A Wilcoxon-Mann-Whitney spatial scan statistic for functional data

Zaineb Smida, Lionel Cucala, Ali Gannoun, Ghislain Durif

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167947321002127
Manuscript_077e13b1bf96f137db2ca6ecd0e4485e

A Wilcoxon-Mann-Whitney spatial scan statistic for functional data

Zaineb SMIDA*, Lionel CUCALA, Ali GANNOUN, Ghislain DURIF

Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France.

Abstract

A nonparametric scan method for functional data indexed in space is introduced. The associated scan statistic is derived from the Wilcoxon-Mann-Whitney test statistic defined for infinite dimensional data. It is completely nonparametric as it does not assume any distribution concerning the functional marks. Whatever the clustering scenario, this scan test seems to be efficient to detect and locate the cluster. This method is applied to a data set for extracting features in Spanish province population growth. A significant spatial cluster of low demographic evolution rates is found, exhibiting a specific phenomenon in the North-West of Spain.

Keywords: Cluster detection, Functional data, Hilbert space, Spatial Scan statistic, Wilcoxon-Mann-Whitney test.
2000 MSC: 46E20, 62G10, 62H11, 62R10.

1. Introduction

Spatial cluster detection has become a fruitful area of statistics that has particulary expanded in recent decades. It is used to identify aggregations of events in a specific area, see Lawson and Denison (2002) for a thorough review. One of the most popular cluster detection technique is the scan statistic which was firstly introduced by Naus (1963). It was defined as the maximum number of events observed within a window with constant size,

*Corresponding author
Email addresses: zaineb.smida@umontpellier.fr (Zaineb SMIDA), lionel.cucala@umontpellier.fr (Lionel CUCALA), ali.gannoun@umontpellier.fr (Ali GANNOUN), ghislain.durif@umontpellier.fr (Ghislain DURIF)

known as the scanning window, as it moves continuously over the studied region. Knowing the distributions of these scan statistics (Alm, 1997) helps to decide whether exceptional or not observing a cluster of events. The field of spatial scan statistics was highly enhanced by the article written by Kulldorff (1997): he proposed scanning the study area with variable size circular windows and selecting the most likely cluster as the one maximizing a likelihood ratio test. He used either Bernoulli or Poisson model and estimated the clusters' statistical significance via a Monte-Carlo procedure. These innovations gave birth to several works in which researchers adapted the spatial scan statistics to different types of data, using different probability models: exponential (Huang et al., 2007), normal (Kulldorff et al., 2009), multivariate Gaussian (Cucala et al., 2017), etc.

All these spatial scan methods are developed for univariate and multivariate data indexed in space. However, the development of sensoring and computing tools brings more and more access to data of functional type coming from various fields of applications such as environmetrics, biometrics, medecine and econometrics (Ramsay and Silverman, 2005). These data are not real random variables or vectors but they are a sample of random curves where each element is considered as a function. Moreover, these functional data are often indexed in space (Delicado et al., 2010) and, even if a few studies have been conducted on modelling (Cronie et al., 2019) or clustering (Gaetan et al., 2017) such data, to our knowledge, there is no spatial cluster detection method designed for this kind of data yet.

In the present work, we develop a scan statistic for functional data indexed in space: thanks to this statistic, we are able to detect spatial clusters in which the observations of a functional random variable are different than elsewhere, and also to compute the significance of these differences. Since no likelihood is associated with functional random variables (Ferraty , 2011), maximising a likelihood ratio test is not possible here. Thus, we follow the idea by Cucala (2017) that any test for equality of two distributions can give birth to a scan statistic.

The rest of this paper is organized as follows. In section 2, we build a non-parametric spatial scan statistic for functional data based on the Wilcoxon-Mann-Whitney statistic proposed by Chakraborty and Chaudhuri (2015) and we evaluate its statistical significance using random permutations. In section 3, first, the spatial scan statistic is compared to other methods on simulated datasets. Then, we apply it to a real dataset illustrating the demographic evolution over time in Spanish provinces and we exhibit a specific behaviour

in the North-West of Spain in the last twenty-two years. We conclude with a discussion and a brief scope for future work.

2. A nonparametric spatial scan statistic for functional data

2.1. Introducing the statistic

Consider a random variable X taking values in a functional space χ. For sake of simplicity, we will suppose that χ is an Hilbert space such as $L^2([0,1],\mathbb{R})$. 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 (Demattei 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. We remark that, since i might be equal to j, the number of potential clusters is n^2.

Following the initial work by Kulldorff (1997), the spatial scan statistics designed for univariate or multivariate marks are most often based on a concentration index derived from a likelihood ratio. This likelihood ratio relies on assuming a specific probability distribution for the marks and testing the null hypothesis H_0 (absence of a cluster) against an alternative one $H_{1,Z}$ (presence of a cluster in Z) for every potential cluster $Z \in \mathcal{S}$. However, for functional random variables, even if approximations have been proposed

(Jacques and Preda, 2013), the notion of probability density generally does not exist. Thus, our clustering index will rely on a nonparametric test for equality of distributions, as proposed by Jung and Cho (2015) and Cucala et al. (2019) in the univariate and multivariate settings respectively.
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 $n_{Z^c} = n - n_Z$. Assume that the marks in Z and Z^c respectively follow probability measures P_Z and P_{Z^c} on χ. We suppose that P_Z and P_{Z^c} differ by a shift $\Delta_Z \in \chi$. For testing the hypothesis $H_0 : \Delta_Z = 0$ (equality of distributions) against $H_{1,Z} : \Delta_Z \neq 0$, a Wilcoxon-Mann-Whitney test statistic in such space is defined by Chakraborty and Chaudhuri (2015) as:

$$T_{\mathrm{WMW}}(Z) = \frac{1}{n_Z n_{Z^c}} \sum_{\{i:s_i \in Z\}} \sum_{\{j:s_j \in Z^c\}} \frac{X_j - X_i}{\|X_j - X_i\|_\chi},$$

where $\|.\|_\chi$ stands for a norm on χ. Remark that this statistic is a natural extension to the functional setting of the well-known statistics introduced by Wilcoxon (1945) and Mann and Whitney (1947) in the univariate setting, since every element of the first sample is compared to every element of the second one. The statistic $T_{\mathrm{WMW}}(Z)$, taking values in χ, cannot be used directly as a concentration index since its distribution highly depends on n_Z, the size of the potential cluster Z. Thus, as recomended by Cucala (2017), we introduce the standardized concentration index

$$\mathrm{U}(Z) := \sqrt{\frac{n_z n_{Z^c}}{n}} T_{\mathrm{WMW}}(Z)$$

which is designed to compare potential clusters having different population sizes, as claimed by the following lemma.

Lemma 1. *The null limiting distribution of $U(Z)$ is the same for any potential cluster $Z \in \mathcal{S}$.*

The proof comes directly from the convergence theorem by Chakraborty and Chaudhuri (2015) stating that, under H_0, if $n_Z/n \to \gamma \in [0,1]$ as $n_Z, n_{Z^c} \to \infty$,

$$(n_Z n_{Z^c}/n)^{1/2} T_{\mathrm{WMW}}(Z) \text{ converges weakly to } G(0,\Gamma), \tag{1}$$

where $G(m, C)$ is the distribution of a Gaussian random element in χ with mean $m \in \chi$ and covariance C. Since the covariance operator Γ does not depend on n_Z and n_{Z^c}, the result holds. □

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

$$\Lambda_{\text{WMWFSS}} = \max_{Z \in \mathcal{S}} \|\mathrm{U}(Z)\|_\chi$$

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

$$\hat{\mathrm{C}} = \arg\max_{Z \in \mathcal{S}} \|\mathrm{U}(Z)\|_\chi.$$

This latter is called the most likely cluster.

2.2. Computing the scan statistic

- The computation of the scan statistic Λ_{WMWFSS} involves the computation of the concentration index $U(Z)$ for every potential cluster $Z \in \mathcal{S}$ and, since this index is issued from a sum of $n_Z \times n_{Z^c}$ terms, a naive computation can be very time-consuming. However, here are two computational tricks to address this problem:
  - all concentration indices $U(Z)$, for every $Z \in \mathcal{S}$, rely on the computation of
    $$R_{i,j} = \frac{X_j - X_i}{\|X_j - X_i\|_\chi}$$
    for every $1 \leq i < j \leq n$. Thus, these $(n-1)^2/2$ terms must be calculated at the very beginning of the process and stored. Remark that the computation of the $R_{i,j}$'s will be different whether the functional marks $X_1, \ldots, X_n$ are known explicitly (for example by their decomposition on a certain basis) or only partially observed. See Ramsay and Silverman (2005) for more details.
  - In order to optimize the computation process, we decided to calculate the indices $U(Z)$ in a very specific order. Here is an example: let Z and Z' be any potential clusters such that $Z' = Z \cup s_k$.

