A nonparametric spatial scan statistic for functional data

Zaineb Smida, Lionel Cucala, Ali Gannoun

A nonparametric spatial scan statistic for functional data

Zaineb SMIDA *,[1], Lionel CUCALA[1], and Ali GANNOUN[1]

[1]Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France.

Abstract

This paper introduces a nonparametric scan method for functional data indexed in space. The scan statistic that we introduce 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. This scan test seems to be powerful against any clustering alternative. We apply this method to a data set for extracting features in Spanish province population growth.

Keywords: Cluster detection, Functional data, Hilbert space, Spatial Scan statistic, Wilcoxon-Mann-Whitney test.

AMS Subject Classifications: 46E20, 62G10, 62H11, 62R10.

1 Introduction

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 time and/or space. There are several cluster detection methods used since the work of Naus (1963) and they are applied in a wide variety of research subjects such as spatial epidemiology (see, Hjalmars et al. (1996)). We can also notice that they can be used to identify low and high rate spatial clusters of the new Coronavirus identified in January 2020 in China. In the field of public health, it has become an important part of the agenda of public health authorities. It allows them for example to determine the causal factors of temporal or spatial cancer clusters (see, Ramis et al. (2013)). It is also applied in quality control to determine the clusters of defective products.
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 variable window, known as the scanning window, as it moves over the studied region with a continuous manner. These scan statistics are used to decide whether exceptional or not observing

*Corresponding author: zaineb.smida@umontpellier.fr

a cluster of events. Specifically, they are random variables used as test statistics to check the membership to a given distribution of observations, against an alternative hypothesis favouring the existence of cluster within the observation domain.

During the last decades, since the work of Kulldorff and Nagarwalla (1995) and Kulldorff (1997), numerous researchers are committed to develop scan statistics. The principle idea in the article of Kulldorff (1997) lies in the fact of scanning the studied area with circular shaped windows and select the most likely cluster as the one maximizing the statistic of a likelihood ratio test. He used either Bernoulli or Poisson model, and tested 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 other models, such as exponential (see, Huang et al. (2007)), normal (see, Kulldorff et al. (2009)), etc. These likelihood-based methods were adapted for random variables indexed in time and/or space.

In the multivariate case, Kulldorff et al. (2007) proposed an extension of the scan statistics in the spatial and spatio-temporal framework which is a combination of the univariate scan statistics but does not take into account the correlation structure of the observed variables. This problem was recently tackled by Shen and Jiang (2014) and Cucala et al. (2017). However, in these latter, the scan statistics based on the likelihood ratio are computed when the data follow a Gaussian model. A natural question arises: how can we detect a spatial cluster when the data are not Gaussian? In order to overcome this problem, researchers consider the nonparametric procedures which are applicable in many cases where the data are not drawn from a population with a specific distribution.

In the last few years, Jung and Cho (2015) and Cucala (2016) proposed separately a spatial scan statistic that does not need to assume hypotheses on the shape of the distribution. In their works, they introduced a scan statistic in the univariate setting which is based on the Wilcoxon-Mann-Whitney test. This test is one of the most popular distribution-free rank tests and it is applied to verify that two datasets come from identical populations using only the ranks of the observations. Very recently, Cucala et al. (2019) proposed a nonparametric scan statistic in the multivariate setting using the multivariate extension of the Wilcoxon-Mann-Whitney test introduced by Oja and Randles (2004).

Currently, the development of the sensoring and computing tools allows us to work with huge datasets. Hence, we have more and more access to data of functional type coming from various fields of applications like environmetrics, biometrics, medecine and econometrics (for more details see, Ramsay and Silverman (2005)). These types of data are not real random variables or vectors but they are a collection of random curves (elements) where each sample is considered as a function.

