Tests de comparaison de deux populations et statistiques de balayage spatial pour données fonctionnelles

Zaineb Smida

▶ **To cite this version:**

Zaineb Smida. Tests de comparaison de deux populations et statistiques de balayage spatial pour données fonctionnelles. Probabilités [math.PR]. Université Montpellier, 2021. Français. NNT : 2021MONTS138 . tel-04604008

HAL Id: tel-04604008

https://theses.hal.science/tel-04604008v1

Submitted on 6 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique

École doctorale I2S - Information, Structures, Systèmes

Unité de recherche UMR 5149 - IMAG - Institut Montpelliérain Alexander Grothendieck

Tests de comparaison de deux populations et statistiques de balayage spatial pour données fonctionnelles

Présentée par Zaineb SMIDA

Le 30 novembre 2021

Sous la direction de Ali GANNOUN et Lionel CUCALA

Devant le jury composé de

Ali GANNOUN	Professeur, Université de Montpellier	Directeur
Lionel CUCALA	Maître de conférences, Université de Montpellier	Co-encadrant
Liliane BEL	Professeure, AgroParisTech	Rapportrice
Cristian PREDA	Professeur, Université de Lille	Rapporteur
Jean-Noël BACRO	Professeur, Université de Montpellier	Président
Anne RUIZ-GAZEN	Professeure, Université de Toulouse 1 Capitole	Examinatrice
Jérôme SARACCO	Professeur, Université de Bordeaux	Examinateur

UNIVERSITÉ DE MONTPELLIER

Remerciements

Au terme de ce travail, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de cette thèse.

Je désire, pour commencer, témoigner de ma reconnaissance envers mes deux directeurs de thèse, Ali GANNOUN et Lionel CUCALA.

Depuis ton encadrement du Master à Monastir jusqu'à ma thèse à Montpellier, Si Ali tu as toujours su me guider et m'orienter. J'ai eu vraiment un immense privilège de travailler sous ta direction durant des années. MERCI de ne t'être jamais lassé de moi, MERCI pour ta présence, tes conseils judicieux, ta bonne humeur communicative et ta disponibilité. MERCI aussi pour ton soutien indéfectible et tes encouragements inépuisables pendant les périodes les plus difficiles de cette thèse.
MERCI à toi Lionel pour le sujet et le suivi. Tu m'as fait confiance et j'ai tout fait pour être à la hauteur. Oui, j'ai usé et abusé d'envoyer des mails mais tu n'as jamais perdu la patience et tu as toujours su répondre à mes doléances avec calme et sérénité. Merci d'avoir toujours trouvé les bons mots pour que je ne perde jamais ma motivation même face à la normalité asymptotique qui m'a empêchée de dormir plusieurs nuits. Sache que tes nombreux travaux sur les statistiques de balayage ont été une source d'inspiration pour la réalisation de cette thèse. Je suis très fière d'être ta première doctorante en thèse.

Je remercie sincèrement Monsieur le Professeur Jean-Noël BACRO pour m'avoir fait l'honneur de présider le jury de ma thèse. Il était aussi le président de tous mes comités de suivi. Ses conseils, sa gentillesse et ses encouragements m'ont été d'une très grande utilité pour arriver là où je suis. Merci Professeur du fond du cœur.

J'adresse mes vifs remerciements à Madame la Professeure Liliane BEL et Monsieur le Professeur Cristian PREDA pour avoir pu dégager un peu de leur temps précieux pour évaluer ce travail. Leurs travaux ont été une source d'inspiration pour moi et je suis ravie de les voir dans mon jury de thèse.

Que Madame la Professeure Anne RUIZ-GAZEN et Monsieur le Professeur

Jérôme SARACCO trouvent ici toute l'expression de ma gratitude et de ma reconnaissance pour m'avoir fait l'honneur de participer à mon jury. Je vous suis très reconnaissante.

Je voudrais aussi remercier tous les collègues avec lesquels j'ai collaboré pour la construction du package R `HDSpatialScan` qui figure dans ce manuscrit et plus particulièrement Camille. C'est une chance pour moi d'avoir travaillé avec eux après avoir apprécié leurs travaux au début de ma thèse sans les connaître.

Je souhaite adresser mes sincères remerciements à Ghislain DURIF qui était toujours disponible pour répondre avec patience à mes questions de programmation. Ses conseils ont contribué fortement à la réalisation de cette thèse.

