A non parametric spatial scan statistic for functional data

Zaineb Smida, Lionel Cucala, Ali Gannoun

A NONPARAMETRIC SPATIAL SCAN STATISTIC FOR FUNCTIONAL DATA

Zaineb SMIDA & Lionel CUCALA & Ali GANNOUN

Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier, France.

E-mail: zaineb.smida@umontpellier.fr ; lionel.cucala@umontpellier.fr ; ali.gannoun@umontpellier.fr

Résumé. Dans ce travail, nous introduisons une méthode non paramétrique de balayage pour des données fonctionnelles indexées dans l'espace. Nous présentons une statistique de balayage construite en utilisant la statistique de test de Wilcoxon-Mann-Whitney pour des données de dimension infinie. Cette dernière est totalement non paramétrique car elle ne suppose aucune distribution concernant les marques fonctionnelles. Ce test de balayage semble puissant contre toute alternative d'agrégation. Nous appliquons cette méthode à un ensemble de données pour extraire des caractéristiques de l'évolution démographique de provinces espagnoles.

Mots-clés. Détection d'Agrégats, Données Fonctionnelles , Espace de Hilbert, Statistique de Balayage Spatiale, Test de Wilcoxon-Mann-Whitney.

Abstract. In this work, we introduce a nonparametric scan method for functional data indexed in space. The scan statistic we present 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.

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. One of the most popular cluster detection technique is the scan statistic which was firstly introduced by Naus (1963). These scan statistics are used to decide whether exceptional or not observing a cluster of events.
During the last decades, Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed

spatial scan statistics based on Bernouilli and Poisson models. They presented a method based on the likelihood ratio and they tested the clusters' statistical significance via a Monte-Carlo procedure. In the multivariate case, scan statistics based on likelihood ratio were recently tackled by Shen and Jiang (2014) and Cucala et al. (2017). However, in these latter, the likelihood ratio used to construct the scan statistics 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 nonparametric spatial scan statistic. In their works, they introduced a scan statistic in the univariate setting which is based on the Wilcoxon-Mann-Whitney test. Very recently, Cucala et al. (2019) proposed a nonparametric scan statistic in the multivariate setting using the Wilcoxon-Mann-Whitney test introduced by Oja and Randles (2004).
Currently, the development of the sensoring allows us to work with huge datasets. Hence, we have more and more access to functional data coming from various fields of applications like environmetrics, medecine and econometrics (see, Ramsay and Silverman (2005), Ferraty and Vieu (2006)).
In the present work, we develop a nonparametric spatial scan statistic for functional data. 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 scan statistic. 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 and real datasets.

2 Nonparametric spatial scan statistic in functional data

2.1 Statistic construction

Consider X a random element in a separable Hilbert space 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 construct a 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)).
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.
Recently, Chakraborty and Chaudhuri (2015) proposed an extension of the Wilcoxon-Mann-Whitney test in the functional case using a spatial sign function defined as $\mathbf{SGN}_x = x/\|x\|_\chi$ for any non zero $x \in \chi$ and $\mathbf{SGN}_x = 0$ if $x = 0$.
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 χ. We suppose that P and Q differ by a shift $\Delta \in \chi$ in the location. For testing the hypothesis $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$, a Wilcoxon-Mann-Whitney test statistic extension in such space is defined as:

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

Chakraborty and Chaudhuri (2015) studied the asymptotic distribution of T_{WMW} and 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

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

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 nonparametric functional scan statistic (NPFSS) is

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

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

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

It is named the most likely cluster (MLC).

2.2 Rule of decision

After computing the scan statistic Λ_{NPFSS} and the most likely cluster $\hat{\mathrm{C}}$, 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, we used a strategy called "random labelling", which was already used in numerous scan methods (see for example, Cucala et al. (2019), Cucala (2017)). This method is based on random permutations and leads to an unbiased estimation of the significance value, whatever the distribution of the data.

3 Application

3.1 Simulation study

In this section, we compared Λ_{NPFSS} with the univariate spatial scan statistic introduced by Cucala (2016), denoted by Λ_{NPUSS}, which is applied to the mean values of the curves. 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 $\chi = L^2[0,1]$. For all $i \in [1,94]$, the functional marks X_i are generated using the Karhunen-Loève decomposition and they are measured at 101 equispaced points in $[0,1]$. We have considered two different cases: (i) a Gaussian distribution $\mathcal{N}(0,1)$ and (ii) a Student distribution $t(5)$. The probability measures of the marks inside and outside the cluster C differ by a shift $\Delta(t) = \mathbf{c}t, \mathbf{c} \geq 0$ for all $t \in [0,1]$. We generated 100 simulated datasets to see the performance of Λ_{NPFSS} and Λ_{NPUSS} and we computed three distinct criteria: the power to detect a significant cluster, the true positive (TP) rate and the false positive (FP) rate where a type I error equal to 5% was condidered for the rejection of H_0. The following Table 1 gives the results obtained.