The statistical progress in Functional data analysis (FDA) deals with several kind of interesting datasets like spatial functional data and spatial-temporal data (see, Ferraty (2011)). In this paper, we focus only on the case of functional data presenting spatial dependence which are defined using process $\{X_s, s \in D \subset \mathbb{R}^d\}$, where s is a generic data location in the d-dimensional Euclidean space (we consider usually d equal to 2), D is a subset of $\mathbb{R}^d$ which is fixed and X_s are functional random curves (elements) valued in a functional space. The main difficulty of such data is the case of the infinite dimension of the space where data belong (like the Banach and the Hilbert spaces). Then, appropriate statistical tools are necessary to handle these type of data, for example to decide whether two samples of curves are issued from the same distribution.

In this context, Horvàth et al. (2013) proposed two test statistics using functional data for testing the equality of mean functions. Among these two parametric tests, one is the same as the Hotelling T^2 test in finite dimensional space. In the nonparametric setting, Chakraborty and Chaudhuri (2015) proposed an extension of the Wilcoxon-Mann-Whitney test which is based on functional ranks.
In the present work, we develop a nonparametric scan statistic for spatial functional data and more precisely for processes valued in infinite dimensional Banach space and particularly in Hilbert one.
The rest of this paper is organized as follows. In section 2, we explain how the use of the Wilcoxon-Mann-Whitney statistic proposed by Chakraborty and Chaudhuri (2015) can give birth to a nonparametric spatial scan statistic for functional data. Then, to evaluate its statistical significance, we introduce a test procedure based on permutations. In section 3, we apply the spatial scan statistic to simulated datasets firstly to see its performance. Then, we apply it to a real dataset (demographic evolution over time in Spanish provinces) in order to analyse the Spanish population structures.

2 Nonparametric spatial scan statistic for functional data

2.1 Statistic construction

Consider X a random element in a separable Banach space χ and let χ^* be the dual of χ which is the Banach space of real valued continuous linear functions on χ. We denote by $\|.\|_\chi$ a norm on χ. Let $X_1, \ldots, X_n$ be observations of X measured in 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 to 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 marks are abnormally higher or abnormally lower than elsewhere. In order to do that, we will consider the scan statistic which is usually defined to be the maximum of a concentration index observed in a collection of potential clusters using a variable window (see, 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 the windows is based only on distances between the spatial locations (see for example, Demattei et al. (2007), Assunção et al. (2006) and 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.
Most of the spatial scan statistics which have been designed for univariate or multivariate

marks are derived from a likelihood ratio, following the initial work by Kulldorff (1997). The determination of this likelihood ratio is based on a null hypothesis H_0 (absence of a cluster) against an alternative one H_1 (presence of a cluster). For example, in the article of Cucala et al. (2017), all the multivariate marks are supposed to be normally-distributed and independent. Hence, the null hypothesis considers that all the marks have equal mean vectors and covariance matrices while the alternative one considers equal covariance matrices and different means inside and outside the cluster. In the multivariate case, we remark that the likelihood ratio used in the scan statistic proposed by Cucala et al. (2017) is equivalent to the two-sample T^2 test statistic defined by Hotelling (1931). The proof can be found in the book of Anderson (2003) (page 171). However, as stated by Cucala (2017), the use of likelihood ratios is not the best method to compare all the potential clusters. Consequently, we decided to consider the nonparametric procedures.
In the univariate case, one of the most known nonparametric procedure for testing whether two samples follow the same distribution is the rank sum test proposed by Wilcoxon (1945). Another version of this test is due to Mann and Whitney (1947). These two tests are equivalent and lead to the same test named Wilcoxon-Mann-Whitney (see, Hàjek et al. (1999)).
Recently, Chakraborty and Chaudhuri (2015) proposed an extension of the Wilcoxon-Mann-Whitney test in the functional case. To construct this test, they defined the functional ranks associated to the marks $X_1, \ldots, X_n$ in χ. This latter depends on the spatial sign function. To define this function, they assumed that the space χ is smooth, i.e., $\|.\|_\chi$ is Gâteaux differentiable at each $x \neq 0, x \in \chi$ with Gâteaux derivative called $\mathrm{SGN}_x \in \chi^*$. Consequently, the sign function is defined as follows: $\forall h \in \chi$

$$\mathrm{SGN}_x(h) = \begin{cases} \lim_{t\to 0} \frac{\|x+th\|_\chi - \|x\|_\chi}{t} & \text{if } \ x \neq 0 \\ 0 & \text{if } \ x = 0 \end{cases}.$$