Que Monsieur le Professeur Mounir ZILI soit remercié ici. Il m'a donné le goût de la recherche et la passion des mathématiques. Je tiens à lui exprimer ma plus profonde gratitude pour son soutien moral au cours de ces cinq dernières années. Merci également à mes chères enseignantes de Monastir Mouna AYACHI et Leila BEN ABDELGHANI qui m'ont ouvert la porte de la collaboration internationale et qui m'ont assuré la meilleure des formations mathématiques pour affronter la thèse dans les meilleures conditions.

J'ai effectué ma thèse à l'Université de Montpellier au sein de l'IMAG. Je tiens à remercier tout le personnel permanent spécialement Benoîte DE SAPORTA, Xavier BRY, Véronique LADRET et Pascal AZERAD pour leurs précieux conseils et informations. MERCI aussi à Christophe CRAMBES et Gilles DUCHARME d'avoir fait partie de mes comités de suivi individuel.

MERCI aussi au personnel administratif, tout particulièrement Nathalie COLLAIN pour avoir géré efficacement mon arrivée en France et même après, Sophie CAZANAVE pour son accueil chaleureux, sa gentillesse et son écoute et Baptiste CHAPUISAT pour son soutien informatique efficace.

Ces trois années ont été une nouvelle et une belle aventure dans ma vie. C'était très difficile de m'adapter au début mais j'ai fini par m'épanouir dans ce beau pays grâce à de très belles rencontres. Je remercie chaleureusement mes collègues et amis du bureau Pascal et Alain : un grand MERCI fraternel à vous deux pour tous les bons moments que nous avons partagés. A Pascal (Matteollip): MERCI de m'avoir tout le temps encouragée par la fameuse phrase "ça va aller Zainouba ! on va y arriver". Je veux te dire que heureusement tu es là Boula7ya. A Alain (Beroda) : malgré que tu aies soutenu quelques mois avant nous, Merci d'avoir

gardé contact et d'être resté toujours à nos côtés (Pascal et moi). Tu resteras pour toujours notre *chames* de bureau et notre chef de muse. J'étais très chanceuse d'avoir le même bureau que vous. Les moments que nous avons passés ensemble (les chansons, le jeu de fléchettes, la slackline, Gryffondor, le basketball, les séances de chorégraphie) resteront gravés dans mon cœur. I want to thank also our new work colleagues Xiao (the best photographer) and Junyi for his support. It is really a pleasure to meet you guys.

Je remercie chaleureusement mes amies (ma famille) de Montpellier (les Daltonnes : Rado, Kenzo, Thizit) avec lesquelles j'ai partagé et passé mes meilleurs moments dans cette ville, surtout pendant le confinement. J'adresse mes remerciements également à Maryouma surtout pour son soutien au cours de cette dernière année. Merci aussi à Antonia, Sonia, Morgane, Khadija, Dina et Safa pour leur amour. Je vous ADORE la Team.

MERCI aussi à tous les (post-)doctorants spécialement à Robert pour sa disponibilité et pour son aide administrative. MERCI aussi à Julien, Gwenaël, Tom et Alan pour leur gentillesse et MERCI à Salomé, Faustine, Samia, Tiffany, Florent, Bart, Thibault (GodinT), Nathan, Raphaël, Juliette, Guillaume, Pablo, Victor, Hermès, André,... Je tiens à remercier vivement Anis.F, Med-Iheb, Paul et Ibrahim pour leur présence.

Enfin, un MERCI énorme et TRÈS spécial et TRÈS particulier à mes parents, mon frère et ma sœur. Votre soutien et votre amour m'ont permis de réussir et ont abouti à la réalisation de cette thèse. Un grand MERCI pour votre soutien continuel et sans relâche pendant toutes mes études. C'est grâce à vous et pour vous que j'existe. C'est avec émotion qu'à mon tour je vous dévoile le fruit de mes efforts. Malgré que je ne vous ai pas vus depuis longtemps, je suis certaine que :

« Dans une famille, on est attachés les uns aux autres par des fils invisibles qui nous ligotent, même quand on les coupe. » Jean-Michel Guenassia.

Vous êtes toute ma VIE.

Résumé long de la thèse

Les tests d'hypothèses sont parmi les procédures de décision les plus utilisées en statistiques et ils peuvent répondre à de nombreuses questions. Par exemple, les chercheurs peuvent s'interroger sur l'existence d'un agrégat de cas de Covid dans une zone géographique, les experts en santé cherchent à savoir si le vaccin testé est fiable, les biologistes s'interrogent sur l'efficacité d'un nouveau mode de culture bactérienne, etc. Par conséquent, le choix du test dépend de la question posée et de la nature des données utilisées. Les tests élementaires les plus employés dans la littérature sont les tests paramétriques et non paramétriques. Pour les tests paramétriques, les observations sont supposées suivre un modèle connu contrairement aux tests non paramétriques qui ne nécessitent pas d'hypothèse sur la loi des observations.

