# Statistique de balayage spatial non paramétrique pour données fonctionnelles

Zaineb Smida [1] & Lionel Cucala [2] & Ali Gannoun[3] & Ghislain Durif [4]

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

*E-mail:* [1] *zaineb.smida@umontpellier.fr ;* [2] *lionel.cucala@umontpellier.fr ;* [3] *ali.gannoun@umontpellier.fr;* [4] *ghislain.durif@umontpellier.fr;*

**Résumé.** Dans ce travail, nous introduisons une méthode non paramétrique de balayage pour données fonctionnelles indexées dans l'espace. Dans un premier temps, nous décrivons sa construction basée sur la statistique de test de Wilcoxon-Mann-Whitney pour des données de dimension infinie. Cette méthode est totalement non paramétrique car elle ne suppose aucune distribution concernant les marques fonctionnelles. Dans un deuxième temps, nous appliquons cette technique à un ensemble de données simulées pour étudier sa significativité. Puis, nous l'utilisons pour extraire des caractéristiques de l'évolution démographique de la population espagnole. Un agrégat spatial significatif de faible taux d'évolution démographique a été trouvé dans le Nord-Ouest de l'Espagne.

**Mots-clés.** Détection d'agrégats, Données fonctionnelles, Statistique de balayage spatial non paramétrique, Test de Wilcoxon-Mann-Whitney.

**Abstract.** In this work, we introduce a nonparametric scan method for functional data indexed in space. First, we describe its construction which is based on the Wilcoxon-Mann-Whitney test statistic defined for infinite dimensional data. This method is completely nonparametric as it does not assume any distribution concerning the functional marks. Second, we apply this technique to a set of simulated data to study its performance. Then, we use it to extract characteristics of the demographic evolution of the Spanish population. A significant spatial cluster with a low rate of demographic change was found in North-West of Spain

**Keywords.** Cluster detection, Functional data, Nonparametric spatial scan statistic, Wilcoxon-Mann-Whitney test.

# 1 Introduction

La détection d'agrégats est un domaine statistique qui s'est développé au cours des dernières décennies. Il est utilisé pour identifier des agrégats d'événements dans le temps et/ou dans l'espace. L'une des techniques les plus connues pour détecter des agrégats est l'utilisation de statistiques de balayage (ou de scan) qui sont apparues pour la première

fois dans les années soixantes (Naus, 1963).
Au cours des dernières décennies, Kulldorff et Nagarwalla (1995) et Kulldorff (1997) ont proposé des statistiques de balayage spatial basées sur les modèles de Bernouilli et de Poisson et utilisant le rapport de vraisemblance. Kulldorff et al. (2009) ont introduit une statistique de balayage basée sur le modèle Gaussien pour le cas où l'on observe une variable aléatoire réelle en chaque localisation spatiale. Cucala et al. (2017) l'ont étendue au cas où la variable aléatoire observée est vectorielle.
Avec le développement des capteurs, nous avons de plus en plus accès à des données qui ne sont pas, comme généralement en statistique, des observations de variables aléatoires réelles ou vectorielles mais des fonctions aléatoires : des courbes, des images, etc. (Ferraty and Vieu, 2006). Dans ce travail, nous construisons une statistique de balayage spatial non paramétrique adaptée à des données de type fonctionnel. Dans la Section 2, nous détaillons sa construction. Dans la Section 3, nous l'appliquons à des données simulées et réelles. Une conclusion est présentée dans la Section 4.

# 2 Méthodologie

Considérons $X$ un élément aléatoire dans un espace de Hilbert séparable $\chi$ qui est muni d'une norme $\|.\|_\chi$. Soit $X_1, \ldots, X_n$ des observations de $X$ mesurées sur $n$ localisations spatiales $s_1, \ldots, s_n$ inclues dans $D \subset \mathbb{R}^2$. On appelle $D$ le domaine d'observation et $X_i$ la marque associée à la location spatiale $s_i$, pour tout $i = 1, \ldots, n$. Dans la suite, nous supposons que $X_1, \ldots, X_n$ sont des observations indépendantes de $X$ (hypothèse classique pour construire des statistiques de balayage).
Notre but est de détecter un agrégat $Z \subset D$ dans lequel les marques ont un profil différent des autres. Pour cela, nous allons construire une statistique de balayage qui est définie comme le maximum d'un indice de concentration observé sur un ensemble d'agrégats potentiels. Dans ce travail, nous considérons l'ensemble des agrégats potentiels circulaires (Kulldorff, 1997) $\mathcal{S}$ défini par:

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

où $D_{i,j}$ est le disque centré sur $s_i$ et passant par $s_j$.
Pour la construction de l'indice de concentration, nous nous appuyons sur la statistique de test de Wilcoxon-Mann-Whitney pour données fonctionnelles introduite par Chakraborty et Chaudhuri (2015). Soit $Z \in \mathcal{S}$ un agrégat potentiel de taille $n_Z$, où $n_Z = \sum_{i=1}^{n} \mathbb{1}(s_i \in Z)$ et $Z^c$ son complémentaire de taille $n_{Z^c} = n - n_Z$. Supposons que les marques dans $Z$ et $Z^c$ suivent respectivement les mesures de probabilité $P$ et $Q$ sur $\chi$ et que $P$ et $Q$ diffèrent par un décalage $\Delta \in \chi$. Pour tester l'hypothèse $H_0 : \Delta = 0$ contre $H_1 : \Delta \neq 0$, la statistique de Wilcoxon-Mann-Whitney dans $\chi$ est définie par :