In the multivariate case, when $\chi = \mathbb{R}^p$, $p \in \mathbb{N}$, this sign function value is just a direction (a point on the unit p sphere) (see, Oja and Randles (2004)).
Using this sign function, the functional rank of $x \in \chi$ is thus defined as $S_x = \mathbb{E}(\mathrm{SGN}_{\{x-X\}})$ with respect to the distribution of the random element $X \in \chi$.

Now, we suppose that $X_1, \ldots, X_n$ are independent observations of 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 and Q on a smooth Banach space χ. We suppose that P and Q differ by a shift $\Delta \in \chi$ in the location. For testing the hypothesis $H_0 : \Delta = 0$ (equality of distributions) against $H_1 : \Delta \neq 0$, a Wilcoxon-Mann-Whitney test statistic extension in such space is defined by Chakraborty and Chaudhuri (2015) as:

$$T_{\mathrm{WMW}} = \frac{1}{n_Z n_{Z^c}} \sum_{\{i:s_i \in Z\}} \sum_{\{j:s_j \in Z^c\}} \mathrm{SGN}_{\{X_j - X_i\}}.$$

Assuming that the dual χ^* is a separable and type 2 Banach space (see, Araujo and Giné (1980)), Chakraborty and Chaudhuri (2015) proved the following convergence theorem :

under H_0, if $n_Z/n \to \gamma \in [0,1]$ as $n_Z, n_{Z^c} \to \infty$, then

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

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 function Γ does not depend on n_Z and n_{Z^c}, we can use

$$\mathrm{U}(Z) := (n_Z n_{Z^c}/n)^{1/2} T_{\mathrm{WMW}}$$

as a concentration index to compare potential clusters having different population sizes. Thus, the scan statistic can be defined as the maximum of the concentration index on the set of potential clusters $\mathcal{S}$ previously defined. The nonparametric functional scan statistic (NPFSS) is

$$\Lambda_{\mathrm{NPFSS}} = \max_{Z \in \mathcal{S}} \mathrm{U}(Z)$$

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

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

It is named the most likely cluster.

We remark that, in the particular case when the space χ is assumed to be an Hilbert space, the sign function is equal to $\mathrm{SGN}_x = \frac{x}{\|x\|_\chi}$. Then, the Wilcoxon-Mann-Whitney test statistic can be rewrited as

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

Since the Hilbert space χ is a type 2 Banach space, and assuming that it is a separable space, the convergence result in (1) still holds and the concentration index becomes

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

2.2 Computing the scan statistic and its significance

The computation of the scan statistic Λ_{NPFSS} involves the computation of the concentration index $U(Z)$ for every potential cluster $Z \in \mathcal{S}$. However, this index $U(Z)$ is issued from a sum of $n_Z \times n_{Z^c}$ terms so that a naive computation can be very time-consuming. 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.

After computing the scan statistic Λ_{NPFSS} 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 in comparing the observed scan statistic to the scan statistics obtained using simulated datasets. The only way to obtain these datasets is by running a method called random labelling (see, Cucala (2014)): a simulated dataset is obtained by randomly associating the marks X_i to the spatial locations s_i. Based on T random permutations, let $\Lambda^{(1)}_{\mathrm{NPFSS}},\ldots,\Lambda^{(T)}_{\mathrm{NPFSS}}$ be the observations of the scan statistics associated to the simulated datasets. Then, as stated by Dwass (1957), the p_{value} of the scan statistic Λ_{NPFSS}, observed in the initial sample, is given by

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

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

In this section, we decided to run a simulation study to evaluate the performance of the functional scan statistic Λ_{NPFSS} proposed in the previous section. Then, we compared it with the univariate spatial scan statistic introduced by Cucala (2016) and denoted by Λ_{NPUSS}. In order to do that, we replaced curves by their mean values.