Une grande partie de la littérature s'intéresse aux tests non paramétriques. Ils ont l'avantage de ne pas utiliser d'hypothèses particulières sur la forme de la distribution d'origine et de se contenter d'hypothèses générales de régularité mathématique. Ces tests sont présents notamment dans la comparaison de deux échantillons indépendants. En premier lieu, la communauté scientifique a développé des tests non paramétriques à destination des variables aléatoires réelles (ce qu'on va appeler dans la suite le cadre univarié). Ces tests d'hypothèses sont construits en utilisant des statistiques de tests basées sur les rangs. Citons par exemple, le test le plus connu de Wilcoxon (1945) qui est équivalent au test de Mann and Whitney (1947). Ces deux tests utilisent des statistiques équivalentes qui mène au même test nommé : test de Wilcoxon-Mann-Whitney. Il existe aussi dans la littérature un test moins populaire que celui de Wilcoxon-Mann-Whitney appelé test de la médiane. Il a été introduit par Mood (1950). Ce test est présent dans de nombreux ouvrages (voir par exemple Van Der Vaart , 1998 et Capéraà and Cutsem , 1988) à travers une statistique de test de rang qui sera utilisée dans cette thèse. En deuxième lieu, ces deux derniers tests ont été étendus dans le cadre multivarié (en utilisant des vecteurs aléatoires réels). Leurs extensions se trouvent dans les travaux de Oja and Randles (2004) et Oja (2010).

Le domaine de détection d'agrégats est devenu un domaine vaste de la statistique dans les dernières décennies. Parmi les méthodes connues pour détecter des agrégats, on peut citer les statistiques de balayage (ou de scan), basées sur une collection de tests statistiques, qui sont apparues initialement dans les années

soixante (Naus , 1963). Ces techniques permettent de détecter des agrégats dans une zone géographique, de déterminer leurs allures (normale ou anormale, élévée ou faible, etc.) et de mesurer leurs significativités. La principale étape de la méthode de détection se fait par balayage continu d'une fenêtre de taille variable (ou fixe) sur l'ensemble du domaine d'observation. Ce balayage mène dans un premier temps à construire un ensemble des agrégats appelé l'ensemble des agrégats potentiels (ou candidats) puis à utiliser une statistique de test pour tester l'hypothèse nulle (absence d'un cluster) contre une hypothèse alternative selon laquelle il existe un cluster au sein de la zone géographique étudiée. Cette dernière est appelée la statistique de balayage qui est définie simplement comme étant le maximum d'un indice de concentration sur l'ensemble des agrégats potentiels. Cet indice de concentration mesure l'écart entre les distributions de deux échantillons: celui à l'intérieur de l'agrégat potentiel et celui qui se situe à l'extérieur de cet agrégat. Plus précisément, il correspond à une statistique de test entre deux hypothèses : l'hypothèse nulle selon laquelle les deux échantillons suivent la même loi et l'hypothèse alternative selon laquelle les deux échantillons sont distribués différemment. D'une manière paramétrique, un indice de concentration pour la construction d'une statistique de balayage est défini à partir du test de rapport de vraisemblance. Des statistiques de balayage basées sur des nombreux modèles, dans le cadre univarié, ont été traité par Kulldorff (1997), Kulldorff et al. (2009), Huang et al. (2007), etc. Une extension de la statistique de balayage de Kulldorff et al. (2009), basée sur le modèle Gaussien, a été proposée par Cucala et al. (2017) dans le cadre multivarié. D'une manière non paramétrique, une statistique de balayage dans le cadre univarié a été introduite par Cucala (2016). Cette dernière a été construite à partir de la statistique de Wilcoxon-Mann-Whitney (Wilcoxon , 1945). Une extension de cette statistique dans le cadre multivarié a été aussi proposée par Cucala et al. (2019) en se basant sur la statistique de Wilcoxon-Mann-Whitney définie par Oja and Randles (2004) pour des vecteurs aléatoires réels.

Avec le développement technologique et le progrès en traitement des données, nous avons accès aujourd'hui 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. Dans ce cadre fonctionnel, il s'agit de données de dimension infinie, c'est-à-dire des élements aléatoires à valeurs dans un espace fonctionnel. L'essentiel des résultats, des méthodes et des exemples sur ce type de données, de manière très détaillée, se trouvent dans les monographies de Ramsay and Silverman (2005) et Ferraty and Vieu (2006). Ce type de données est utilisé fréquemment dans des nombreux domaines tels que la climatologie, la médecine, la biologie, etc. Aussi, plusieurs données de

type fonctionnel sont disponibles sur le package `fda` : on trouve par exemple, `CanadianWeather`, `growth`, `pinch`, etc. La difficulté de traitement des données fonctionnelles nécessite une connaissance mathématique solide, relative aux espaces fonctionnels (par exemple les espaces de Banach, de Hilbert, etc.) et aux probabilités adaptées à l'étude des processus en temps continu.