Then, the concentration index for Z' can be obtained from

$$\begin{aligned}
(n_{Z'} n_{Z'^c} n)^{1/2} \mathrm{U}(Z') &= \sum_{\{i:s_i \in Z'\}} \sum_{\{j:s_j \in Z'^c\}} \frac{X_j - X_i}{\|X_j - X_i\|_\chi} \\
&= \sum_{\{i:s_i \in Z\}} \sum_{\{j:s_j \in Z^c\}} \frac{X_j - X_i}{\|X_j - X_i\|_\chi} \\
&+ \sum_{\{j:s_j \in Z^c\}} \frac{X_j - X_k}{\|X_j - X_k\|_\chi} - \sum_{\{i:s_i \in Z\}} \frac{X_k - X_i}{\|X_k - X_i\|_\chi} \\
&= (n_Z n_{Z^c} n)^{1/2} \mathrm{U}(Z) \\
&+ \sum_{\{j:s_j \in Z^c\}} \frac{X_j - X_k}{\|X_j - X_k\|_\chi} - \sum_{\{i:s_i \in Z\}} \frac{X_k - X_i}{\|X_k - X_i\|_\chi}
\end{aligned}$$

This set-up requires to iterate over only n elements instead of $(n_Z - 1) \times (n - n_Z + 1)$ and dramatically decreases the computational cost.

- When $\chi = L^2([a, b], \mathbb{R})$ where $a, b \in \mathbb{R}$ and $a < b$, the algorithm used to derive the WMWFSS and its associated most likely cluster $\hat{\mathrm{C}}$ is as follows:

Algorithm 1 Computing the WMWFSS and the most likely cluster MLC

Data: $\{(s_1, X_1), \ldots, (s_n, X_n)\}$

For all $i, j \in \{1, \ldots, n\}$ compute

$$R_{i,j} = \frac{X_j - X_i}{\|X_j - X_i\|_{L^2}}$$

and let $R = \{R_{i,j}\}_{i,j \in \{1,\ldots,n\}}$.

For all $i, j \in \{1, \ldots, n\}$ compute the distance $d_{i,j}$ between locations s_i and s_j and let $d = \{d_{i,j}\}_{i,j \in \{1,\ldots,n\}}$.

function TWMW (computing the WMW test statistic)

  Input: R, $A \subseteq \{1, \cdots, n\}$, $B \subseteq \{1, \cdots, n\}$

  Output: $\text{WMW} \in \mathbb{R}^+$

  Initialization: $\text{WMW} = 0$

  for $i \in A$ **do**

    for $j \in B$ **do**

      $\text{WMW} = \text{WMW} + R_{i,j}$

  $\text{WMW} = \frac{\text{WMW}}{\text{length}(A)\text{length}(B)}$

function ORDER

  Input: $v \in \mathbb{R}^n$

  Output: $p =$ permutation vector of $(1, \ldots, n)$

  for $k = 1$ to n **do**

    $p_k =$ order of value v_k in v following ascending order

function WMWFSS (computing the WMWFSS scan statistic)

  Input: R, d

  Output: $\tilde{c}$ (WMWFSS value), MLC (most likely cluster)

  Initialization: $\tilde{c} = -\infty$, $\tilde{i} = 0$ and $\tilde{j} = 0$.

  for $i = 1$ to n **do**

    $O = \text{ORDER}(\{d_{i,j}\}_{j \in \{1,\ldots,n\}})$

    for $j = 1$ to $(n-1)$ **do**

      $v_{in} = \{O_k\}_{k \in \{1,\ldots,j\}}$ and $v_{out} = \{O_k\}_{k \in \{j+1,\ldots,n\}}$

      $c = \sqrt{\frac{j(n-j)}{n}} \times \|\text{TWMW}(R, v_{in}, v_{out})\|_{L^2}$

      if $c > \tilde{c}$ **then**

        $\tilde{c} = c$, $\tilde{i} = i$, $\tilde{j} = j$

  $\tilde{O} = \text{ORDER}(\{d_{\tilde{i},j}\}_{j \in \{1,\ldots,n\}})$

  $\text{MLC} = \{\tilde{O}_k\}_{k \in \{1,\ldots,\tilde{j}\}}$

2.3. Computing the statistical significance

After computing the scan statistic Λ_{WMWFSS} and the most likely cluster $\hat{C}$, it is necessary to evaluate its significance. However, the distribution, under H_0, of a variable window scan statistic has no analytical form. To overcome this problem, Dwass (1957) proposed a test procedure based on Monte-Carlo simulations allowing to give an approximation of the null distribution. This method was subsequently extended by Bernard (1963) and Hope (1968). It relies on comparing the observed scan statistic to scan statistics issued from datasets simulated under H_0. Here, since no assumption is made on the distribution of the functional marks, the only way to obtain such datasets is by running a method called random labelling (Cucala, 2014): a simulated dataset is obtained by randomly associating the functional marks X_i to the spatial locations s_i. Based on T random permutations, let

$$\Lambda^{(1)}_{\text{WMWFSS}}, \ldots, \Lambda^{(T)}_{\text{WMWFSS}}$$

be the observations of the scan statistics associated with the simulated datasets. Then, as stated by Dwass (1957), the p-value of the scan statistic Λ_{WMWFSS}, observed in the initial sample, is given by

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

Of course, the larger the number of permutations T, the better the estimation of the p-value of the scan statistic. However, since the computational cost cannot be neglected, one needs to find a trade-off between the two aspects. The most likely cluster $\hat{C}$ is said to be significant if p_{value} is less than the type I error α.

3. Applications

3.1. Simulation study

We decided to run a simulation study to evaluate the performance of the functional scan statistic Λ_{WMWFSS} proposed in the previous section. We generated artificial datasets using the geographic locations of the administrative centers of the 94 french administrative areas named as *départements*. The

simulated true cluster, denoted by C, is defined as a set of *départements* in the Parisian area according to two configurations: (i) 8 *départements* and (ii) 10 *départements*. Maps of the simulated clusters are given in Fig. 1.

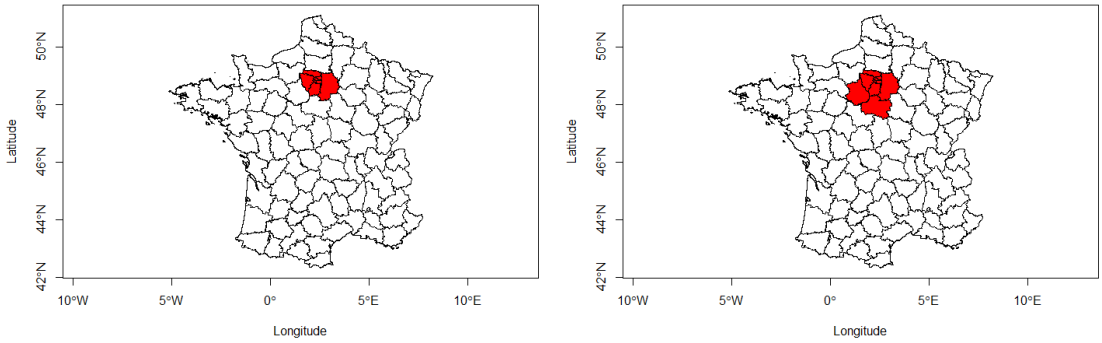

Figure 1: The 94 French *départements*. In red: simulated clusters (8 and 10 *départements*).

The functional marks associated with each location take values in $\chi = L^2([0,1],\mathbb{R})$ and are defined as follows:

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

where for all $k \geqslant 1$, $e_k(t) = \sqrt{2}\sin(t/\sigma_k)$ is an orthonormal basis of χ, $\sigma_k = ((k-0.5)\pi)^{-1}$ and $Z_{i,k}$'s are independent random variables which correspond to the projection of X_i on the Karhunen-Loève basis (Karhunen, 1947; Lévy and Loève, 1948). The decomposition of the functional marks above is based in the Karhunen-Loève expansion which is widely used in several issues related to image processing and functional data analysis (Ahmed et al., 2017).

We have investigated two different cases, namely a standard Brownian motion (sBm) process: $Z_{i,k}/\sigma_k$ having a $\mathcal{N}(0,1)$ distribution and a centered Student-t process with five degrees of freedom: $Z_{i,k}/\sigma_k$ having a $t(5)$ distribution.

The probability measures of the functional marks inside and outside the

cluster C differ by a shift Δ. Three types of shifts are studied: $\Delta_1(t) = ct$, $\Delta_2(t) = ct(1-t)$ and $\Delta_3(t) = c\sin(2\pi t)$, $c > 0$ for all $t \in [0,1]$. The parameter c is called the cluster intensity: remark that, since the functional marks are independent, this parameter totally controls their level of spatial heterogeneity. Different values of this parameter were considered for each Δ. The range of Δ_2 being smaller than the ranges of Δ_1 and Δ_3, it is combined with larger values of c.