Artificial datasets were generated by using the geographic locations of the 94 french administrative areas named as "*départements*". Each location associated to each "*département*" was

defined as its administrative center. The true cluster, denoted by C, is a set of 8 "*départements*" in the Parisian area. We set $\chi = L^2[0,1]$ and in each location $i = 1, \ldots, 94$ we consider the functional marks associated as follows

$$X_i(t) = \sum_{k=1}^{\infty} Z_k e_k(t) + \Delta(t)\mathbb{1}_{\{s_i \in C\}},$$

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_k's are independent random variables which correspond to the projection of X on the Karhunen-Loève basis (see, Karhunen (1947), Lévy and Loève (1948)). These marks are observed at 101 equispaced points in $[0,1]$. We have considered two different cases, namely, Z_k/σ_k having a $\mathcal{N}(0,1)$ distribution (the standard brownian motion) and a t distribution with 5 degrees of freedom.
The probability measures of the marks inside and outside the cluster C differ by a shift Δ. Two types of shifts Δ are considered namely, $\Delta_1(t) = ct$ and $\Delta_2(t) = ct(1-t)$, $c > 0$ for all $t \in [0,1]$. The parameter c is called the cluster intensity.
For each distribution and for different values of the cluster intensity, we generated 100 simulated datasets to see the performance of the proposed scan statistic Λ_{NPFSS} and the univariate one Λ_{NPUSS}. To do that, we computed three distinct criteria: the power to detect a significant cluster, the true positive (TP) rate (also called the sensitivity) and the false positive (FP) rate (also called the specificity). These three criteria were calculated as follows

- The power of the test 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 permuted samples.

- 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 "*departements*" 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 negative (FP) "*départements*" i.e, the number of "*départements*" included in the most significant cluster $\hat{\text{C}}$ but not in the true cluster C divided by the number of "*départements*" not included in C.

The following Table 1 and Table 2 gives the results obtained in this simulation study.

Normal distribution				**Student distribution**			
c		Λ_{NPFSS}	Λ_{NPUSS}	c		Λ_{NPFSS}	Λ_{NPUSS}
0.0	Power	**0.060**	0.060	0.0	Power	0.040	0.040
	%TP	**0.500**	0.500		%TP	0.750	0.750
	%FP	**0.475**	0.508		%FP	**0.512**	0.689
0.5	Power	0.110	0.110	1.0	Power	**0.170**	0.150
	%TP	0.580	0.580		%TP	**0.743**	0.725
	%FP	**0.325**	0.328		%FP	**0.276**	0.300
1.0	Power	**0.210**	0.180	1.25	Power	**0.210**	0.170
	%TP	**0.810**	0.799		%TP	**0.798**	0.787
	%FP	**0.259**	0.307		%FP	**0.172**	0.202
1.25	Power	**0.250**	0.220	1.75	Power	**0.450**	0.330
	%TP	**0.815**	0.801		%TP	**0.914**	0.909
	%FP	**0.209**	0.211		%FP	**0.113**	0.160
1.5	Power	**0.360**	0.330	2.0	Power	**0.580**	0.440
	%TP	**0.889**	0.871		%TP	**0.940**	0.920
	%FP	**0.125**	0.193		%FP	**0.110**	0.115
1.75	Power	**0.620**	0.530	2.5	Power	**0.780**	0.700
	%TP	**0.923**	0.910		%TP	**0.955**	0.954
	%FP	**0.094**	0.115		%FP	**0.078**	0.097
2.0	Power	**0.800**	0.720	3.0	Power	**0.920**	0.880
	%TP	**0.975**	0.951		%TP	**0.977**	0.964
	%FP	**0.072**	0.078		%FP	**0.047**	0.065
2.5	Power	**0.960**	0.900	3.5	Power	**0.980**	0.970
	%TP	**0.969**	0.957		%TP	**0.983**	0.979
	%FP	**0.052**	0.077		%FP	**0.024**	0.041
3.0	Power	**1.000**	0.980	4.0	Power	**1.000**	0.980
	%TP	**0.995**	0.989		%TP	**0.996**	0.990
	%FP	**0.021**	0.051		%FP	**0.021**	0.032