De nombreux tests statistiques pour la comparaison de deux populations ont été développés pour des données fonctionnelles. D'une manière paramétrique, pour tester l'égalité entre deux moyennes de deux groupes d'observations de données fonctionnelles, Horváth et al. (2013) ont proposé deux statistiques de tests basées sur les projections orthogonales de la différence entre les deux moyennes échantillonales. Une de ces deux statistiques de test est une extension de la statistique du test de Hotelling (1931), la généralisation (dans $\mathbb{R}^d$) du test de Student qui est le test le plus populaire dans le cadre paramétrique univarié. Cuevas et al. (2004) ont construit une extension du test de l'analyse de la variance (ANOVA) pour des données fonctionnelles. De même, Zhang et al. (2010) ont introduit un test de comparaison de deux échantillons dont la statistique de test est construite à partir de la norme L^2 de la différence entre les moyennes empiriques des deux groupes. D'une manière non paramétrique, Chakraborty and Chaudhuri (2015) ont introduit une extension du test de Wilcoxon-Mann-Whitney dans des espaces de dimension infinie. Dans leur travail, ils ont développé une statistique de test similaire à celle proposée par Mann and Whitney (1947) dans un espace de Banach puis dans un espace de Hilbert. Ils ont étudié sa performance en la comparant avec d'autres statistiques de tests telles que celles de Horváth et al. (2013) et Cuevas et al. (2004). De plus, ils ont présenté ses propriétés asymptotiques sous les deux hypothèses de test.

Dans cette thèse, nous nous intéressons dans un premier temps à la construction d'une statistique de test non paramétrique pour la comparaison de deux échantillons indépendants pour des données fonctionnelles comparable à celle de Wilcoxon-Mann-Whitney proposée par Chakraborty and Chaudhuri (2015). A notre connaissance, il n'existe pas d'extension de la statistique de la médiane basée sur les rangs pour des données fonctionnelles. Nous nous sommes donc focalisés à construire cette extension. Bien que ce test soit moins utilisé que celui de Wilcoxon-Mann-Whitney, il est avantageux (s'agissant de son efficacité relative asymptotique) dans le contexte des distributions symétriques à queues lourdes (voir page 154 de Capéraà and Cutsem , 1988). Dans un deuxième temps, nous nous concentrons sur la détection des agrégats et plus particulièrement les statistiques de balayage. Dans la littérature, ces dernières ne sont développées que dans les cadres univarié et multivarié. Pour cela, nous nous intéréssons à

construire une statistique de balayage spatial non paramétrique pour des données fonctionnelles. Nous proposons une extension de celles développées par Cucala (2016) dans le cadre univarié et Cucala et al. (2019) dans le cadre multivarié en se basant sur la statistique de Chakraborty and Chaudhuri (2015). Nous proposons vers la fin, en collaboration avec des spécialistes en statistiques de balayage, un package `R` contenant la statistique de balayage non paramétrique pour des données fonctionnelles que nous avons développée ainsi que d'autres statistiques de balayage paramétriques proposées par eux très récemment pour faciliter l'application aux utilisateurs.

Ce manuscrit est composé de 4 chapitres :

Chapitre 1.

Ce chapitre donne un aperçu sur quelques méthodes existantes dans le cadre univarié et dans le cadre fonctionnel. Dans la première partie, nous nous intéressons à l'étude des tests non paramétriques existant dans le cadre univarié. Plus particulièrement, nous décrivons les tests basés sur les rangs tels que le test de Wilcoxon-Mann-Whitney et celui de la médiane qui seront nécessaires pour la compréhension de la suite de cette thèse. Dans la deuxième partie, nous décrivons la méthodologie des statistiques de balayage spatial univarié dans le cadre paramétrique et non paramétrique ainsi que le calcul de leurs significativités, afin d'introduire aisement l'extension proposée pour des données fonctionnelles dans le cadre non paramétrique, proposée dans le chapitre 3 et utilisée aussi dans le chapitre 4. Dans la dernière partie, nous donnons quelques généralités sur les données fonctionnelles qui aideront le lecteur à mieux comprendre les méthodes développées dans les chapitres suivants.

Chapitre 2.