Since, as already said in the Introduction, we do not know any other cluster detection method dedicated to functional data indexed in space, we decided to compare the Wilcoxon-Mann-Whitney functional scan statistic to two univariate scan statistics applied to summaries of the functional marks:

- the first scan method relies on the mean values of the marks

$$\bar{X}_i = \int_0^1 X_i(t)dt, \quad i = 1, \ldots, n.$$

  Each mean value is associated with its location and the univariate Wilcoxon-Mann-Whitney scan statistic introduced by Cucala (2016) is computed, using the same set of potential clusters and same random permutations than the functional one. This mean-based univariate scan statistic is denoted by Λ_{MBUSS}.

- the second one, inspired from the LISA function defined by Mateu et al. (2007), relies on the deviations from the mean function of the marks

$$D_i = \int_0^1 (X_i(t) - \bar{X}(t))^2 dt, \quad i = 1, \ldots, n,$$

  where

$$\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)$$

  is the mean function of the observed functional marks. Each deviation is associated with its location and the univariate Wilcoxon-Mann-Whitney scan statistic is computed. This deviation-based univariate scan statistic is denoted by Λ_{DBUSS}.

To compare the three scan methods, we generated 100 simulated datasets for each distribution of the marks and each value of the cluster intensity c

and we computed three distinct criteria for each method: the alarm rate (AR), the True Positive (TP) rate (also called the sensitivity) and the False Positive (FP) rate. These three criteria were calculated as follows:

- The alarm rate (AR) was defined as the proportion of datasets exhibiting a significant cluster with a type I error equal to 0.05 and based on T = 99 random permutations.

- The TP rate, denoted by %TP, was defined as the mean proportion of the True Positive (TP) *départements* over all simulated datasets. It was calculated as the number of *départements* included both in the significant cluster $\hat{C}$ and in the true cluster C divided by the number of *départements* included in C.

- The calculation of the FP rate, denoted by %FP, is similar to the TP one. It was defined as the average proportion of the False Positive (FP) *départements* i.e, the number of *départements* included in the most significant cluster $\hat{\mathrm{C}}$ but not in the true cluster C divided by the number of *départements* not included in C.

The whole results of this simulation study are given in Appendix B but they are summarized in Table 1 and Table 2 below.

Table 1: Simulation study–AR, %TP and %FP results of Λ_{WMWFSS}, Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_1 = ct$, $\Delta_2 = ct(1-t)$ and $\Delta_3 = c\sin(2\pi t)$ using two distributions: Normal and Student-t. The true cluster contains 8 *départements*. Bold values indicate the best performance in each line.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
0.0	AR	**0.070**	0.050	0.060	0.0	AR	**0.060**	**0.060**	0.020
	%TP	0.554	0.875	**1.000**		%TP	0.479	0.479	**1.000**
	%FP	**0.442**	0.553	0.589		%FP	**0.444**	0.446	0.645
$\Delta_1(t) = ct$									
1.5	AR	**0.380**	0.340	0.150	1.5	AR	**0.240**	0.220	0.090
	%TP	0.908	0.882	**1.000**		%TP	0.885	0.847	**1.000**
	%FP	**0.110**	0.165	0.148		%FP	**0.098**	0.164	0.472
2.0	AR	**0.730**	0.660	0.300	2.0	AR	**0.600**	0.510	0.180
	%TP	**0.967**	0.966	0.933		%TP	0.935	0.939	**0.993**
	%FP	**0.049**	0.087	0.073		%FP	**0.095**	0.142	0.228
2.5	AR	**0.920**	0.890	0.570	2.5	AR	**0.790**	0.730	0.390
	%TP	**0.978**	0.961	0.879		%TP	0.949	0.938	**0.978**
	%FP	**0.056**	0.070	0.077		%FP	**0.045**	0.063	0.137
$\Delta_2 = ct(1-t)$									
4.5	AR	**0.460**	0.320	0.160	4.5	AR	**0.360**	0.310	0.130
	%TP	0.853	0.844	**0.938**		%TP	0.760	0.706	**0.904**
	%FP	**0.101**	0.139	0.262		%FP	**0.121**	0.144	0.286
5.5	AR	**0.700**	0.530	0.260	5.5	AR	**0.450**	0.380	0.150
	%TP	**0.950**	0.934	0.923		%TP	0.908	0.898	**1.000**
	%FP	**0.042**	0.077	0.178		%FP	**0.070**	0.130	0.261
6.5	AR	**0.870**	0.760	0.460	6.5	AR	**0.610**	0.470	0.200
	%TP	**0.991**	0.984	0.929		%TP	0.932	0.910	**0.988**
	%FP	**0.041**	0.068	0.091		%FP	**0.067**	0.097	0.153
$\Delta_3(t) = c\sin(2\pi t)$									
1.0	AR	**0.310**	0.080	0.170	1.0	AR	**0.170**	0.070	0.140
	%TP	**0.895**	0.531	0.882		%TP	0.772	0.571	**0.938**
	%FP	**0.156**	0.552	0.347		%FP	**0.126**	0.150	0.273
1.25	AR	**0.660**	0.040	0.350	1.25	AR	**0.390**	0.060	0.200
	%TP	**0.981**	0.781	0.979		%TP	**0.949**	0.667	0.938
	%FP	**0.037**	0.573	0.250		%FP	**0.109**	0.455	0.251
1.5	AR	**0.960**	0.060	0.660	1.5	AR	**0.820**	0.050	0.310
	%TP	**0.988**	0.833	0.981		%TP	**0.970**	0.425	0.960
	%FP	**0.010**	0.271	0.071		%FP	**0.053**	0.490	0.199

Table 2: Simulation study–AR, %TP and %FP results of Λ_{WMWFSS}, Λ_{MBUSS} and Λ_{DBUSS} when Δ_1, Δ_2 and Δ_3 using two distributions: Normal and Student-t. The true cluster contains 10 *départements*. Bold values indicate the best performance in each line.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
0.0	AR	**0.060**	0.050	0.030	0.0	AR	**0.060**	**0.060**	0.030
	%TP	0.317	0.480	**1.000**		%TP	0.200	0.550	**0.667**
	%FP	**0.534**	0.545	0.635		%FP	**0.204**	0.206	0.544
$\Delta_1(t) = ct$									
1.5	AR	**0.650**	0.540	0.160	1.5	AR	**0.420**	0.360	0.120
	%TP	0.926	0.913	**0.963**		%TP	0.883	0.872	**1.000**
	%FP	**0.096**	0.156	0.263		%FP	**0.136**	0.147	0.313
2.0	AR	**0.900**	0.860	0.400	2.0	AR	**0.710**	0.660	0.200
	%TP	**0.960**	0.956	0.913		%TP	**0.956**	0.933	0.950
	%FP	**0.051**	0.076	0.119		%FP	**0.074**	0.091	0.123
2.5	AR	**1.000**	0.980	0.600	2.5	AR	**0.960**	0.910	0.400
	%TP	**0.988**	0.979	0.963		%TP	**0.980**	0.968	0.950
	%FP	**0.029**	0.041	0.083		%FP	**0.040**	0.049	0.085
$\Delta_2 = ct(1-t)$									
4.5	AR	**0.660**	0.520	0.170	4.5	AR	**0.420**	0.380	0.120
	%TP	**0.950**	0.933	0.918		%TP	0.874	0.871	**0.942**
	%FP	**0.090**	0.139	0.218		%FP	**0.118**	0.129	0.289
5.5	AR	**0.930**	0.740	0.230	5.5	AR	**0.650**	0.530	0.170
	%TP	0.974	0.972	**1.000**		%TP	0.906	0.898	**0.994**
	%FP	**0.040**	0.051	0.139		%FP	**0.058**	0.097	0.345
6.5	AR	**0.990**	0.900	0.600	6.5	AR	**0.900**	0.820	0.360
	%TP	**0.980**	0.973	0.960		%TP	0.956	0.952	**0.975**
	%FP	**0.026**	0.049	0.091		%FP	**0.035**	0.055	0.162
$\Delta_3(t) = c\sin(2\pi t)$									
1.0	AR	**0.690**	0.070	0.250	1.0	AR	**0.340**	0.050	0.120
	%TP	**0.948**	0.757	0.912		%TP	0.953	0.800	**1.000**
	%FP	**0.052**	0.388	0.259		%FP	**0.096**	0.429	0.348
1.25	AR	**0.960**	0.040	0.480	1.25	AR	**0.760**	0.020	0.230
	%TP	**0.993**	0.950	0.975		%TP	**0.963**	0.500	0.874
	%FP	**0.015**	0.393	0.143		%FP	**0.042**	0.369	0.151
1.5	AR	**1.000**	0.060	0.720	1.5	AR	**0.950**	0.100	0.350
	%TP	**1.000**	0.983	0.994		%TP	**0.984**	0.840	0.971
	%FP	**0.004**	0.274	0.056		%FP	**0.014**	0.411	0.148

