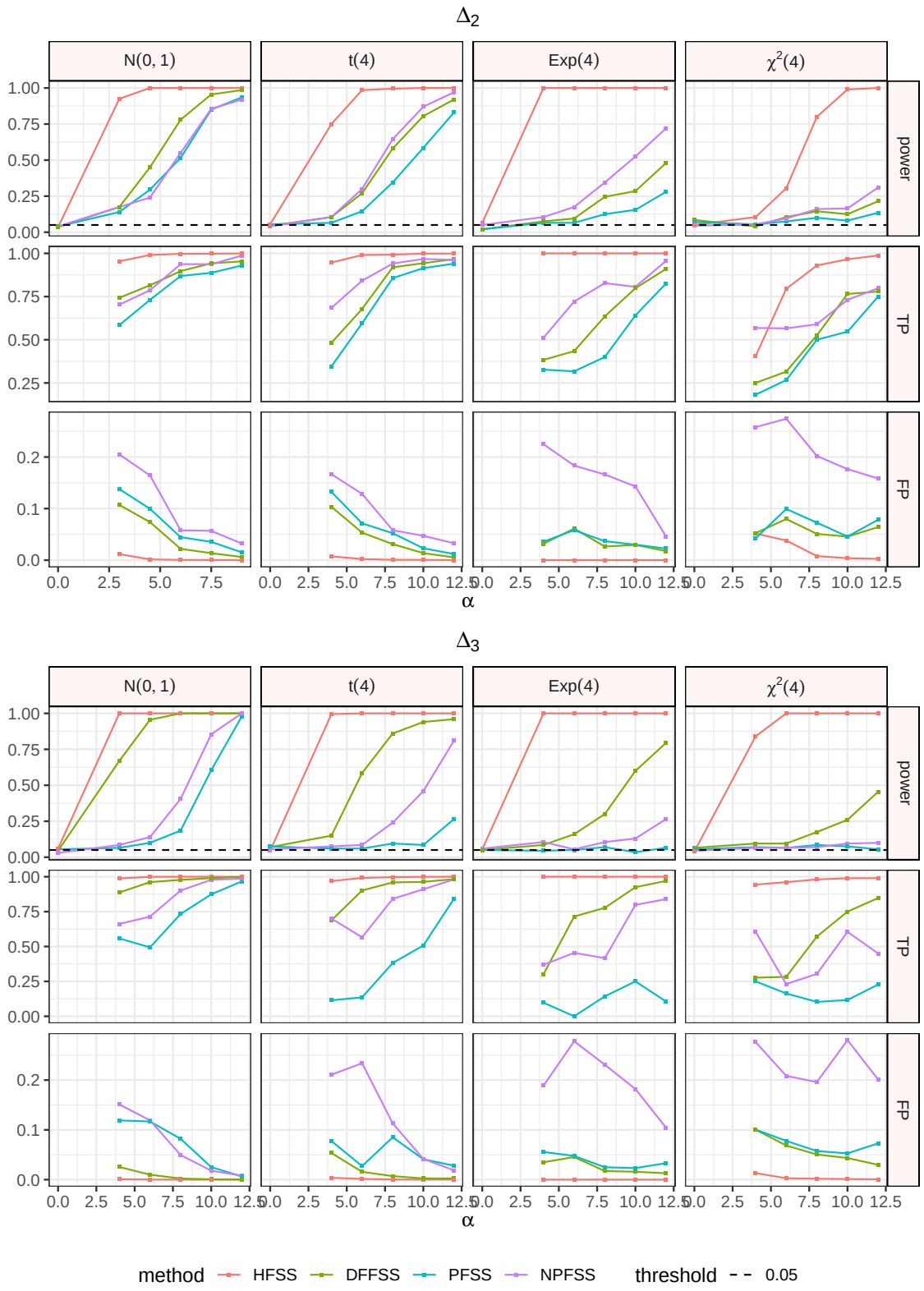

Figure 2: Simulation study- Power, TP and FP results of the HFSS, DFFSS, NPFSS and PFSS methods for the shift $\Delta_2(t) = \alpha t(1-t)$ and $\Delta_3(t) = \alpha \exp\left[-100(t-0.5)^2\right]/3$, using four distributions: Normal, Student-t, Exponential and Chi-squared. The true cluster contains 8 *départements*.

Based on the power curves shown in Fig. 1, and Fig. 2, the sizes of the different methods (i.e., the power when $\alpha = 0$) are close to the nominal type I error rate of 0.05, regardless of the distribution of the marks. The performance of all scan statistics tends to increase with higher cluster intensity, α. Notably, the power of the proposed scan statistic HFSS is consistently higher than that of DFFSS, PFSS, and NPFSS across all cases whatever the size of the true cluster C and the shift Δ, particularly in the case of Δ_3. On the other hand, power grows more slowly with Student-t or Chi-squared distributions compared to normal or exponential ones, due to their heavier tails. Additionally, for a fixed cluster intensity α, the power increases as the cluster size grows from 8 to 10 (see Fig. S4, Fig. S5, and Fig. S6 in online supplementary material), since larger clusters are easier to detect.

The true positive rate and false positive rate also improve as the cluster intensity α increases (higher TP and lower FP). The true positive of the proposed HFSS is consistently higher than that of all other methods, especially for shifts Δ_2 and Δ_3, regardless of the true cluster size C. In terms of false positives, HFSS consistently shows the lowest false positive rates, with a notable difference observed for shifts Δ_2 and Δ_3. According to Frévent et al. (2021), NPFSS exhibits the highest false positive rates. However, it is also evident that NPFSS achieves higher true positive rates for shifts Δ_1 and Δ_2 compared to DFFSS and PFSS (which have the lowest true positive rates) whatever the size of the true cluster.

4. Application to real data

In this study, we will employ the scan statistic for functional data to explore two types of applications: economic and climatic data. For the economic data, the scan statistic will be applied to detect spatial clusters of abnormal unemployment rates in Spain over time. For the climatic data, the aim is to monitor environmental changes by identifying spatial clusters of unusual climate change patterns over time in Great Britain, Nigeria, Pakistan, and Venezuela.

Comparison of the procedures. To identify spatial clusters of low or high unemployment rates in Spain and unusual climate change in Great Britain, Nigeria, Pakistan, and Venezuela, we used the HFSS, DFFSS, NPFSS, and PFSS methods to detect the MLCs with high concentration indices. Additionally, we examined the presence of the most likely secondary clusters (2MLC) following the approach outlined by Zhang et al. (2010) (see also Smida et al., 2022*b*). For all four methods, the set of potential clusters $\mathcal{S}$ is constructed as described in subsection 2.1. The significance of the MLCs and 2MLCs was evaluated through $T = 999$ Monte Carlo permutations, with statistical significance defined as a p-value below 0.05, as detailed in subsection 2.3.

4.1. Application to unemployment rates in Spain

Unemployment rate data in Spain. Given that Spain has consistently been one of the countries with the highest unemployment rates in the European Union, we conducted an analysis of unemployment rate data sourced from the Spanish Institute of Statistics (`www.ine.es`). This data spans each quarter from 2002 to 2022 (a total of 80 quarters) across all 47 provinces of Spain (see Fig. S7 in online supplementary material, left panel). The right panel of Fig. S7 presents a chart illustrating significant fluctuations in unemployment rates over time. Fig. S9 in online supplementary material shows the spatial distribution of the average unemployment rate from 2002 to 2022, highlighting that high unemployment rates were mainly in the south of Spain,

while lower rates were found in the north. Additionally, Fig. S8 in the online supplementary material, which displays the mean unemployment rate every two years, indicates a diverse pattern across provinces. Thus, we propose using functional spatial scan statistics to identify spatial clusters of unemployment at the provincial level.