Table 1: Simulation study–Power, %TP and %FP results of the functional scan statistic Λ_{NPFSS} and the univariate one Λ_{NPUSS} when $\Delta_1(t) = ct$ using two distributions : Normal and Student.

Normal distribution				Student distribution			
c		Λ_{NPFSS}	Λ_{NPUSS}	c		Λ_{NPFSS}	Λ_{NPUSS}
0.5	Power	**0.080**	0.070	4.0	Power	**0.200**	0.190
	%TP	**0.609**	0.553		%TP	**0.863**	0.816
	%FP	**0.467**	0.518		%FP	**0.219**	0.229
2.5	Power	**0.100**	0.070	4.5	Power	0.240	0.240
	%TP	**0.788**	0.696		%TP	**0.859**	0.839
	%FP	**0.285**	0.334		%FP	**0.172**	0.195
3.5	Power	**0.210**	0.150	5.0	Power	**0.410**	0.360
	%TP	**0.821**	0.792		%TP	**0.854**	0.799
	%FP	**0.187**	0.191		%FP	**0.168**	0.187
4.0	Power	**0.310**	0.230	5.5	Power	**0.480**	0.410
	%TP	**0.891**	0.859		%TP	**0.935**	0.912
	%FP	**0.171**	0.198		%FP	**0.098**	0.134
4.5	Power	**0.410**	0.270	6.0	Power	**0.510**	0.450
	%TP	**0.909**	0.870		%TP	**0.912**	0.900
	%FP	**0.112**	0.134		%FP	**0.094**	0.115
5.0	Power	**0.600**	0.480	6.5	Power	**0.620**	0.600
	%TP	**0.931**	0.917		%TP	**0.952**	0.944
	%FP	**0.084**	0.129		%FP	**0.065**	0.095
5.5	Power	**0.650**	0.540	7.0	Power	**0.730**	0.640
	%TP	**0.963**	0.962		%TP	**0.969**	0.947
	%FP	**0.054**	0.095		%FP	**0.059**	0.068
6.0	Power	**0.840**	0.680	7.5	Power	**0.860**	0.650
	%TP	**0.973**	0.965		%TP	**0.980**	0.960
	%FP	**0.048**	0.095		%FP	**0.029**	0.052
6.5	Power	**0.870**	0.700	8.0	Power	**0.860**	0.750
	%TP	**0.955**	0.941		%TP	**0.985**	0.983
	%FP	**0.056**	0.123		%FP	**0.029**	0.046
7.0	Power	**0.950**	0.820	8.5	Power	**0.940**	0.840
	%TP	**0.991**	0.980		%TP	**0.992**	0.985
	%FP	**0.035**	0.063		%FP	**0.026**	0.049
7.5	Power	**0.990**	0.910	9.5	Power	**0.990**	0.920
	%TP	**0.996**	0.989		%TP	**0.990**	0.982
	%FP	**0.028**	0.054		%FP	**0.021**	0.045

Table 2: Simulation study–Power, %TP and %FP results of the functional scan statistic Λ_{NPFSS} and the univariate one Λ_{NPUSS} when $\Delta_2(t) = ct(1-t)$ using two distributions : Normal and Student.