From Table 1 and Table 2, the sizes of the different methods (i.e. the alarm rates when c=0) are close to the correct type I error which is equal to 0.05, regardless of the distribution of the marks.
As expected, the performances of all scan statistics tend to increase with high cluster intensity c and we can remark that the alarm rate of Λ_{WMWFSS} is higher than Λ_{MBUSS} and Λ_{DBUSS} in all different cases: this is expected as the first one relies on the whole information of the curves, the second one is only based on their mean value and the third one is derived from the distances between each curve and the mean curve. It should be noted that:

- When c increases, the alarm rate of all scan methods increases whatever the shift Δ and the size of the true cluster C. However, when the process is Student-t distributed, the alarm rate increases more slowly than when it is normally distributed. This difference can be explained by the fact that the Student-t distribution is more heavy-tailed than the Gaussian one. The relation between the alarm rate and the cluster intensity c seems to be the following: the alarm rate slowly increases when c is small but then, when c reaches a certain threshold, the slope gets steeper and the alarm rate very quickly gets close to 1. Since this threshold is different depending on the distribution of the functional marks, the discrepancy between the alarm rates of Normal and Student-t distributions is far from being constant. Remark also that, for equal values of the cluster intensity c, the alarm rate is larger when the size of the cluster goes from 8 to 10: it is always easier to detect a larger cluster.
  The difference in alarm rates between Λ_{WMWFSS} and Λ_{MBUSS} is slight when the shift between the marks inside and outside the cluster is linear, but it increases when this shift is quadratic and moreover when it is sinusoidal (see Table B.7 and Table B.10 in Appendix B): we can see that the alarm rate of Λ_{MBUSS} does not exceed 10% (is close to the nominal level 5%) using Δ_3 whatever the size of the true cluster and the distribution of the processes, since the sinusoidal shift has absolutely no consequence on the mean value of the process. The deviation-based scan statistic Λ_{DBUSS} is more adapted to this sinusoidal shift but still displays lower alarm rates than Λ_{WMWFSS}.

- The true positive and false positive rates also improve when the cluster intensity c increases (increasing for %TP and decreasing for %FP). As for the alarm rate, the recovering of the location of the cluster is harder

when the process is Student-t distributed than normally distributed but the size of the cluster has no great impact on %TP and %FP.
The whole information included in the functional marks is as useful for detecting the presence of a cluster than for recovering its exact location. Thus, unsurprisingly, the %TP and %FP rates obtained by the functional method $\Lambda_{\mathrm{WMWFSS}}$ are globally better than the ones obtained by the univariate methods. The difference is more obvious concerning the false positive rates: more often than the functional one, the univariate methods tend to exhibit clusters larger than the true cluster C.

3.2. Application to real data

Here, we give an example of the use of our scan statistic to extract features in Spanish province population growth, as presented by Cronie et al. (2019). In order to study the structure of the Spanish population, we considered one of the most important population characteristics which is the demographic evolution. This latter can change over time because of factors like birth and death rates, immigration rate or economical situations. The Spanish province population is provided by the *Spanish Institute of Statistics* (`www.ine.es`) and the boundary and centre coordinates of the 47 provinces of Spain (see Fig. 2) by the `R` package `raster` (Hijmans, 2019). For geographical reasons, we decided to exclude from the study *Baleares* and *Canarias* islands as well as the Spanish autonomous cities (*Melilla* and *Ceuta*) which are located on the Northwest coast of Africa and sharing a border with Morocco. To each point (centre) i, for $i = 1, \ldots, 47$, we associated the functional mark X_i, i.e. the demographic evolution over time, for 22 distinct years starting from 1998 to 2019 (see Fig. 3). The demographic evolution in each province was defined as the total population over the years 1998 to 2019 divided by the total population in 1998.

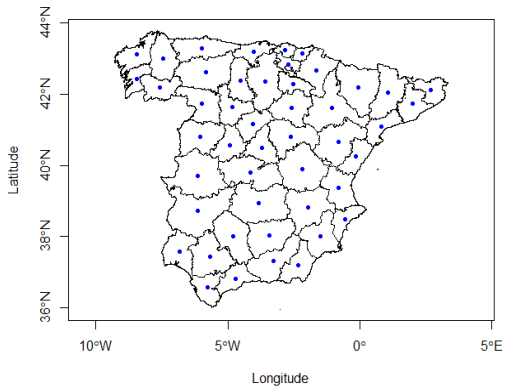

Figure 2: The 47 Spanish provinces and their geometrical centres.

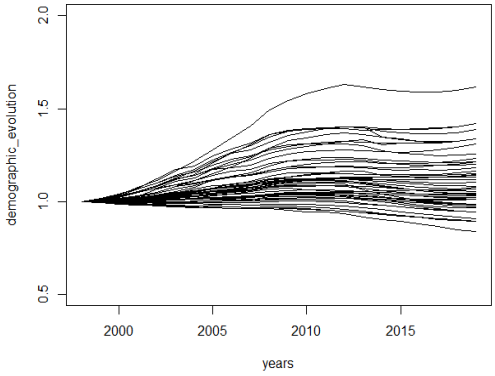

Figure 3: Demographic evolution in the 47 provinces from 1998 to 2019.

3.2.1. Analysis of the real dataset

Our objective here is to detect a spatial area where the demographic evolution would be significantly different. In order to identify such a cluster, we computed the functional scan statistic on this dataset: $\Lambda_{\mathrm{WMWFSS}} = 2.72025$. Remark that here the computation of the scan statistic is slightly different from what is done in the simulation study since it is estimated from 22 observation points. Based on $T = 999$ permutations, the value of the statistic is highly significant ($p_{value} = 0.001$) and the most likely cluster $\hat{C}$ is plotted in Fig. 4.

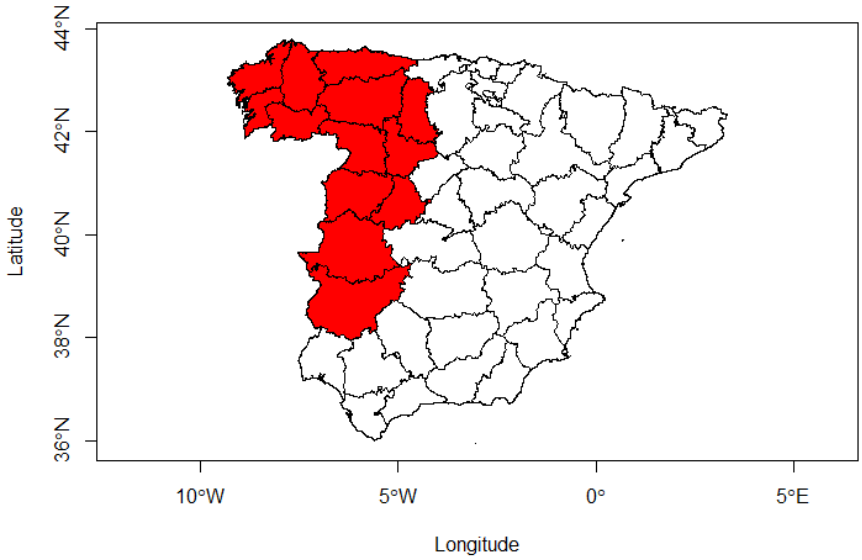

Figure 4: The most likely cluster detected by the functional scan statistic Λ_{WMWFSS}.

This cluster includes 13 provinces in the west of Spain (*Asturias*, *Galicia*, *Extremadura* and the west of *Castilla y León*) in which the marks are significantly lower than in the rest of the observation domain. In the west part of *Castilla y Leòn*, the most likely cluster includes the *región leonesa* and the west of the *Castilla la Vieja* (*Ávila*, *Palencia* and *Valladolid*).
We can see the demographic evolution curves associated with the most likely cluster in Fig. 5.