		Normal distribution		**Student distribution**	
c		Λ_{NPFSS}	Λ_{NPUSS}	Λ_{NPFSS}	Λ_{NPUSS}
0.0	Power	**0.060**	0.060	0.040	0.040
	%TP	**0.500**	0.500	0.750	0.750
	%FP	**0.475**	0.508	**0.512**	0.689
1.0	Power	**0.210**	0.180	**0.170**	0.150
	%TP	**0.810**	0.799	**0.743**	0.725
	%FP	**0.259**	0.307	**0.276**	0.300
2.0	Power	**0.800**	0.720	**0.580**	0.440
	%TP	**0.975**	0.951	**0.940**	0.920
	%FP	**0.072**	0.078	**0.110**	0.115
3.0	Power	**1.000**	0.980	**0.929**	0.880
	%TP	**0.995**	0.989	**0.977**	0.964
	%FP	**0.021**	0.051	**0.047**	0.065

Table 1: Power, %TP and %FP results of Λ_{NPFSS} and Λ_{NPUSS} when $\Delta(t) = \mathbf{c}t$ in the cases (i) and (ii).

As expected, both methods perform better when the cluster intensity $\mathbf{c}$ becomes larger and our functional scan statistic gives better results since it exploits the whole information contained in the curves (not only the mean values).

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).
We considered the demographical evolution of the Spanish province population provided

by the *Spanish Institute of Statistics* (`www.ine.es`). The boundary and centre coordinate data of the 47 provinces of Spain are obtained from the *R* package *raster* .
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 and $\hat{\text{C}}$ is plotted in Figure 1A. We can see the demographic evolution curves associated to $\hat{\text{C}}$ in the Figure 1B.

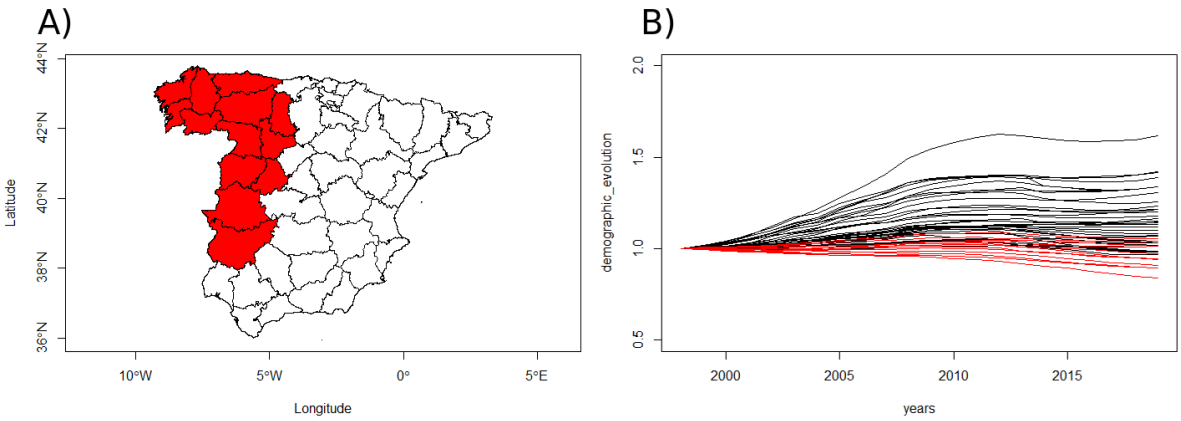

Figure 1: A) : The MLC is presented in red. B) : The demographic evolution curves (from 1998 to 2019) in each province are presented. In red : curves correspond to provinces inside the MLC. In black : curves correspond to provinces outside the MLC.

We remark that the MLC 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. 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 a higher mortality rate and a lower birth rate in these regions which have been highly affected by the economic crisis.

4 Conclusion

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 to detect clusters in 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 one introduced by Cucala et al. (2019) in the multivariate case. First, we proposed a nonparametric scan statistic for functional data in Hilbert space. Second, we defined how to compute its significance using a Monte-Carlo procedure which provides an approximation

to the null distribution. Then, we used artificial and real datasets to see the performance of this scan test.
Recently, Frévent et al. (2020) proposed a parametric spatial scan statistic, denoted by Λ_{PFSS}, which is derived from the functional ANOVA test. In their work, they compared Λ_{NPFSS} with their statistic. They conclude, with simulation studies, that the nonparametric methods performs better against non Gaussian data.

Bibliography

Chakraborty, A. and Chaudhuri, P. (2015). A Wilcoxon-Mann-Whitney type test for infinite-dimensional data. *Biometrika.***102,** 1, 239–246.
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. (2016). A Mann-Whitney scan statistic for continuous data. *Communications in Statistics - Theory and Methods.***45,** 321–329.
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.
Ferraty, F. and Vieu, Ph. (2006). *Nonparametric Functional Data Analysis (Theory and practice).* Springer-Verlag, New York.
Frévent. C., Ahmed. M.S., Marbac. M. and Genin. M. (2020). Detecting spatial clusters on functional data: a parametric scan statistic approach. *arXiv:2011.03482.*
Jung, I. and Cho, H. (2015). A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics.* **14,** 30.
Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine.* **14,** 799–810.
Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods.* **26,** 1481–1496.
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.
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.