As expected, the performance of this scan statistic is better with high cluster intensity c and we can remark that the power of Λ_{NPFSS} is higher than Λ_{NPUSS} in the different cases : this sounds logical as the first one relies on the whole information of the curves and the second one only on their mean values. It should be noted that, when c increases:

- The power of the proposed scan statistic in the different cases increases. However, using the Student distribution, the power of the model increases more slowly than the other case. This difference can be explained by the fact that the Student distribution is more heavy-tailed than the Gaussian one.

- The true positive rate %TP in the different cases increases: the detected cluster contains almost all the true "*départements*" (8) of the true cluster when c is large.

- The false positive rate %FP in the different cases decreases: a few number of false "*départements*" are assigned to the detected cluster when c is large.

3.2 Application to real data

Here, we numerically illustrate how our scan statistic model can be applied to real data. In particular, we considered data for extracting features in Spanish province population growth presented in the study of Cronie et al. (2019).
In order to study the structure of population, we considered one of the most important population characteristics which is the demographic evolution. This latter can change over time because of some factors like birth and death rates, immigration rate or economical situations.
To derive demographical evolution of the Spanish province population, we have used statistical informations provided by the *Spanish Institute of Statistics* (`www.ine.es`). We have taken the boundary and centre coordinate data of the 47 provinces of Spain (see Figure 1) from the R package *raster* (see, Hijmans (2019)). For geographical reasons, we decided to exclude from the study *Baleares* and *Canarias* islands as well as the Spanish autonomous cities (*Mellila* 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 have associated the functional mark X_i, i.e. the demographic evolution of the Spanish population for 22 distinct years starting from 1998 to 2019 which changes over time (see Figure 2). The demographic evolution was defined as the number of the total population in each province over the years 1998 to 2019, divided by the number of the total population in each province in 1998.

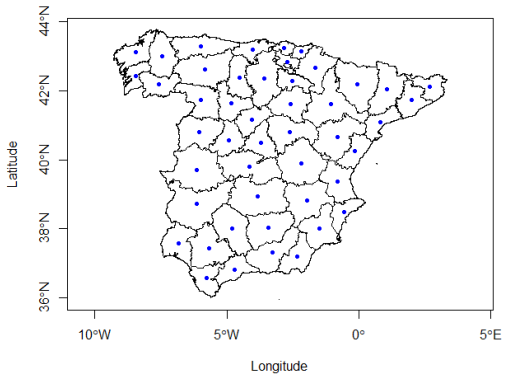

Figure 1: The 47 Spanish provinces and their geometrical centres.

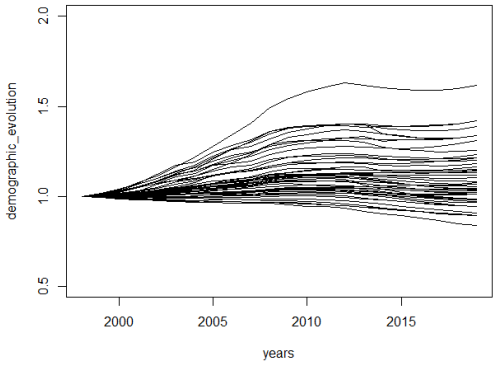

Figure 2: Demographic evolution in 47 provinces for the years 1998 to 2019

Our objective here is to detect a spatial area where the demographic evolution would be significantly higher or lower. In order to identify such a cluster, we computed the functional scan statistic on this dataset: $\Lambda_{\text{NPFSS}} = 2.72025$. Based on $T = 999$ permutations, this value is highly significant ($p_{value} = 0.001$) and the most likely cluster $\hat{C}$ is plotted in Figure 3.
This cluster includes 13 locations 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 geographical area studied. In the west part of *Castilla y León*, 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 to the most likely cluster in the Figure 4.