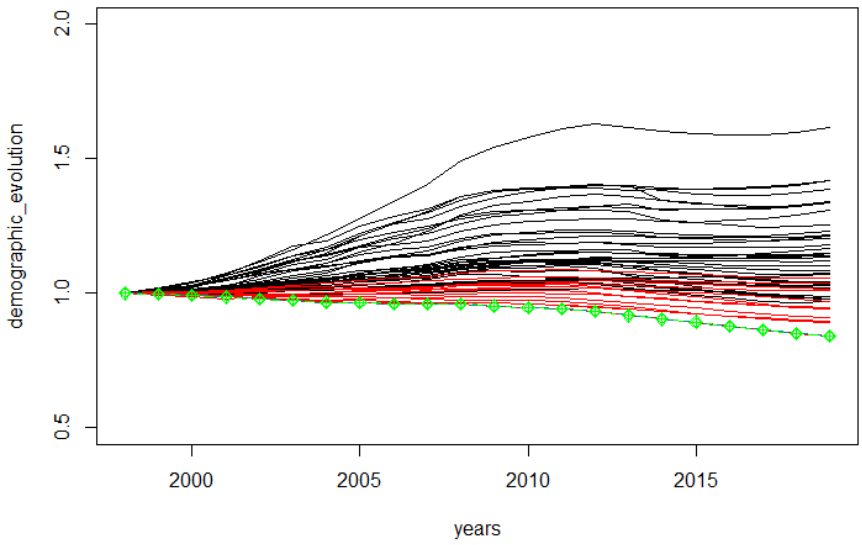

Figure 5: The demographic evolution curves (from 1998 to 2019) in each province are presented. Curves in red correspond to provinces inside the cluster, curves in black correspond to provinces outside the cluster and the curve in green corresponds to *Zamora* which is inside the cluster too.

We can see that this cluster includes the provinces which have the lowest demographic evolution compared to the rest of Spain. This can be explained by the increase in mortality rate and the decrease in birth rate in these regions. Between years 2006 and 2018, according to the *Spanish Institute of Statistics*, the 4 autonomous communities detected in the cluster are the territories which have the lowest birth rates (per 1000 inhabitants) compared to the other autonomous communities in Spain. In particular, the last 2 provinces with the lowest birth rate (per 1000 inhabitants) are *Ourense* (6.12 in 2006 and 4.82 in 2018) and *Zamora* (6.08 in 2006 and 5.13 in 2018). Moreover, the mortality rate (per 1000 inhabitants) is higher in the provinces belonging to the detected cluster and in particular *Zamora* has the highest mortality rate (12.46 in 2006 and 15.75 in 2018). This explains why *Zamora* has the lowest evolution demographic (see Fig. 5) and is close to becoming a demographic desert. Such a demographic decrease can be explained by the emigration in the last years of the youngest population abroad and to

other regions of Spain like *Cataluña* and *Madrid* where the average hourly wage is higher and the unemployment rate is lower than the autonomous communities detected by the functional scan statistic (for more details, see the website of the *Spanish Institute of Statistics*).

Then, using the same dataset, we also computed the univariate scan statistics Λ_{MBUSS} and Λ_{DBUSS} and their p-values and we recorded the computation time. The results are given in Table 3.

Table 3: The p-values and computation time (in seconds) of the different scan methods using different number of permutations.

	T=99		**T=999**	
Method	p_{value}	time	p_{value}	time
Λ_{WMWFSS}	0.01	1.14	0.001	10.92
Λ_{MBUSS}	0.01	0.62	0.003	7.20
Λ_{DBUSS}	0.56	0.34	0.527	6.83

Contrary to the deviation-based univariate scan statistic, the mean-based univariate scan statistic detects a very significant cluster which is very similar to the one detected by Λ_{WMWFSS} (see Fig. 6). This is not surprising since the main difference between the curves inside this cluster and the curves outside is their mean level rather than their shape. Moreover, the cluster detected by Λ_{MBUSS} is larger than the one detected by Λ_{WMWFSS}. As said in the simulation study, contrary to the functional scan method, the univariate scan methods tend to exhibit larger clusters than the true one, as noticed by the %FP rate. Thus, we believe that the most likely cluster detected by our functional scan method, based on the analysis of the curves on the whole time period, should be investigated first.

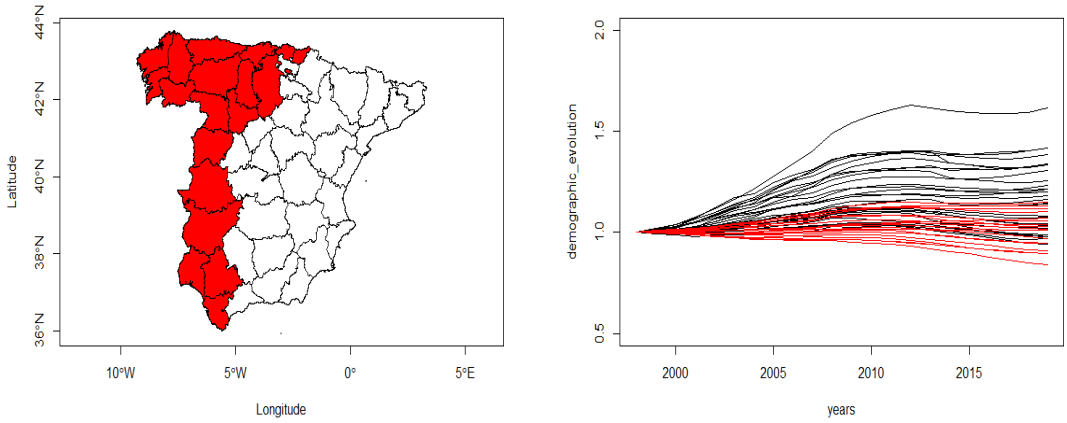

Figure 6: The most likely cluster detected by Λ_{MBUSS} and demographic evolution curves associated.

Concerning the computation time, we remark that the functional scan statistic, even if it takes advantage of the whole information of the data, is not that time-consuming compared to the univariate ones. This performance was achieved thanks to the use of the fonction `NPFSS` from the `R` package `HDSpatialScan` introduced very recently by Frévent et al. (2021).

Finally, after identifying the most likely cluster, we have tested the presence of a secondary cluster, following the method by Zhang et al. (2010): once a significant cluster is found, remove the data included in that cluster and restart the analysis. However, on this dataset and using the functional scan statistic, the secondary cluster is not significant since its p-value equals 0.282, using $T = 999$ permutations.

3.2.2. Analysis of the sensitivity of the method

Secondly, we decided to add noise to the preceding real dataset in order to test the sensitivity of the proposed method. We also investigated the choice of the number of permutations. We considered the noisy data

$$X_i'(t) = X_i(t) + \alpha\epsilon_i(t), \quad \forall i = 1, \ldots, 47, \quad \forall t \in [1998, 2019],$$

where the X_i's are the initial functional marks (the demographic evolution of the Spanish population measured in each province), the ϵ_i's are independent

centered sBm processes and α is the parameter controlling the variance of the added noise. We simulated 100 noisy datasets with different levels of variance α and computed the functional scan statistic and its p-value based on different numbers of random permutations T. As in section 3.1, Table 4 presents the alarm rates, TP and FP rates we obtained. In this case, the TP and FP rates are not computed based on the true cluster (which is unknown) but on the most likely cluster obtained without noise (see Fig. 4).

Table 4: Real data plus noise –Alarm rate, %TP and %FP results of the functional scan statistic Λ_{WMWFSS} for different variance level α and number of permutations T.

α		**T=29**	**T=59**	**T=99**	**T=999**
0.05	AR	0.990	1.000	1.000	1.000
	%TP	1.000	1.000	1.000	1.000
	%FP	0.000	0.000	0.000	0.000
0.1	AR	0.980	1.000	1.000	1.000
	%TP	0.998	0.988	0.984	0.985
	%FP	0.013	0.031	0.038	0.035
0.15	AR	0.940	1.000	1.000	1.000
	%TP	0.980	0.948	0.949	0.943
	%FP	0.096	0.100	0.130	0.129
0.2	AR	0.870	0.970	0.990	0.990
	%TP	0.893	0.898	0.904	0.890
	%FP	0.229	0.248	0.266	0.266
0.25	AR	0.560	0.630	0.680	0.690
	%TP	0.771	0.866	0.769	0.817
	%FP	0.380	0.350	0.407	0.383
0.3	AR	0.220	0.270	0.290	0.290
	%TP	0.664	0.795	0.618	0.695
	%FP	0.418	0.416	0.444	0.449

As expected, when the level of noise added to the initial data increases, the alarm rate of the test decreases since the presence of a significant cluster becomes less and less obvious. Moreover, for moderate level of noise, the clusters detected are not that different from the most likely cluster obtained without noise (the TP rate is close to 1 and the FP rate close to 0) but there is an evolution when α increases. It seems that, as already described by McDonough and Whalen (1995), when the noise level α is small, the signal

to noise ratio is large enough so that the scan method still works. However, when α reaches a certain threshold around 0.25, the signal to noise ratio becomes too small and the scan method fails. On the other hand, we can remark that the influence of the number of permutations T is quite limited: we may just mention that choosing $T = 29$ random permutations leads to less accurate p-values, so that the alarm rate obtained with that value of T might be slightly different from the others.

4. Discussion

Nowadays and with the development of modern technology, scientists often observe functional data instead of univariate or multivariate ones. As a consequence, there is a need for testing procedures adapted to these infinite dimensional data. To this end, this paper proposes a nonparametric spatial scan statistic based on the Wilcoxon-Mann-Whitney two-sample test for functional data introduced by Chakraborty and Chaudhuri (2015). As shown in the application to simulated and real data, this scan procedure is much more suitable for functional data than existing ones, and its implementation in the `R` package `HDSpatialScan` makes it easy and quick to compute.