Cluster detection results. Here, the size of $\mathcal{S}$ is $n^2 = 47^2 = 2209$, whereas the size of $\tilde{\mathcal{S}}$ after removing duplicates is $\tilde{N} = 1613$. For the HFSS method, we select $K = 2$ based on the CPV criteria (see Fig. S10 in online supplementary material for plots of $C\bar{P}V$), which explains approximately 90% of the total variance. Fig. 3 shows the significant spatial clusters (MLCs and 2MLCs) identified by the HFSS, DFFSS, NPFSS, and PFSS methods. For the MLCs, the p-values for all methods are zero up to two decimal places (or 0.001). For the most likely secondary clusters, the HFSS method has the most significant p-value (0.003), though the p-values for the other methods are also below the type I error threshold of 0.05. The complete results are presented in Table 1 in online supplementary material.

The HFSS, DFFSS, and PFSS methods identified exactly the same MLC of 13 provinces in southern Spain. In contrast, the NPFSS detected a larger MLC of 17 provinces, encompassing the same 13 provinces plus 4 additional ones. The unemployment rate in these detected regions was consistently higher than the rest of Spain throughout the period. These MLCs were homogeneous, as they included provinces with unemployment rates above the national average (except for two provinces at the end of the study period and the larger cluster detected by NPFSS).

For the secondary clusters, the HFSS and NPFSS methods detected the same 2MLC consisting of 15 provinces. The PFSS method identified a slightly larger 2MLC, which included these 15 provinces plus one additional province. In contrast, the DFFSS method detected a much larger 2MLC of 22 provinces. All 2MLCs were located in northern Spain, except for the DFFSS cluster, which also extended into the northeast. All these clusters detect the provinces with the lower unemployment rates between 2002 and 2022. Notably, the unemployment rate curves for the HFSS and NPFSS secondary clusters were below the national average curve.

To conclude, this first application revealed that the detected clusters (MLCs or 2MLCs) correspond to curves consistently positioned above or below the others, with all four methods performing comparably. However, the HFSS and PFSS methods have the advantage of producing more reasonable cluster sizes.

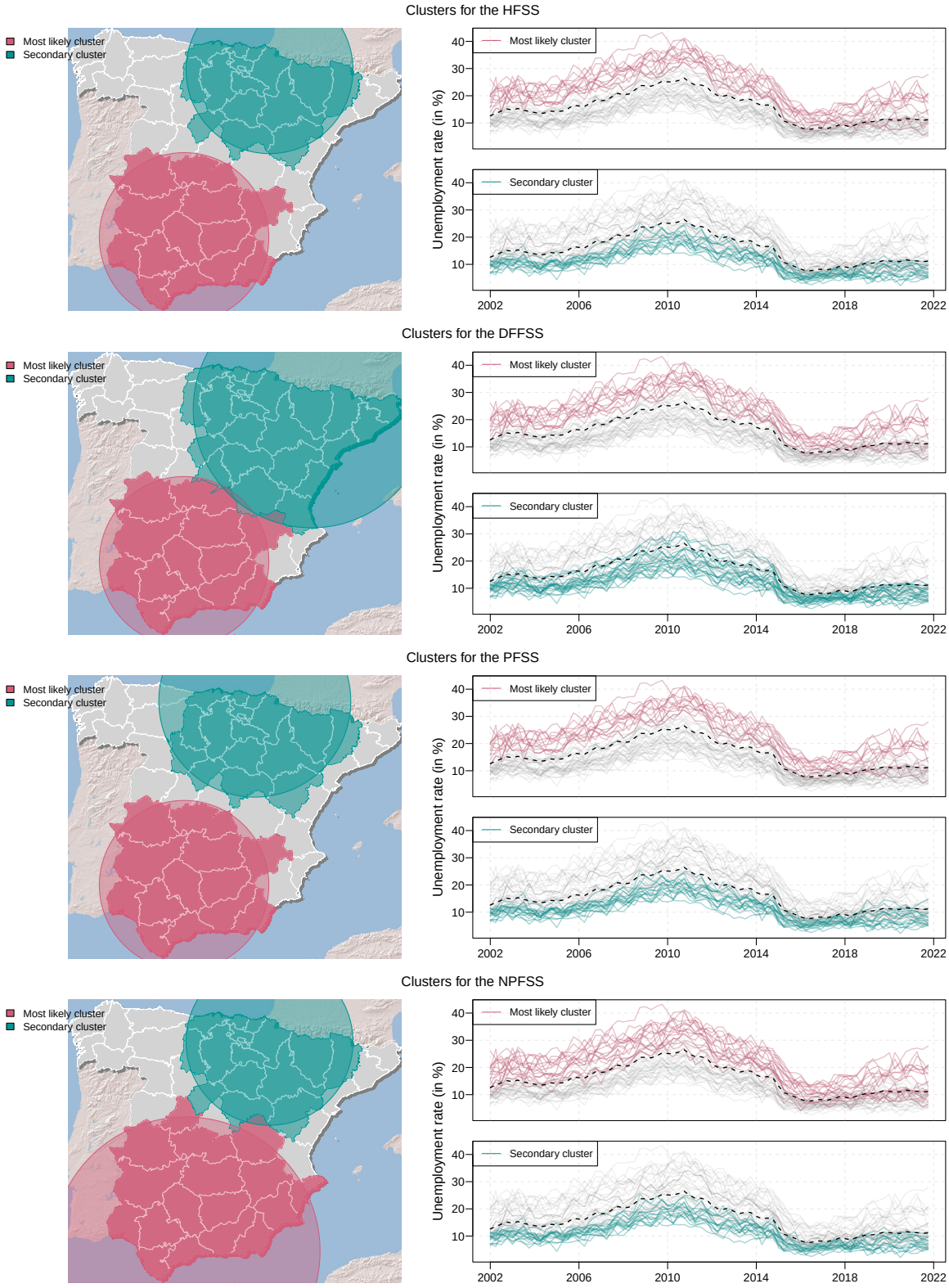

Figure 3: Clusters for the HFSS, DFFSS, PFSS, NPFSS methods, for the variable Unemployment rate (in %), observed quarterly from 2002 to 2022 in Spain.

4.2. Application to climate data

Some changes, such as droughts, heat waves, and extreme rainfall, are occurring faster than scientists estimated. In this work, we focus on annual indicators of global climate change from 2010 to 2023 (i.e., 24 measurement points), such as the number of heat waves days, the difference in temperatures from seasonal norms, and extreme precipitation events. Additionally, we selected one country per continent (Great Britain in Europe, Nigeria in Africa, Pakistan in Asia, and Venezuela in America) to determine whether we could detect regions where climate change is expressed differently. To achieve this, we used the Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2), provided by the National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) from 1981 to 2023. MERRA-2 is fully described in Gelaro et al. (2017) and has been used in many recent scientific articles (see, for instance, `https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/pubs/`). All of the MERRA-2 atmospheric variables are provided at $0.5^\circ \times 0.625^\circ$ spatial resolution.