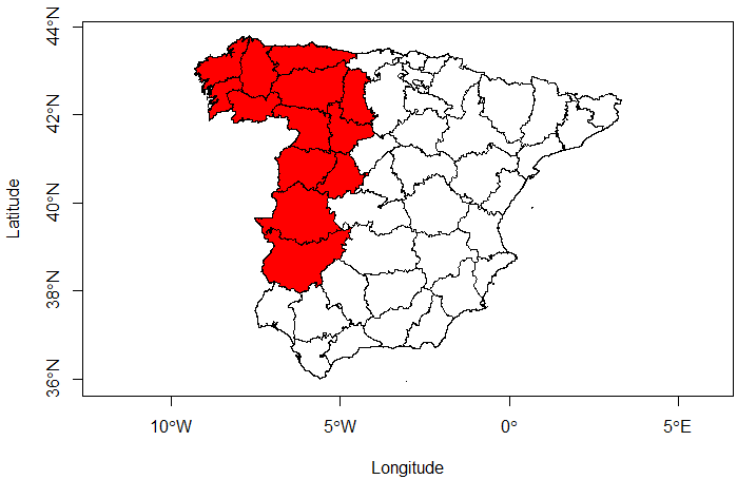

Figure 3: The most likely cluster detected by the functional scan statistic.

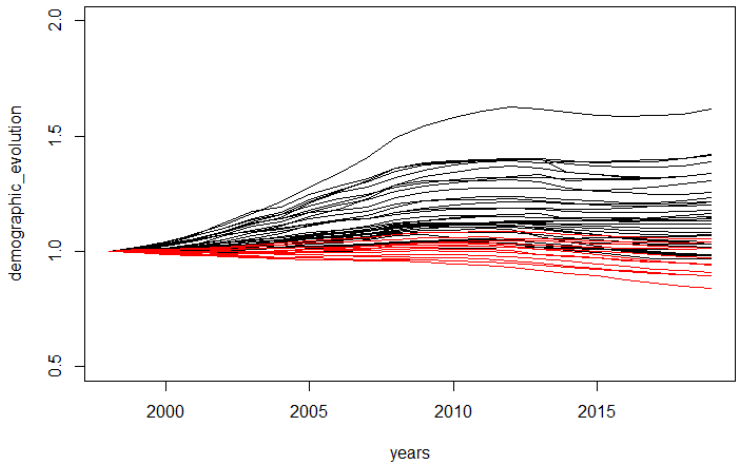

Figure 4: The demographic evolution curves (from 1998 to 2019) in each province are presented. Curves in red correspond to provinces inside the cluster, curves in black correspond to provinces outside the cluster.

We can see that this cluster includes the provinces which have the lowest demographic evolution compared to the other provinces of Spain. This can be explained by the increase in mortality rate and the decrease in birth rate in these regions.
Between the years 2006 and 2018, according to the *National Spanish Statistics Institute*, the 4 autonomous communities detected in the cluster are the territories which have the lowest birth rates (per 1000 inhabitants) compared to all 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). Also, the mortality rate (per 1000 inhabitants) increases in the provinces belonging to the detected cluster (this information is provided by the *National Spanish Statistics Institute* (INE)) and in particular *Zamora* has the highest mortality rate (12.46 in 2006 and 15.75 in 2018). Specially, these arguments can illustrate the fact that *Zamora* has the lowest evolution demographic (see Figure 5) and is close to becoming a demographic desert.