For sake of simplicity, we decided to focus on functional data belonging to an Hilbert space. We must mention that extending this work to data belonging to a more general Banach space is straightforward since the Wilcoxon-Mann-Whitney statistic of Chakraborty and Chaudhuri (2015) can be generalized to such a space.

This scan statistic allows to detect clusters using functional data indexed by space without assuming anything about their distribution. Another functional spatial scan statistic could be proposed using any other two-sample test statistic for functional data (Zhang and Chen, 2007; Zhang et al., 2010) as long as its asymptotic distribution is known. In a preprint, Frévent et al. (2020) recently proposed a parametric spatial scan statistic which is derived from the functional ANOVA test introduced by Cuevas et al. (2004). In their work, they compared our scan statistic $\Lambda_{\mathrm{WMWFSS}}$ with their statistic. They conclude, with simulation studies, that our nonparametric method performs better against non Gaussian data. `R` codes of this parametric extension are also available in the package `HDSpatialScan` (Frévent et al., 2021).

The scan method we propose allows to detect multiple clusters. If two "opposite" clusters (for example one exhibiting higher rates than expected and the other lower rates than expected) exist in two disjoint areas of the observation

domain D, the scan method first computes the concentration index for both of them and decides which one is the most significant one. Secondly, the sequential procedure may exhibit the other cluster. Actually, two "opposite" clusters can cancel out with each other only if the intersection of their areas is not null but this seems very unlikely to happen.
When the functional marks associated with the spatial locations are time series, another possibility would be considering spatio-temporal cluster detection such as Kulldorff et al. (2005). Our approach is completely different since each functional mark is taken as a whole and cannot be split: the goal is to highlight the functional marks exhibiting a different behaviour on the entire temporal observation domain.
Our work is based on the frequent assumption in the literature of spatial scan statistics that the observations in different spatial locations are independent. One should be aware that this sometimes unrealistic assumption is just a means to introduce mathematical tools that can be applied even if data are spatially correlated, as explained by Glaz (2017). However, taking into account this spatial correlation in our method could be envisaged, as Loh and Zhu (2007) did in the univariate case.
Finally, since we may observe different curves in each spatial location (for example the temporal variation of different atmospheric pollutants), another perspective would be to develop a functional extension of the multivariate Gaussian scan statistic introduced by Cucala et al. (2017).

Appendix A. Examples of the generated data in subsection 3.1

The following figures show examples of simulated data using sBm process with different types of shifts.

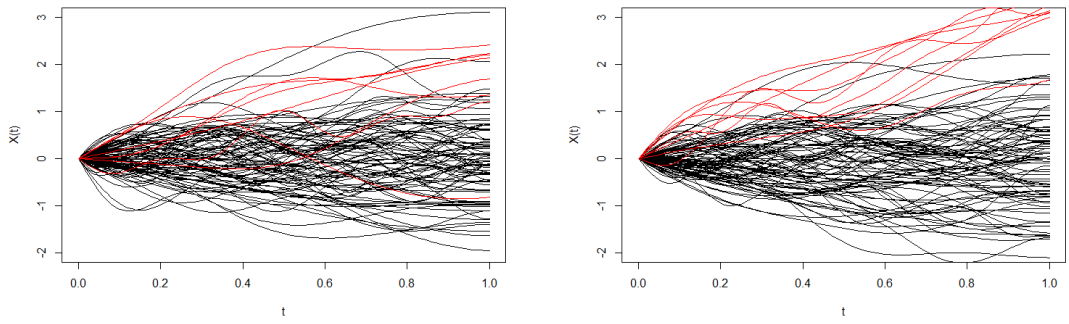

Figure A.7: An example of the simulated data for the sBm process with $\Delta_1(t) = t$ (left panel) and $\Delta_1(t) = 3t$ (right panel). Curves in red correspond to the observations in the cluster.

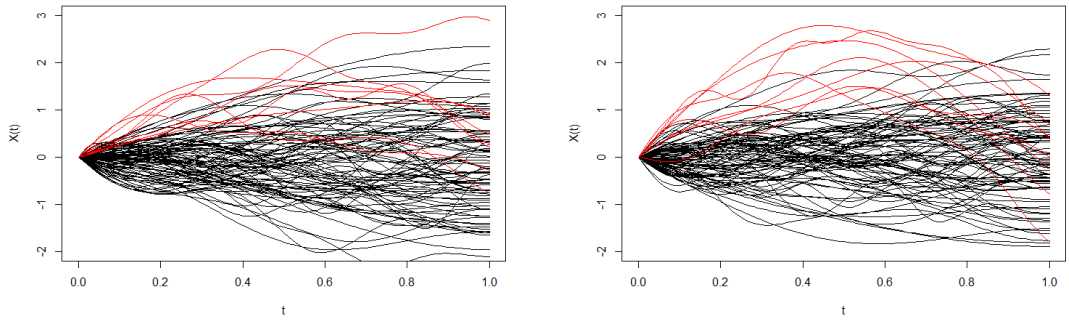

Figure A.8: An example of the simulated data for the sBm process with $\Delta_2(t) = 4t(1-t)$ (left panel) and $\Delta_2(t) = 7t(1-t)$ (right panel). Curves in red correspond to the observations in the cluster.

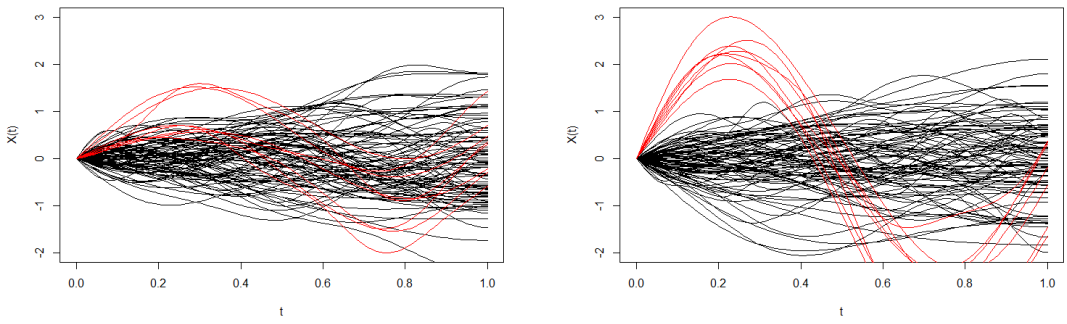

Figure A.9: An example of the simulated data for the sBm process with $\Delta_3(t) = \sin(2\pi t)$ (left panel) and $\Delta_3(t) = 2.5\sin(2\pi t)$ (right panel). Curves in red correspond to the observations in the cluster.

Appendix B. Results of the simulation study subsection 3.1

- When the true cluster is a set of 8 *départements*:
  The following Table B.5, Table B.6 and Table B.7 give the results obtained in this simulation study. Bold values indicate the best performance in each line.

Table B.5: Simulation study–AR , %TP and %FP results of the functional scan statistic Λ_{WMWFSS} and the univariate ones Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_1(t) = ct$ using two distributions: Normal and Student-t. The true cluster contains 8 *départements*.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
0.0	AR	**0.070**	0.050	0.060	0.0	AR	**0.060**	**0.060**	0.020
	%TP	0.554	0.875	**1.000**		%TP	0.479	0.479	**1.000**
	%FP	**0.442**	0.553	0.589		%FP	**0.444**	0.446	0.645
1.25	AR	**0.310**	0.260	0.110	1.5	AR	**0.240**	0.220	0.090
	%TP	0.887	0.880	**1.000**		%TP	0.885	0.847	**1.000**
	%FP	**0.188**	0.199	0.403		%FP	**0.098**	0.164	0.472
1.5	AR	**0.380**	0.340	0.150	2.0	AR	**0.600**	0.510	0.180
	%TP	0.908	0.882	**1.000**		%TP	0.935	0.939	**0.993**
	%FP	**0.110**	0.165	0.148		%FP	**0.095**	0.142	0.228
1.75	AR	**0.590**	0.450	0.160	2.5	AR	**0.790**	0.730	0.390
	%TP	**0.962**	0.956	0.953		%TP	0.949	0.938	**0.978**
	%FP	**0.074**	0.085	0.197		%FP	**0.045**	0.063	0.137
2.0	AR	**0.730**	0.660	0.300	3.0	AR	**0.920**	0.870	0.520
	%TP	**0.967**	0.966	0.933		%TP	**0.967**	0.945	0.964
	%FP	**0.049**	0.087	0.073		%FP	**0.035**	0.036	0.094
2.5	AR	**0.920**	0.890	0.570	3.5	AR	**0.980**	0.940	0.800
	%TP	**0.978**	0.961	0.879		%TP	**0.974**	0.973	0.956
	%FP	**0.056**	0.070	0.077		%FP	**0.035**	0.050	0.048
3.0	AR	**1.000**	**1.000**	0.870	4.0	AR	**0.990**	0.980	0.920
	%TP	**0.996**	0.986	0.951		%TP	**0.990**	0.980	0.942
	%FP	**0.019**	0.027	0.057		%FP	**0.021**	0.031	0.044
3.5	AR	**1.000**	**1.000**	0.910	4.5	AR	**1.000**	0.990	0.980
	%TP	**1.000**	**1.000**	0.968		%TP	**0.995**	0.990	0.941
	%FP	**0.012**	0.022	0.032		%FP	**0.013**	0.021	0.029