4.2.1. Difference from average temperatures in Great Britain

Temperature Data for Great Britain. We consider the island of Great Britain, which includes the countries England, Scotland and Wales. The island is divided into 145 cells, covering around 209 331 km^2 (Fig. S11, left panel). We focus on the difference from average temperatures, observed from 2010 to 2023 (Fig. S11, right panel). This difference is defined as the deviation between the average yearly temperature and the normal yearly temperature, which is computed from the period 1981-2010.

The spatial distributions of the average temperature over the entire observed time period (Fig. S13) were heterogeneous. Lower temperatures tended to aggregate in northern Scotland, while higher temperatures were observed in southeastern England. Additionally, in Fig. S12 in online supplementary material, we map the average of the variable aggregated over a two-year window. This mapping reveals that the period 2021–2023 was particularly warmer across almost all regions. During 2018–2020, the southeast experienced the most significant rise in temperatures. The periods 2000–2002, 2009–2011 and 2012–2014 were not far from normal temperatures. Finally, between 2003 and 2006, the north of Great Britain experienced temperatures higher than normal compared to the rest of the country. Moreover, according to Christidis et al. (2020), in addition to the noticeable contrast between warmer summers in the south and cooler temperatures in the north, southeast England stands out as the region where high temperature extremes are most likely to occur. Here, we are exploring whether scan statistics can identify results comparable to those observed.

Cluster detection results. In this instance, $\mathcal{S}$ contains $n^2 = 145^2 = 21025$ elements, while after eliminating duplicates, the size of $\tilde{\mathcal{S}}$ decreases to $\tilde{N} = 13644$. For running the HFSS method, we selected $K = 6$ based on the CPV criteria (refer to Fig. S14 in online supplementary material for the plots of $C\bar{P}V$), which explains aapproximately 94% of total variance. The MLCs and 2MLCs detected by all methods are significant at level 1‰ (see, Table 2 online supplementary material). In Figure 4, we observe that the MLCs identified by the NPFSS, PFSS and DFFSS are very similar. The MLCs contain 73 cells for NPFSS and PFSS, and 104 cells for DFFSS. The observations are concentrated in a large southern part of Great Britain, where the temperatures are essentially higher than in other areas throughout the observation period. The 2MLCs identified by NPFSS, PFSS, and DFFSS contain observations located just above the MLCs.

The sizes of these secondary clusters are smaller than the MLCs, with 30, 52, and 20 cells, respectively.

Moreover, the MLC obtained by the HFSS method is of reasonable size (50 cells). It is located at the extreme north of the island, where the differences from average temperatures are the smallest throughout the observation period, except for the period 2003–2005, when the increases in temperature compared to normal were greater. In contrast, the 2MLC (31 cells in size) is located in the southeastern part of the island, where the difference from average temperatures are the highest, except for the period 2003–2005, when the increases in temperature compared to normal were greater.

Finally, the HFSS method stands out by identifying two clusters that differ significantly in both size and geographic location. In this case, the detected clusters no longer correspond to curves consistently positioned above or below the others. Instead, the identified curves alternate between periods where they are sometimes above and other times below the remaining curves.

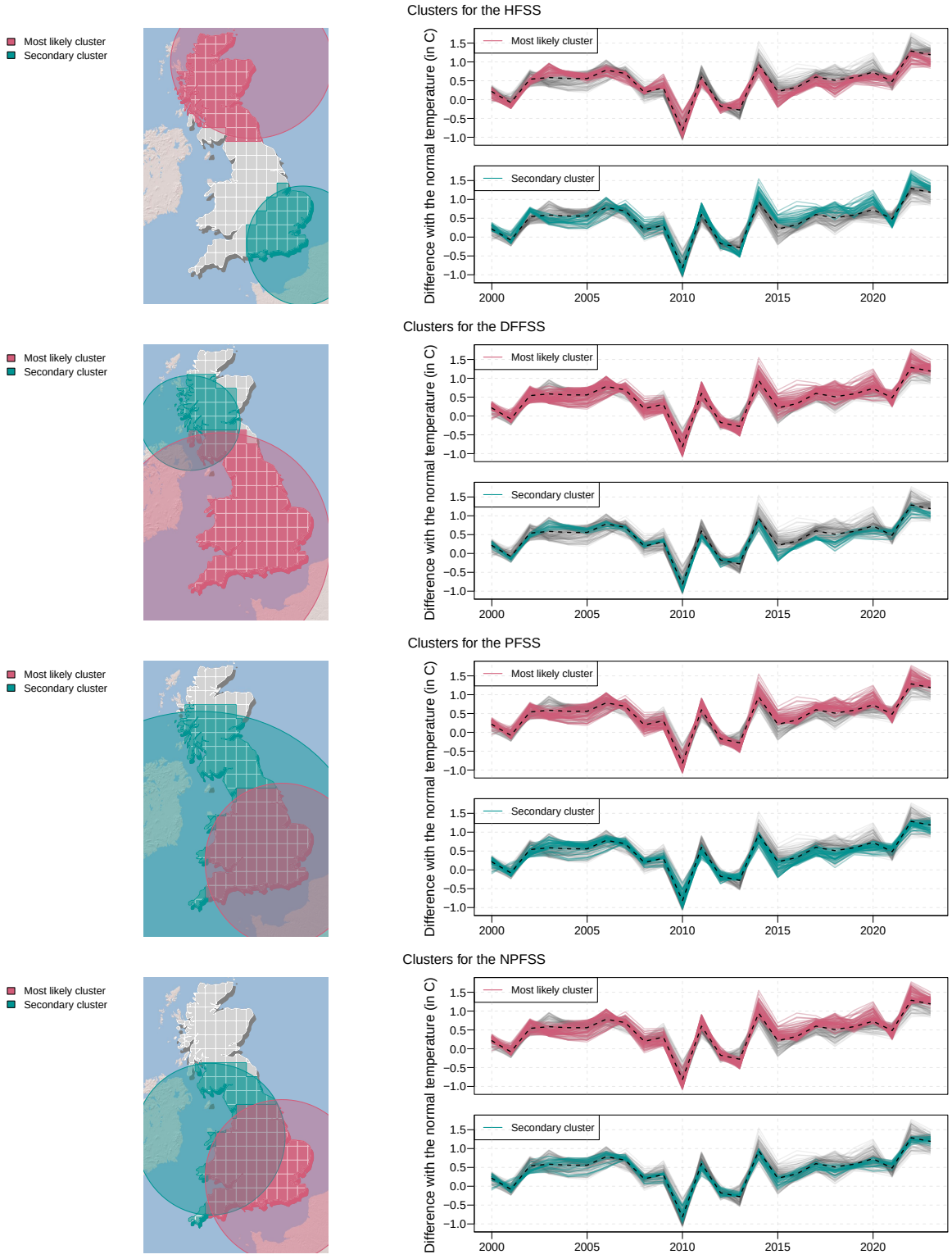

Figure 4: Clusters for the HFSS, DFFSS, PFSS, NPFSS methods, for the variable difference with the normal temperatures (in degree Celsius) in Great-Britain.