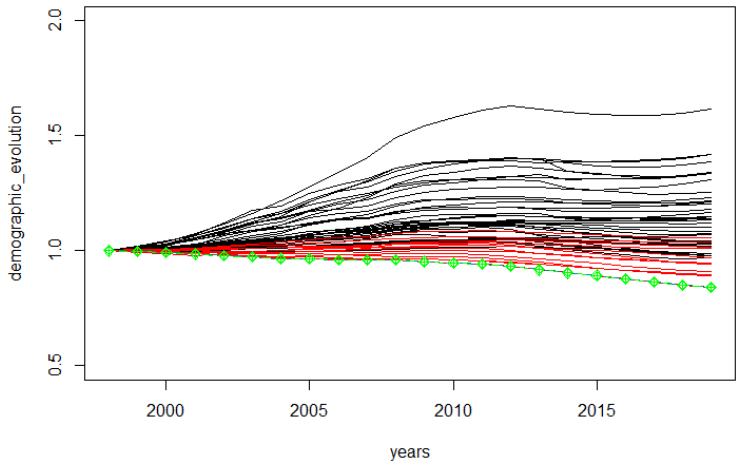

Figure 5: The demographic evolution curves (from 1998 to 2019) in each province are presented. Curves in red correspond to provinces inside the cluster, curves in black correspond to provinces outside the cluster and curve in green correspond to *Zamora* which is inside the cluster too.

Moreover, this 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 cities detected by the nonparametric scan statistic(for more details, see the website of (INE)).

4 Discussion

Nowadays and with the development of modern technology, scientists often observe functional data instead of univariate and multivariate ones. As a consequence, there is a need for testing procedures adapted to these infinite dimensional data.
In this work, we have proposed a nonparametric spatial scan statistic using the Wilcoxon-Mann-Whitney two-sample test for functional data (see, Chakraborty and Chaudhuri (2015)). This scan statistic allows as to detect clusters using functional data indexed by space without assuming anything about their distribution.
To do that, we decided to construct a nonparametric spatial scan statistic in the functional case, similar to the one proposed by Cucala (2016) in the univariate case and the other one introduced by Cucala et al. (2019) in the multivariate case. First, we proposed a nonparametric scan statistic for functional data (Λ_{NPFSS}) in Banach space then its equivalent in a particular case which is Hilbert space. Second, we defined how we can compute its significance using a Monte-Carlo procedure which provides an approximation to the null distribution. Then, we used simulation (artificial dataset) and real applications to see the performance of this Λ_{NPFSS}. In the simulation study, we can see that the Λ_{NPFSS} has a high power when the cluster intensity is large. It should be noted also that for the Student distribution, the power of the Λ_{NPFSS}

increases more slowly then the Gaussian distribution but it remains also high for a large cluster intensity.
We remark that, when the functional marks associated to 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 splitted: the goal is to highlight the functional marks exhibiting a different behaviour on the entire observation domain.
In this work we only focus on the nonparametric procedure for detecting the most likely cluster using a spatial scan statistic. A perspective would be to develop a functional extension of the multivariate Gaussian scan statistic introduced by Cucala et al. (2017). Another functional spatial scan statistic could be proposed using any other two-sample test statistic for functional data (see, Cuevas et al. (2004), Zhang and Chen (2007), Zhang et al. (2010), Horvàth et al. (2013), etc.) as long as its asymptotic distribution is known.

References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis, third ed.* Wiley series in probability and statistics.

Araujo, A. and Giné, E. (1980). *The central limit theorem for real and Banach valued random variables.* John Wiley & Sons.

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.

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.

Hàjek, J., Ŝidák, Z. and Sen, K. (1999). *Theory of Rank Tests (Second edition).* Academic Press, United States of America.

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.

Hjalmars, U. L. F., Kulldorff, M., Gustafsson, G. and Nagarwalla, N. (1996). Childhood leukaemia in Sweden: using gis and a spatial scan statistic for cluster detection. *Statistics in medicine.* **15,** 707–715.

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

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

Hotelling, H. (1931). The generalization of student's ratio. *Ann. Math. Statist.* **2,** 360–378.

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

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

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

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.

Mann, H.B., 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.

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.

Oja, R. and Randles, H.R. (2004). Multivariate nonparametric tests. *Statistical Science.* **19,** 598–605.

Ramis, R., Gomez-Barroso, D. and López-Abente, G. (2013). Cluster detection of diseases in heterogeneous populations: an alternative to scan methods. *Geospatial health.* **8,** 517–526.

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

Shen, X. and Jiang, W. (2014). Multivariate normal spatial scan statistic for detecting the most severe cluster of a disease. *Journal of Management Analytics.* **1,** 130–145.

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.