Table B.6: Simulation study–AR, %TP and %FP results of the functional scan statistic Λ_{WMWFSS} and the univariate ones Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_2(t) = ct(1-t)$ using two distributions: Normal and Student-t. The true cluster contains 8 *départements*.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
4.0	AR	**0.410**	0.330	0.120	4.5	AR	**0.360**	0.310	0.130
	%TP	**0.869**	0.867	0.750		%TP	0.760	0.706	**0.904**
	%FP	**0.193**	0.243	0.214		%FP	**0.121**	0.144	0.286
4.5	AR	**0.460**	0.320	0.160	5.5	AR	**0.450**	0.380	0.150
	%TP	0.853	0.844	**0.938**		%TP	0.908	0.898	**1.000**
	%FP	**0.101**	0.139	0.262		%FP	**0.070**	0.130	0.261
5.0	AR	**0.560**	0.470	0.210	6.5	AR	**0.610**	0.470	0.200
	%TP	0.944	0.910	**0.946**		%TP	0.932	0.910	**0.988**
	%FP	**0.077**	0.111	0.162		%FP	**0.067**	0.097	0.153
5.5	AR	**0.700**	0.530	0.260	7.5	AR	**0.850**	0.760	0.340
	%TP	**0.950**	0.934	0.923		%TP	**0.951**	0.950	0.901
	%FP	**0.042**	0.077	0.178		%FP	**0.065**	0.099	0.119
6.0	AR	**0.830**	0.590	0.290	8.5	AR	**0.960**	0.820	0.570
	%TP	**0.973**	0.958	0.957		%TP	**0.990**	0.988	0.982
	%FP	**0.034**	0.046	0.126		%FP	**0.023**	0.040	0.074
6.5	AR	**0.870**	0.760	0.460	9.5	AR	**0.990**	0.910	0.730
	%TP	**0.991**	0.984	0.929		%TP	**0.991**	0.984	0.945
	%FP	**0.041**	0.068	0.091		%FP	**0.020**	0.035	0.073
7.0	AR	**0.960**	0.810	0.530	10.5	AR	**0.990**	0.930	0.890
	%TP	**0.992**	0.986	0.981		%TP	**0.997**	0.995	0.980
	%FP	**0.026**	0.047	0.075		%FP	**0.015**	0.027	0.055

Table B.7: Simulation study–AR, %TP and %FP results of the functional scan statistic Λ_{WMWFSS} and the univariate ones Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_3(t) = c\sin(2\pi t)$ using two distributions: Normal and Student-t. The true cluster contains 8 *départements*.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
1.0	AR	**0.310**	0.080	0.170	1.0	AR	**0.170**	0.070	0.140
	%TP	**0.895**	0.531	0.882		%TP	0.772	0.571	**0.938**
	%FP	**0.156**	0.552	0.347		%FP	**0.126**	0.150	0.273
1.25	AR	**0.660**	0.040	0.350	1.25	AR	**0.390**	0.060	0.200
	%TP	**0.981**	0.781	0.979		%TP	**0.949**	0.667	0.938
	%FP	**0.037**	0.573	0.250		%FP	**0.109**	0.455	0.251
1.5	AR	**0.960**	0.060	0.660	1.5	AR	**0.820**	0.050	0.310
	%TP	**0.988**	0.833	0.981		%TP	**0.970**	0.425	0.960
	%FP	**0.010**	0.271	0.071		%FP	**0.053**	0.490	0.199
1.75	AR	**1.000**	0.070	0.940	1.75	AR	**0.880**	0.030	0.460
	%TP	**1.000**	0.911	0.899		%TP	**0.972**	0.833	0.959
	%FP	**0.009**	0.400	0.058		%FP	**0.015**	0.217	0.096
2.0	AR	**1.000**	0.060	**1.000**	2.0	AR	**0.990**	0.070	0.760
	%TP	**1.000**	**1.000**	0.993		%TP	**0.996**	0.893	0.991
	%FP	**0.007**	0496	0.041		%FP	**0.009**	0.387	0.070
2.25	AR	**1.000**	0.020	**1.000**	2.25	AR	**1.000**	0.070	0.890
	%TP	**1.000**	**1.000**	0.984		%TP	**1.000**	**1.000**	0.997
	%FP	**0.005**	0.052	0.029		%FP	**0.003**	0.561	0.063
2.5	AR	**1.000**	0.050	**1.000**	2.5	AR	**1.000**	0.040	0.950
	%TP	**1.000**	**1.000**	0.995		%TP	**1.000**	**1.000**	0.996
	%FP	**0.003**	0.481	0.027		%FP	**0.002**	0.311	0.046

- When the true cluster is a set of 10 *départements*:
  The following Table B.8, Table B.9 and Table B.10 give the results obtained in this simulation study. Bold values indicate the best performance in each line.

Table B.8: Simulation study–AR, %TP and %FP results of the functional scan statistic Λ_{WMWFSS} and the univariate ones Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_1(t) = ct$ using two distributions: Normal and Student-t. The true cluster contains 10 *départements*.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
0.0	AR	**0.060**	0.050	0.030	0.0	AR	**0.060**	**0.060**	0.030
	%TP	0.317	0.480	**1.000**		%TP	0.200	0.550	**0.667**
	%FP	**0.534**	0.545	0.635		%FP	**0.204**	0.206	0.544
1.0	AR	**0.210**	0.200	0.050	1.0	AR	**0.210**	0.190	0.060
	%TP	0.795	0.785	**1.000**		%TP	0.786	0.774	**1.000**
	%FP	**0.185**	0.230	0.362		%FP	**0.186**	0.200	0.514
1.25	AR	**0.360**	0.300	0.050	1.25	AR	**0.310**	0.270	0.080
	%TP	0.922	0.903	**1.000**		%TP	0.771	0.744	**0.850**
	%FP	**0.218**	0.254	0.355		%FP	**0.171**	0.173	0.405
1.5	AR	**0.650**	0.540	0.160	1.5	AR	**0.420**	0.360	0.120
	%TP	0.926	0.913	**0.963**		%TP	0.883	0.872	**1.000**
	%FP	**0.096**	0.156	0.263		%FP	**0.136**	0.147	0.313
1.75	AR	**0.750**	0.630	0.260	1.75	AR	**0.580**	0.470	0.150
	%TP	0.933	0.922	**0.950**		%TP	0.328	0.298	**1.000**
	%FP	**0.064**	0.071	0.280		%FP	**0.072**	0.106	0.316
2.0	AR	**0.900**	0.860	0.400	2.0	AR	**0.710**	0.660	0.200
	%TP	**0.960**	0.956	0.913		%TP	**0.956**	0.933	0.950
	%FP	**0.051**	0.076	0.119		%FP	**0.074**	0.091	0.123
2.25	AR	**0.950**	0.870	0.480	2.25	AR	**0.860**	0.750	0.330
	%TP	**0.960**	0.960	0.944		%TP	**0.968**	0.968	0.918
	%FP	**0.047**	0.061	0.089		%FP	**0.064**	0.069	0.131
2.5	AR	**1.000**	0.980	0.600	2.5	AR	**0.960**	0.910	0.400
	%TP	**0.988**	0.979	0.963		%TP	**0.980**	0.968	0.950
	%FP	**0.029**	0.041	0.083		%FP	**0.040**	0.049	0.085