4.2.2. Maximum consecutive 5-days precipitation in Nigeria

Precipitation Data in Nigeria. Nigeria, situated in West Africa, has experienced severe flooding in recent years, displacing millions of people, destroying homes, and causing damage worth billions of naira (local currency). According to Dike et al. (2020), the maximum consecutive 5-day rainfall totals have increased across Nigeria's three climatic zones: the Guinea coast (south), Sub-Sahel (central), and Sahel (north) regions. To examine these changes in detail, we analyze data from 285 grid cells that cover approximately 923 768 km^2 (Fig. S15, left panel). We focus on the variable maximum consecutive 5-days precipitation (in mm), observed from 2010 to 2023 (Fig. S15, right panel). This variable represents the highest amount of cumulative precipitation observed over any five consecutive days within a year.

The spatial distribution of the average maximum consecutive 5-day precipitation over the entire observed period was heterogeneous (see Fig. S17 in online supplementary material). Higher values of maximum consecutive 5-day precipitation were generally observed in the southeast and northern regions near the center, while lower values tended to be in the northeastern part of the country. To further explore these trends, Fig. S16 maps the average of this variable over a three-year window. This mapping shows that, before 2018, heavy rainfall mainly impacted the southern region of Nigeria. However, starting in 2019, heavy rainfall began to impact the entire country, with particular intensity in the central northern regions. Here, we aim to determine whether scan statistics can identify these regions and evaluate whether heavy precipitation patterns have changed differently in various areas or if some regions are at greater risk than others.

Cluster detection results. In this example, the size of $\mathcal{S}$ is $n^2 = 285^2 = 81225$, but after removing duplicates, the size of $\tilde{\mathcal{S}}$ reduces to $\tilde{N} = 70172$. For the HFSS method, we select the value $K = 6$ based on the CPV criteria (see Fig. S18 in online supplementary material for plots of $C\bar{P}V$), which explains approximately 94% of the total variance. The MLCs and 2MLCs detected by all methods are significant at level 1‰ (see Table 3 in online supplementary material). Fig. 5 displays the significant spatial clusters identified by the HFSS, DFFSS, NPFSS, and PFSS methods. It shows that HFSS and PFSS produce similar results, both identifying a small cluster (9/285 observations) in Bauchi State (north-central) with peaks in 2020 and 2022. In September 2022, this area experienced severe flooding that affected 2185 individuals due to heavy rain and strong winds (International Organization for Migration, 2022). This region is also identified as 2MLC by the DFFSS method. Additionally, the MLC detected by the DFFSS method (which corresponds to secondary cluster identified by the PFSS, HFSS, and NPFSS methods) is located in southern Nigeria. This region, characterized by an equatorial climate, receives the highest rainfall and has a hot, humid climate with heavy precipitation.

To conclude, the HFSS, DFFSS, and PFSS methods successfully identified a cluster with curves showing a sudden shift in behavior, similar to a shock in time series. However, the MLC detected by the NPFSS method is less interesting due to its large size (200 out of 285), making it less useful compared to the other methods.

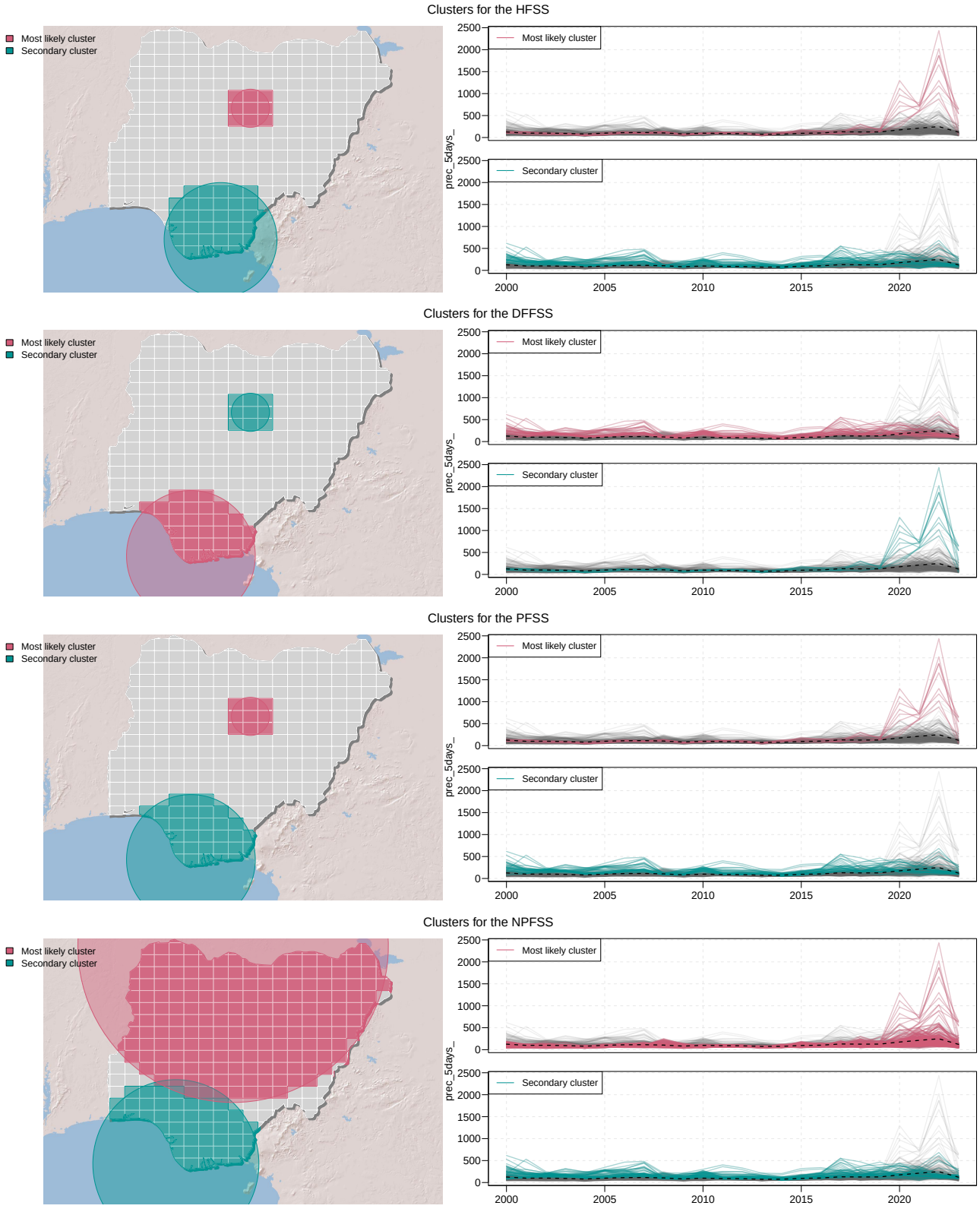

Figure 5: Clusters for the HFSS, DFFSS, PFSS, NPFSS methods, for the variable Maximum consecutive 5-days precipitation (in mm), in Nigeria.

4.2.3. Maximum consecutive 5-days precipitation in Pakistan

Precipitation Data in Pakistan. According to United Nations Human Settlements Programme (2023), Pakistan ranks among the top 10 countries most vulnerable to climate change. During the monsoon season, from July to August, the country typically receives an average of 255 mm of rainfall per month. However, in the summer of 2022 (June–September), the average daily rainfall surged to 3.95 mm, which was 283% higher than the 42-year average of 1.03 mm/day and exceeded the inter-annual standard deviation by sevenfold, based on the Climate Prediction Center's (CPC) unified precipitation analysis (You et al., 2024). To better understand these variations, we divided Pakistan into 306 cells, covering an area of approximately 881,913 km^2 (Fig. S19, left panel). Our analysis focuses on the "Maximum consecutive 5-day precipitation" variable (in mm), observed from 2010 to 2023 (Fig. S19, right panel), as previously outlined.