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

Afin de mesurer le caractère atypique des marques dans $Z$, nous utilisons l'indice de concentration

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

En effet, Chakraborty and Chaudhuri (2015) ont montré que sa distribution asymptotique sous l'hypothèse $H_0$ ne dépend pas de la taille de l'agrégat potentiel $n_Z$. Ainsi, nous introduisons la statistique de balayage spatial non paramétrique

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

et l'agrégat le plus probable

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

# 3 Application

## 3.1 Données simulées

Dans cette section, nous étudions la performance de la statistique de balayage proposée dans la section précédente sur des données simulées. Nous l'avons comparée avec (i) la statistique de balayage spatial univarié non paramétrique (Cucala, 2016) basée sur les moyennes des marques fonctionnelles et (ii) la même statistique basée sur les deviations des marques fonctionnelles par rapport à la fonction moyenne. Les localisations spatiales $s_1, \ldots, s_n$ considérées sont les centres de $n = 94$ départements français. Le vrai agrégat, noté $C$, est un ensemble de départements dans la région parisienne. Nous avons testé deux tailles pour $C$ : a) 8 départements et b) 10 départements. Pour tout $i \in \{1, \ldots, 94\}$, les courbes aléatoires $X_i$ appartiennent à $\chi = L^2([0,1], \mathbb{R})$ et sont générées suivant la décomposition de Karhunen-Loève. Deux distributions sont considérées : (i) une distribution gaussienne $\mathcal{N}(0,1)$ et (ii) une distribution de Student $t(5)$. Trois choix de décalage $\Delta$ entre la distribution à l'intérieur et à l'extérieur de l'agrégat $C$ sont étudiés : $\Delta_1(t) = ct$, $\Delta_2(t) = ct(1-t)$ et $\Delta_3(t) = c\ sin(2\pi t)$, $c > 0$ pour tout $t \in [0,1]$. Pour chaque distribution, pour chaque choix de $\Delta$ et pour différentes valeurs de $c$, nous avons généré 100 données simulées et nous avons calculé la p-valeur estimée basée sur $T = 99$ permutations aléatoires. L'erreur de type I est fixée à $\alpha = 5\%$.
Il en découle que la statistique de balayage $\Lambda_{\mathrm{WMWFSS}}$ présente de meilleures performances que les deux autres. La différence entre les puissances des différents tests est grande surtout en présence du décalage sinusoïdal $\Delta_3$.

## 3.2 Données réelles

Dans cette section, nous appliquons la statistique de balayage $\Lambda_{\mathrm{WMWFSS}}$ à des données réelles, afin d'extraire les caractéristiques de l'évolution démographique de la population

espagnole. Les localisations $s_1, \ldots, s_{47}$ considérées sont les centres de 47 provinces espagnoles et les données associées $X_1, \ldots, X_{47}$ sont les courbes de l'évolution démographique dans chaque province mesurées entre 1998 et 2019. Notre objectif est de détecter un agrégat spatial présentant une évolution démographique significativement différente.
Un agrégat significatif, composé de 13 provinces du Nord-Ouest de l'Espagne a été trouvé: il se caractérise par l'évolution démographique la plus faible par rapport aux autres provinces espagnoles.

# 4 Conclusion

Dans ce travail, nous avons proposé une statistique de balayage non paramétrique pour données fonctionnelles (pour plus de détails, voir Smida et al., 2021). Cette statistique permet de détecter des agrégats pour données fonctionnelles indexées dans l'espace sans émettre d'hypothèse sur leur distribution.
Il est à noter que, dans le cadre paramétrique, Frévent et al. (2021) ont proposé une statistique de balayage spatial pour données fonctionnelles. Ils ont conclu que les méthodes non paramétriques fonctionnent mieux dans le cas de données non gaussiennes.

# Bibliographie


Chakraborty, A. et Chaudhuri, P. (2015). A Wilcoxon-Mann-Whitney type test for infinite-dimensional data. *Biometrika*, 102, pp. 239–246.
Cucala, L. (2016). A Mann-Whitney scan statistic for continuous data. *Communications in Statistics - Theory and Methods*, 45, pp. 321–329.
Cucala, L., Genin, M., Lanier, C. et Occelli, F. (2017). A Multivariate Gaussian scan statistic for spatial data. *Spatial Statistics*, 21, pp. 66-74.
Ferraty, F. et Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics, Springer-Verlag, New York.
Frévent. C., Ahmed. M.S., Marbac. M. et Genin. M. (2021). Detecting spatial clusters on functional data: a parametric scan statistic approach. À paraître dans *Spatial Statistics.*
Kulldorff, M. et Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine.* 14, pp. 799–810.
Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods.* 26, pp. 1481–1496.
Kulldorff, M., Huang, L. et Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *International journal of health geographics,* 8, pp. 58.
Naus, J. I. (1963). *Clustering of random points in the line and plane.* Ph.D. Thesis. Rutgers University, New Brunswick, NJ.
Smida, Z., Cucala, L., Gannoun, A. et Durif, G. (2021). A Wilcoxon-Mann-Whitney spatial scan statistic for functional data. *Computational Statistics & Data Analysis*, 167.