Table B.9: Simulation study–AR, %TP and %FP results of the functional scan statistic Λ_{WMWFSS} and the univariate ones Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_2(t) = ct(1-t)$ using two distributions: Normal and Student-t. The true cluster contains 10 *départements*.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
4.0	AR	**0.440**	0.410	0.100	4.5	AR	**0.420**	0.380	0.120
	%TP	**0.911**	0.907	0.900		%TP	0.874	0.871	**0.942**
	%FP	**0.115**	0.155	0.217		%FP	**0.118**	0.129	0.289
4.5	AR	**0.660**	0.520	0.170	5.5	AR	**0.650**	0.530	0.170
	%TP	**0.950**	0.933	0.918		%TP	0.906	0.898	**0.994**
	%FP	**0.090**	0.139	0.218		%FP	**0.058**	0.097	0.345
5.0	AR	**0.800**	0.550	0.180	6.5	AR	**0.900**	0.820	0.360
	%TP	0.956	0.956	**0.983**		%TP	0.956	0.952	**0.975**
	%FP	**0.044**	0.085	0.097		%FP	**0.035**	0.055	0.162
5.5	AR	**0.930**	0.740	0.230	7.5	AR	**0.960**	0.840	0.530
	%TP	0.974	0.972	**1.000**		%TP	**0.974**	0.964	0.958
	%FP	**0.040**	0.051	0.139		%FP	**0.023**	0.058	0.120
6.0	AR	**0.980**	0.880	0.440	8.5	AR	**0.990**	0.940	0.700
	%TP	**0.977**	0.973	0.939		%TP	**0.989**	0.983	0.986
	%FP	**0.028**	0.050	0.122		%FP	**0.019**	0.056	0.076
6.5	AR	**0.990**	0.900	0.600	9.5	AR	**1.000**	0.980	0.880
	%TP	**0.980**	0.973	0.960		%TP	**0.990**	**0.990**	0.958
	%FP	**0.026**	0.049	0.091		%FP	**0.016**	0.030	0.068
7.0	AR	**0.990**	0.950	0.700	10.5	AR	**1.000**	0.990	0.980
	%TP	**0.996**	0.992	0.990		%TP	**0.996**	0.993	0.981
	%FP	**0.020**	0.032	0.078		%FP	**0.012**	0.024	0.055

Table B.10: Simulation study–AR, %TP and %FP results of the functional scan statistic Λ_{WMWFSS} and the univariate ones Λ_{MBUSS} and Λ_{DBUSS} when $\Delta_3(t) = c\sin(2\pi t)$ using two distributions: Normal and Student-t. The true cluster contains 10 *départements*.

Normal distribution					**Student-t distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
0.5	AR	**0.120**	0.100	0.090	0.5	AR	**0.110**	0.080	0.090
	%TP	0.442	0.650	**0.989**		%TP	0.755	0.725	**0.822**
	%FP	**0.356**	0.430	0.413		%FP	**0.443**	0.571	0.500
0.75	AR	**0.260**	0.060	0.100	0.75	AR	**0.110**	0.050	0.100
	%TP	0.877	0.333	**1.000**		%TP	0.855	0.820	**1.000**
	%FP	**0.138**	0.343	0.239		%FP	**0.134**	0.238	0.307
1.0	AR	**0.690**	0.070	0.250	1.0	AR	**0.340**	0.050	0.120
	%TP	**0.948**	0.757	0.912		%TP	0.953	0.800	**1.000**
	%FP	**0.052**	0.388	0.259		%FP	**0.096**	0.429	0.348
1.25	AR	**0.960**	0.040	0.480	1.25	AR	**0.760**	0.020	0.230
	%TP	**0.993**	0.950	0.975		%TP	**0.963**	0.500	0.874
	%FP	**0.015**	0.393	0.143		%FP	**0.042**	0.369	0.151
1.5	AR	**1.000**	0.060	0.720	1.5	AR	**0.950**	0.100	0.350
	%TP	**1.000**	0.983	0.994		%TP	**0.984**	0.840	0.971
	%FP	**0.004**	0.274	0.056		%FP	**0.014**	0.411	0.148
1.75	AR	**1.000**	0.060	0.990	1.75	AR	**0.990**	0.050	0.770
	%TP	**1.000**	**1.000**	**1.000**		%TP	**0.989**	0.900	0.973
	%FP	**0.003**	0.294	0.050		%FP	**0.013**	0.560	0.122
2.0	AR	**1.000**	0.070	**1.000**	2.0	AR	**1.000**	0.070	0.920
	%TP	**1.000**	**1.000**	**1.000**		%TP	**1.000**	**1.000**	0.997
	%FP	**0.002**	0.332	0.031		%FP	**0.006**	0.425	0.057

References

Alm, S.(1997). On the Distributions of Scan Statistics of a Two-Dimensional Poisson Process. *Advances in Applied Probability.* **29,** 1–18.

Ahmed, M.S, Attouch, M.K. and Dabo-Niang, S.(2017). Binary Functional Linear Models under Choice-Based Sampling. *Econometrics and Statistics.* **7,** 134–152.

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

Barnard, G.(1963). Discussion of professor bartlett's paper. *Journal of the Royal Statistical Society. Series B (Methodological).* **25B,** 294.

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

Chen, J. and Glaz, J.(2009). *Approximations for Two-Dimensional Variable Window Scan Statistics.* Springer.

Cronie, O., Ghorbani, M., Mateu, J. and Yu, J.(2019). Functional marked point processes – A natural structure to unify spatio-temporal frameworks and to analyse dependent functional data. *arXiv:1911.13142v1 [math.ST].*

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,** 321–329.

Cucala, L.(2017). Variable Window Scan Statistics: Alternatives to Generalized Likelihood Ratio Tests. *In: Glaz J., Koutras M. (eds) Handbook of Scan Statistics. Springer, New York, NY.*

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,** 111–122.

Delicado, P., Giraldo, R., Comas, C. and Mateu, J.(2010). Statistics for spatial functional data: some recent contributions *Environmetrics.* **21,** 224–239.

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

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

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

Ferraty, F. (Ed.). (2011). *Recent advances in functional data analysis and related topics.* Springer Science & Business Media.

Frévent, C., Ahmed, M.S., Marbac, M. and Genin, M.(2020). Detecting spatial clusters on functional data: new scan statistic approaches. *arXiv:2011.03482.*

Frévent, C., Ahmed, M.S., Soula, J., Smida, Z., Cucala, L., Dabo-Niang, S. and Genin, M.(2021). HDSpatialScan: Multivariate and Functional Spatial Scan Statistics. *https://CRAN.R-project.org/package=HDSpatialScan.*

Gaetan, C., Girardi, P. and Pastres, R.(2017). Spatial clustering of curves with an application of satellite data. *Spatial Statistics.* **20,** 110–124.

Glaz, J.(2017). Research on probability models for cluster of points before the year 1960. *In: Glaz J., Koutras M. (eds) Handbook of Scan Statistics. Springer, New York, NY.*

Hijmans, R. J.(2019). raster: Geographic data analysis and modelling. *R package version 2, 8-19.* doi: https://cran.r-project.org/web/packages/raster/raster.pdf.

Hope, A.(1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological).* **30,** 582–598.

Huang, L., Kulldorff, M. and Gregorio, D.(2007). A spatial scan statistic for survival data. *Biometrics.* **63,** 109–118.

Jacques, J. and Preda, C.(2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing.* **112,** 164–171.

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

Karhunen. K.(1947). Uber lineare methoden in der wahrscheinlichkeitsrechnung. *Annales Academiae Scientarum Fennicae.* **37,** 3-79.

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. *Journal of the American Statistical Association.* **101,** 1289–1305.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R. and Mostashari, F.(2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine.* **2,** 216–224.

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. and Nagarwalla, N.(1995). Spatial disease clusters: detection and inference. *Statistics in medicine.* **14,** 799–810.

Lawson, A. and Denison, D.(2002). *Spatial cluster modelling.* CRC Press, London.

Lévy, P. and Loève, M.(1948). *Processus stochastiques et mouvement brownien.* Gauthier-Villars, Paris.

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

Loh, J.M. and Zhu, Z.(2007). Accounting for spatial correlation in the scan statistic. *The Annals of Applied Statistics.* **1,** 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 order. *Ann. Math. Statist.* **18,** 50–60.

Mateu, J., Lorenzo, G. and Porcu, E.(2007). Detecting Features in Spatial Point Processes with Clutter via Local Indicators of Spatial Association. *Journal of Computational and Graphical Statistics.* **16,** 968–990.

McDonough, R. and Whalen, A.(1995). *Detection of signals in noise.* Elsevier Science.

Nagarwalla, N.(1996). A scan statistic with a variable window. *Statistics in medicine.* **15,** 845–850.

Naus, J.(1963). *Clustering of random points in the line and plane.* Ph.D. Thesis. Rutgers University, New Brunswick, NJ.

Ramsay, J.O. and Silverman, B.W.(2005). *Functional Data Analysis (Second edition).* Springer-Verlag New York.

Wilcoxon, F.(1945).Individual comparaisons by ranking methods. *Biometrics.* **1,** 80–83.

Zhang, C., Peng, H. and Zhang, J.-T.(2010). Two samples tests for functional data. *Communications in statistics. Theory and Methods.* **39,** 559–578.

Zhang, J.-T. and Chen, J.(2007). Statistical inferences for functional data. *The Annals of Statistics.* **35,** 1052–1079.

Zhang, Z., Assunção, R. and Kulldorff, M.(2010). Spatial Scan Statistics Adjusted for Multiple Clusters. *Journal of Probability and Statistics.* **2010,** 1-11.