The spatial distribution of the average maximum consecutive 5-day precipitation over the entire period was varied, as shown in Fig. S21 in online supplementary material. Additionally, Fig. S20 in online supplementary material presents the average precipitation over a three-year period. This figure indicates that before 2018, the southeastern and northern regions experienced slightly higher precipitation compared to the rest of the country. However, starting in 2019, heavy rainfall began to impact nearly every region of Pakistan. Thus, we seek to assess whether scan statistics can reveal regions with varying levels of impact from heavy rainfall.

Cluster detection results. In this scenario, the size of $\mathcal{S}$ is $n^2 = 306^2 = 93636$, while after removing duplicates, the size of $\tilde{\mathcal{S}}$ is $\tilde{N} = 71053$. For the HFSS method, we chose $K = 6$ based on the CPV criteria (refer to Fig. S22 in online supplementary material for $C\bar{P}V$ plots), which explains about 90% of the total variance. All MLCs and 2MLCs detected by the methods are significant at the 1‰ level (see Table 4 in online supplementary material). Figure 6 shows that the same MLC is identified by both the NPFSS and PFSS methods. This MLC is quite large (226/305) and corresponds to the eastern part of the country. The DFSSS and HFSS methods deliver particularly valuable results, as they identify clusters of reasonable size in notable regions. The MLC detected by the DFSSS method corresponds to Kashmir, which has seen significant heavy rainfall in the past, particularly in 2014. The 2MLC identified by the same method is in the Sindh region (including Karachi), which has experienced heavy rainfall in several years, including 2020, 2017, and 2015. In contrast, the HFSS method identifies the highest peak observed over the years, which occurred in 2021 in northern Balochistan. This MLC is relatively small (11/305) and is situated near Quetta. In 2021, the total precipitation over a five-day period in this area exceeded 600 mm. Additionally, the 2MLC (of size 31) corresponds to a region heavily affected by the floods in 2020 and 2022 (see, for instance, the map produced by Joint Research Center, 2022, for the European Union).

In this example, the HFSS method, similar to the DFFSS and PFSS methods, was effective in detecting clusters that represent sudden changes in the functional data. This capability enables the identification of regions experiencing rapid climate change.

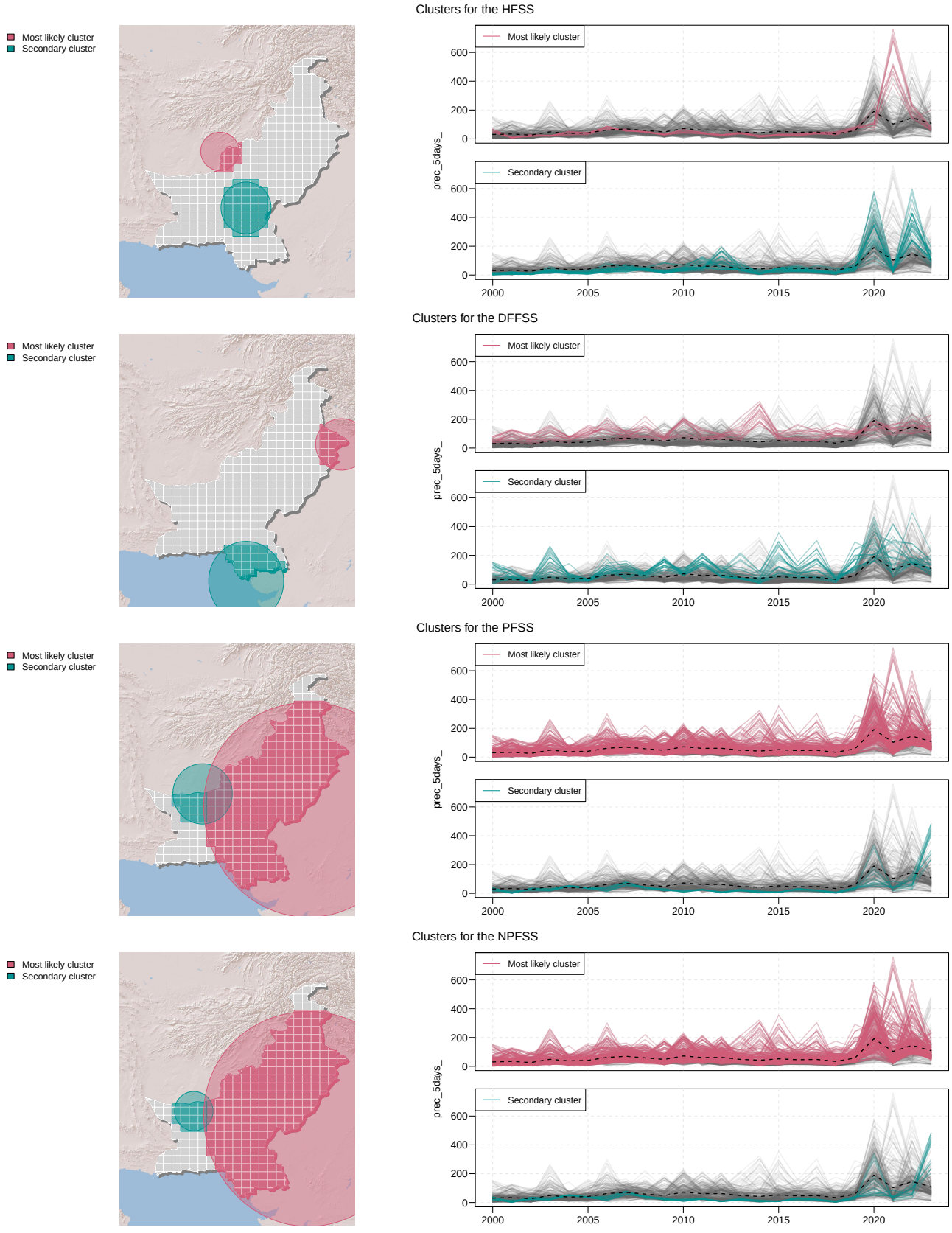

Figure 6: Clusters for the HFSS, DFFSS, PFSS, NPFSS methods, for the variable Maximum consecutive 5-days precipitation (in mm), in Pakistan.

4.2.4. Heat wave days frequency in Venezuela

Heat wave Data in Venezuela. The impacts of climate change in Venezuela are evident through rising air temperatures, melting glaciers, shrinking polar ice caps, and increased desertification. Additionally, the country is experiencing more frequent extreme weather events, including heat waves, droughts, floods, and storms. According to Feron et al. (2019), the proportion of extremely warm days in Northern South America during the austral summer (December to February) has more than doubled in recent decades. Here, the country is divided into 304 cells, covering approximately 916 445 km^2 (Fig. S23, left panel). We analyze the number of heat wave days per year from 2010 to 2023 (Fig. S23, right panel). A heat wave is defined as a period of at least three consecutive days when the daily mean temperature exceeds the 90th percentile (or falls below the 10th percentile) for a 15-day running window during the baseline period (1981–2010). The average spatial distribution of heat waves across the observation period (Fig. S25) varied considerably, with the northwestern and eastern regions of the country being notably vulnerable. In Fig. S24 in online supplementary material, we present a three-year average of heat waves. This mapping shows that before 2018, the southeastern region was somewhat more affected. Between 2018 and 2020, the northwestern region also began experiencing significant heat waves. In the most recent period, heat waves were widespread, with the northwestern region facing the greatest impact. Therefore, we use scan statistics to investigate whether climate change affects some regions more than others.

Cluster detection results. In this example, the size of $\mathcal{S}$ is $n^2 = 304^2 = 92416$, and after removing duplicates, the size of $\tilde{\mathcal{S}}$ is $\tilde{N} = 71153$. For the HFSS method, we selected $K = 6$ based on the CPV criteria (see Fig. S26 in online supplementary material for $C\bar{P}V$ plots), which accounts for approximately 90% of the total variance. The MLCs and 2MLCs identified by all methods are significant at the 1‰ level (see Table 5 in online supplementary material). Figure 7 shows that the MLCs and 2MLCs detected by NPFSS, PFSS, and DFFSS are highly comparable. However, the MLCs found by NPFSS are notably larger, covering 105 out of 304 cells, while those identified by the other methods cover 73 out of 304 cells each. In contrast, the HFSS method detects clusters that are both smaller and differently located. The MLC identified by HFSS corresponds to the Amazon rainforest in southeastern Venezuela and covers a more modest area of 35 out of 304 cells (see Table 5 in online supplementary material). According to Flores et al. (2024), this region is increasingly vulnerable to stress from rising temperatures, extreme droughts, deforestation, and fires. The curves corresponding to this MLC show that this region experienced a higher frequency of heat wave days between 2010 and 2020 compared to the rest of the country. In addition, the HFSS method also identifies a 2MLC in the Apure region of northwestern Venezuela. Covering 71 cells, this area has been notably impacted by severe heat waves in recent years, with some cells recording over 200 heat wave days in both 2020 and 2021.

In this final example, all methods successfully identified two regions affected by climate change: one with consistently high heat waves over several years (the Amazon), and another that experienced a rapid increase in heat waves over the past four years.

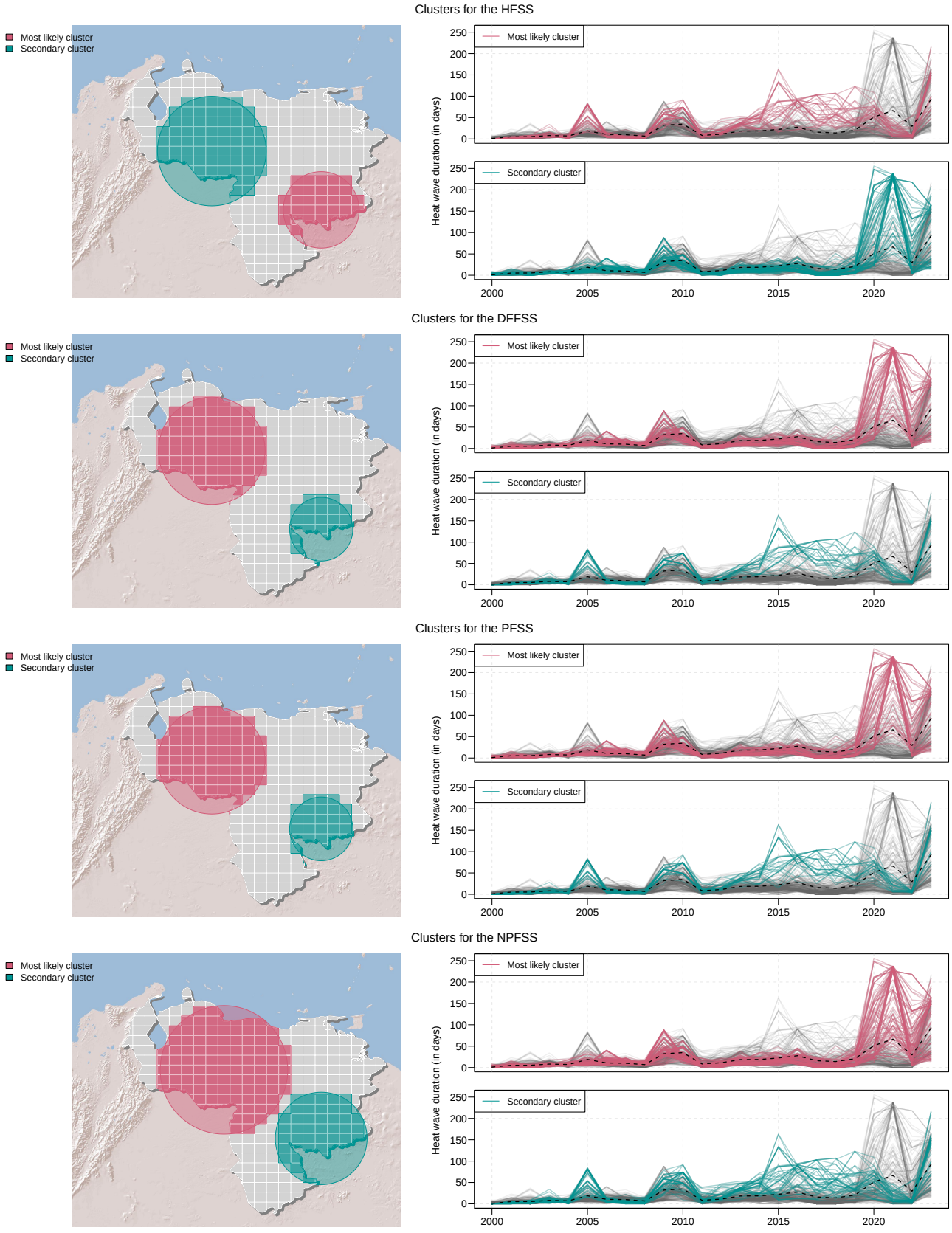

Figure 7: Clusters for the HFSS, DFFSS, PFSS, NPFSS methods, for the variable Heat wave duration (in number of days), in Venezuela.

5. Discussion

In this article, we developed a HFSS scan statistic, for analyzing spatially indexed functional data. It is based on the Hotelling T^2_{FH} test statistic (Joseph et al., 2015) (or Horváth et al., 2013) and utilizes a concentration index with a known distribution. As Kulldorff (2006) noted, the main goal of cluster detection is to raise an alarm, prompting scientists to investigate areas with unusual values. The proposed HFSS method is particularly effective for this when dealing with complex, infinite-dimensional data.

The proposed scan statistic HFSS extends the Gaussian spatial scan statistic introduced by Kulldorff et al. (2009) for univariate data and later adapted for multivariate data by Cucala (2017). Both of these methods rely on a concentration index based on the Gaussian likelihood ratio. As noted by Cucala et al. (2017) and Cucala (2022), the multivariate Gaussian concentration index is equivalent to the T^2_H test statistic introduced by Hotelling (1931), as elaborated by Anderson (2003). This equivalence is demonstrated and detailed separately for both univariate and multivariate cases within the context of scan statistics in this article. Additionally, the advantage of using the Hotelling test, as mentioned by Johnson and Wichern (2002), is that it remains remarkably unaffected by slight departures from normality and the presence of a few outliers when the sample size is moderate to large.

We employed simulations and real-world datasets (economic and climate) to compare our HFSS method with the DFFSS and PFSS methods introduced by Frévent et al. (2021) and the NPFSS method proposed by Smida et al. (2022*b*). When applying the HFSS method, selecting the number of functional principal components is crucial. We suggest using the cumulative percentage of total variance for this purpose, as discussed in detail in this work.

The simulation study results demonstrated that all methods maintained the nominal type I error. However, the HFSS method outperformed the others in terms of power, true positive rate, and false positive rate, regardless of the true cluster size, distribution nature (Gaussian or other), or shift type. Notably, HFSS showed significantly better power in detecting spatial clusters that appear over short periods, particularly in cases of quadratic and exponential shifts. It also proved especially effective when dealing with functions exhibiting extremely high values from light-tailed distributions, such as the exponential distribution. Given that many real datasets, such as those from economics, climate, environmental studies, epidemiology, etc., often deviate from a Gaussian distribution, HFSS may be particularly effective for cluster detection. This method is capable of identifying abrupt changes in functions that other techniques might miss.

In our applications, the HFSS method performed well in identifying significant clusters. These clusters were usually moderate in size compared to those found by the NPFSS method, making them easier to interpret. They usually aligned with the clusters found by the DFFSS and PFSS methods. However, our new HFFS method allowed us to discover different clusters that displayed specific behaviors in their curves. Notably, it was able to detect sudden changes, like shocks, which isn't always the case with other functional scan methods. In particular, it can detect sudden ruptures, such as shocks, which does not appear to be the case for all functional scan methods.

From the perspective of this work, several key considerations arise. Our use of the spatial scan statistic relies on the assumption that observations are independent, a standard approach also adopted by Smida et al. (2022*b*) and Frévent et al. (2021). However, this assumption may be violated in the presence of spatial autocorrelation, potentially leading to an inflated Type I error

rate in the random permutation process, as noted by Ahmed et al. (2021), Lee et al. (2020), and Loh and Zhu (2007). Therefore, addressing the impact of spatial autocorrelation in univariate functional scan statistics is a complex area that requires further exploration. Moreover, while our method, like those of Smida et al. (2022*b*) and Frévent et al. (2021), is specifically designed for circular clusters, it can also be adapted to other shapes, such as elliptical Kulldorff (2006) and graph-based clusters Cucala et al. (2013). Finally, since multiple curves may be observed at each spatial location, another approach could be to develop a functional Hotelling spatial scan for functional data. Comparing this method with the multivariate extensions of DFFSS, PFSS, and NPFSS proposed by Frévent et al. (2023) and implemented in the `HDSpatialScan` `R` package Frévent et al. (2022) could provide valuable insights.

References

Ahmed, M., Attouch, M. and Dabo-Niang, S. (2018), 'Binary functional linear models under choice-based sampling', *Econometrics and Statistics* **7**, 134–152.

Ahmed, M.-S., Cucala, L. and Genin, M. (2021), 'Spatial autoregressive models for scan statistic', *Journal of Spatial Econometrics* **2**(1), 1–20.

Alm, S. E. (1997), 'On the distributions of scan statistics of a two-dimensional poisson process', *Advances in Applied Probability* **29**(1), 1–18.

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics, Wiley.

Aneiros, G., Cao, R., Fraiman, R., Genest, C. and Vieu, P. (2019), 'Recent advances in functional data analysis and high-dimensional statistics', *Journal of Multivariate Analysis* **170**, 3–9. Special Issue on Functional Data Analysis and Related Topics.

Aneiros, G., Horová, I., Hušková, M. and Vieu, P. (2022), 'On functional data analysis and related topics', *Journal of Multivariate Analysis* **189**, 104861.

Assunção, R., Costa, M., Tavares, A. and Ferreira, S. (2006), 'Fast detection of arbitrarily shaped disease clusters', *Statistics in Medicine* **25**(5), 723–742.

Barnard, G, A. (1963), 'Discussion of Professor Bartlett's Paper', *Journal of the Royal Statistical Society: Series B (Methodological)* **25**, 284.

Bhatt, V. and Tiwari, N. (2014), 'A spatial scan statistic for survival data based on Weibull distribution', *Statistics in Medicine* **33**(11), 1867–1876.

Boente, G. and Fraiman, R. (2000), 'Kernel-based functional principal components', *Statistics & Probability Letters* **48**(4), 335–345.

Chakraborty, A. and Chaudhuri, P. (2015), 'A Wilcoxon-Mann-Whitney-type test for infinite-dimensional data', *Biometrika* **102**(1), 239–246.

Chen, J. and Glaz, J. (2009), Approximations for two-dimensional variable window scan statistics, *in* J. Glaz, V. Pozdnyakov and S. Wallenstein, eds, 'Scan Statistics: Methods and Applications', Birkhäuser Boston, Boston, MA, pp. 109–128.

Christidis, N., McCarthy, M. and Stott, P. (2020), 'The increasing likelihood of temperatures above 30 to 40°C in the United Kingdom', *Nature Communications* **11**, 3093.

Cucala, L. (2014), 'A distribution-free spatial scan statistic for marked point processes', *Spatial Statistics* **10**, 117–125.

Cucala, L. (2016), 'A Mann–Whitney scan statistic for continuous data', *Communications in Statistics - Theory and Methods* **45**(2), 321–329.

Cucala, L. (2017), Variable window scan statistics: Alternatives to generalized likelihood ratio tests, *in* J. Glaz and M. V. Koutras, eds, 'Handbook of Scan Statistics', Springer New York, New York, NY, pp. 1–16.

Cucala, L. (2022), Détection d'agrégats temporels, spatiaux et spatio-temporels: contributions aux méthodes de balayage, Hdr, Université de Montpellier.

Cucala, L., Demattei, C., Lopes, P. and Ribeiro, A. (2013), 'A spatial scan statistic for case event data based on connected components', *Computational Statistics* **28**(1), 357–369.

Cucala, L., Genin, M., Lanier, C. and Occelli, F. (2017), 'A multivariate Gaussian scan statistic for spatial data', *Spatial Statistics* **21**, 66–74.

Cucala, L., Genin, M., Occelli, F. and Soula, J. (2019), 'A multivariate nonparametric scan statistic for spatial data', *Spatial Statistics* **29**, 1–14.

Cuevas, A., Febrero, M. and Fraiman, R. (2004), 'An ANOVA test for functional data', *Computational Statistics & Data Analysis* **47**(1), 111–122.

Dabo-Niang, S. and Frévent, C. (2024), 'Uncovering data across continua: An introduction to functional data analysis'.

Dematteï, C., Molinari, N. and Daurès, J.-P. (2007), 'Arbitrarily shaped multiple spatial cluster detection for case event data', *Computational Statistics & Data Analysis* **51**(8), 3931–3945.

Dike, V. N., Lin, Z.-H. and Ibe, C. C. (2020), 'Intensification of summer rainfall extremes over Nigeria during recent decades', *Atmosphere* **11**(10).

Duczmal, L. and Assunção, R. (2004), 'A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters', *Computational Statistics & Data Analysis* **45**(2), 269–286.

Dwass, M. (1957), 'Modified randomization tests for nonparametric hypotheses', *The Annals of Mathematical Statistics* **28**(1), 181–187.

Feron, S., Cordero, R., Damiani, A., Llanillo, P., Jorquera, J., Sepúlveda Araya, E., Asencio, V., Laroze, D., Labbe, F., Carrasco, J. and Torres, G. (2019), 'Observations and projections of heat waves in South America', *Scientific Reports* **9**, 8173.

Ferraty, F., Goia, A., Salinelli, E. and Vieu, P. (2011), Recent advances on functional additive regression, *in* F. Ferraty, ed., 'Recent Advances in Functional Data Analysis and Related Topics', Physica-Verlag HD, Heidelberg, pp. 97–102.

Ferraty, F. and Vieu, P. (2002), 'Functional nonparametric model and application to sprectrometric data', *Computational Statistics* **17**, 545–564.

Flores, B., Montoya, E., Sakschewski, B., Nascimento, N., Staal, A., Betts, R., Levis, C., Lapola, D., Esquivel Muelbert, A., Jakovac, C., Nobre, C., Oliveira, R., Borma, L., Nian, D., Boers, N., Hecht, S., ter Steege, H., Arieira, J., Leite Lucas, I. and Hirota, M. (2024), 'Critical transitions in the Amazon forest system', *Nature* **626**, 555–564.

Frévent, C., Ahmed, M.-S., Dabo-Niang, S. and Genin, M. (2023), 'Investigating spatial scan statistics for multivariate functional data', *Journal of the Royal Statistical Society Series C: Applied Statistics* **72**(2), 450–475.

Frévent, C., Ahmed, M.-S., Marbac, M. and Genin, M. (2021), 'Detecting spatial clusters in functional data: New scan statistic approaches', *Spatial Statistics* **46**, 100550.

Frévent, C., Ahmed, M.-S., Soula, J., Cucala, L., Smida, Z., Dabo-Niang, S. and Genin, M. (2022), 'The R package HDSpatialScan for the detection of clusters of multivariate and functional data using spatial scan statistics', *The R Journal* **14**, 95–120. https://rjournal.github.io/.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M. and Zhao, B. (2017), 'The modern-era retrospective analysis for research and applications, version 2 (merra-2)', *Journal of Climate* **30**(14), 5419–5454.

Hope, A. C. A. (1968), 'A simplified Monte Carlo significance test procedure', *Journal of the Royal Statistical Society. Series B (Methodological)* **30**(3), 582–598.

Horváth, L. and Kokoszka, P. (2012), *Detection of changes in the mean function*, Springer New York, New York, NY, pp. 79–104.

Horváth, L., Kokoszka, P. and Reeder, R. (2013), 'Estimation of the mean of functional time series and a two-sample problem', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **75**(1), 103–122.

Hotelling, H. (1931), 'The generalization of Student's ratio', *The Annals of Mathematical Statistics* **2**(3), 360–378.

Huang, L., Kulldorff, M. and Gregorio, D. (2007), 'A spatial scan statistic for survival data', *Biometrics* **63**, 109–18.

International Organization for Migration (2022), 'Dtm Nigeria - flood flash report - bauchi state (12 september 2022). iom, nigeria'.

Jacques, J. and Preda, C. (2013), 'Funclust: A curves clustering method using functional random variables density approximation', *Neurocomputing* **112**, 164–171. Advances in artificial neural networks, machine learning, and computational intelligence.

Johnson, R. and Wichern, D. (2002), *Applied multivariate statistical analysis*, 5. ed edn, Prentice Hall, Upper Saddle River, NJ.

Joint Research Center (2022), 'Pakistan monsoon rains and EU response'.

Joseph, E., Galeano San Miguel, P. and Lillo Rodríguez, R. E. (2015), Two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance, DES - Working Papers. Statistics and Econometrics. WS ws1503, Universidad Carlos III de Madrid.

Jung, I. and Cho, H. J. (2015), 'A nonparametric spatial scan statistic for continuous data', *International Journal of Health Geographics* **14**(30).

Karhunen, K. (1947), *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*, Annales Academiae Scientiarum Fennicae: Ser. A 1, Kirjapaino oy. sana.

Kulldorff, M. (1997), 'A spatial scan statistic', *Communications in Statistics - Theory and Methods* **26**, 1481–1496.

Kulldorff, M. (2006), 'Tests of spatial randomness adjusted for an inhomogeneity: A general framework', *Journal of the American Statistical Association* **101**, 1289–1305.

Kulldorff, M., Huang, L. and Konty, K. (2009), 'A scan statistic for continuous data based on the normal probability model', *International journal of health geographics* **8**(58).

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W., Kleinman, K. and Platt, R. (2007), 'Multivariate scan statistics for disease surveillance', *Statistics in medicine* **26**(8), 1824–1833.

Kulldorff, M. and Nagarwalla, N. (1995), 'Spatial disease clusters: Detection and inference', *Statistics in Medicine* **14**(8), 799–810.

Lawson, A. and Denison, D. (2002), *Spatial Cluster Modeling*, Chapman and Hall/CRC.

Lee, J., Sun, Y. and Chang, H. H. (2020), 'Spatial cluster detection of regression coefficients in a mixed-effects model', *Environmetrics* **31**(2).

Lévy, P. (1965), *Processus stochastiques et mouvement brownien*, 2nd edn, Gauthier-Villars, Paris.

Lima, M., Duczmal, L., Neto, J. and Pinto, L. (2014), 'Spatial scan statistics for models with overdispersion and inflated zeros', *Statistica Sinica* **25**.

Liu, H. and Houwing-Duistermaat, J. (2024), 'On estimation of covariance function for functional data with detection limits', *Journal of Nonparametric Statistics* **36**(3), 730–748.

Loader, C. R. (1991), 'Large-deviation approximations to the distribution of scan statistics', *Advances in Applied Probability* **23**(4), 751–771.

Loh, J. M. and Zhu, Z. (2007), 'Accounting for spatial correlation in the scan statistic', *The Annals of Applied Statistics* **1**(2), 560–584.

Mann, H. B. and Whitney, D. R. (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *The Annals of Mathematical Statistics* **18**(1), 50–60.

Nagarwalla, N. (1996), 'A scan statistic with a variable window', *Statistics in Medicine* **15**(7-9), 845–850.

Naus, J. (1963), *Clustering of Random Points in Line and Plane*, Harvard University.

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis*, Springer Series in Statistics, 2nd edn, Springer-Verlag New York.

Smida, Z., Cucala, L., Gannoun, A. and Durif, G. (2022*a*), 'A median test for functional data', *Journal of Nonparametric Statistics* **34**, 520–553.

Smida, Z., Cucala, L., Gannoun, A. and Durif, G. (2022*b*), 'A Wilcoxon-Mann-Whitney spatial scan statistic for functional data', *Computational Statistics & Data Analysis* **167**, 107378.

Student (1908), 'The probable error of a mean', *Biometrika* **6**(1), 1–25.

United Nations Human Settlements Programme (2023), 'UN-Habitat Pakistan country report 2023'.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics Bulletin* **1**(6), 80–83.

You, Y., Ting, M. and Biasutti, M. (2024), 'Climate warming contributes to the record-shattering 2022 Pakistan rainfall', *npj Climate and Atmospheric Science* **7**(89).

Zhang, Z., Assunção, R. and Kulldorff, M. (2010), 'Spatial scan statistics adjusted for multiple clusters', *Journal of Probability and Statistics* **2010**(1), 1–11.

Zhenhua Lin, M. E. L. and Müller, H.-G. (2021), 'High-dimensional manova via bootstrapping and its application to functional and sparse count data', *Journal of the American Statistical Association* **118**(541), 177–191.