Ce chapitre décrit la construction d'une extension du test de la médiane dans le cadre fonctionnel en utilisant les généralités sur les données fonctionnelles introduites dans le chapitre 1. Nous présentons de plus une nouvelle statistique de test similaire à celle du test de la médiane ainsi que sa loi asymptotique sous l'hypothèse nulle. Ensuite, nous comparons ces deux statistiques de test avec d'autres statistiques développées dans la littérature dans le cadre paramétrique et non paramétrique en utilisant des données simulées et réelles.
Ce travail fait l'objet d'un article en révision à la revue *Journal of Nonparametric Statistics.*

Chapitre 3.

Ce chapitre introduit une méthode de balayage spatial non paramétrique pour des données fonctionnelles similaire à celles développées dans les cadres univarié et multivarié. Dans un premier temps, nous décrivons sa construction en se basant sur la statistique de test de Wilcoxon-Mann-Whitney pour les espaces de dimension infinie. 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.
Cette étude a donné naissance à un papier accepté à la revue *Computational Statistics & Data Analysis*.

Chapitre 4.

La construction d'une statistique de balayage spatial non paramétrique pour des données fonctionnelles (chapitre 3) a ouvert de nombreuses perspectives et a donné naissance à des travaux qui ont été réalisés très récemment par des chercheurs de l'Université de Lille. En collaboration avec ces auteurs, nous avons construit un package `R` nommé `HDSpatialScan` que nous présentons dans ce chapitre. Ce package permet aux utilisateurs d'appliquer plus rapidement les statistiques de balayage spatial non paramétrique (présentées dans le chapitre 3) et paramétrique (développées par nos collègues de Lille) pour des données fonctionnelles et de visualiser les agrégats détéctés. Nous présentons de plus l'utilisation de ce package illustrée par des exemples de jeux de données inclus dans ce dernier.
Ce chapitre a fait l'objet d'un article en révision à la revue *The R Journal*.

Chapitre 5.

Une conclusion générale, la présentation de travaux en cours pour la construction des nouvelles statistiques de balayage pour des données fonctionnelles (une dans le cadre paramétrique et une autre dans le cadre non paramétrique) ainsi que des perspectives viennent compléter ce manuscrit.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

"In the universe, there are things that are known, and things that are unknown, and in between, there are doors."

— William Blake

Chapter contents

Dans ce chapitre, nous nous focalisons, dans un premier temps (Section 1.1), sur l'étude des tests de comparaison de deux populations dans le cadre univarié. Plus précisement, nous nous intéressons à l'étude des tests non paramétriques basés sur les rangs. Celle-ci nous permet d'élaborer une extension d'un test basé sur les rangs dans le cadre fonctionnel qui sera présentée dans le chapitre 2. Dans un deuxième temps (Section 1.2), nous présentons les statistiques de balayage spatial

existantes dans le cadre univarié. Celles-ci nous permettent de construire une extension non paramétrique dans le cadre fonctionnel qui sera présentée dans le chapitre 3. Puis, nous présentons, dans la Section 1.3, quelques généralités sur les données fonctionnelles afin d'aider le lecteur à mieux comprendre la construction de ces dernières extensions dans le cadre fonctionnel.

1.1 Tests de comparaison de deux populations pour données réelles

Dans cette section, nous nous intéressons à l'étude des tests statistiques non paramétriques pour la comparaison de deux échantillons. Ces tests sont basés généralement sur les rangs des observations contrairement aux tests paramétriques qui se basent sur les valeurs des observations. L'avantage de la méthode non paramétrique est qu'elle ne fait pas d'hypothèses particulières sur la forme de la distribution d'origine. Tout le long de cette partie, on va se concentrer sur les tests basés sur les rangs pour la comparaison de deux échantillons issus de deux variables aléatoires réelles X et Y. Le but de la comparaison de ces deux échantillons est de savoir s'ils proviennent de la même distribution. Pour cela, nous utiliserons des tests qui visent à détecter toute forme de différence. Cette forme peut être introduite par un décalage entre les distributions, une différence de tendance centrale (paramètre de localisation), une différence de dispersion (paramètre d'échelle) ou une assymétrie ...
Nous nous focalisons dans ce travail sur le modèle de localisation. Pour cela, nous considérons deux échantillons indépendants et indentiquement distribués (i.i.d) : $X_1, \ldots, X_m$ de loi F_{δ_1} telle que $F_{\delta_1}(x) = F(x - \delta_1)$ et un deuxième échantillon $Y_1, \ldots, Y_n$ de loi F_{δ_2} telle que $F_{\delta_2}(x) = F(x - \delta_2)$, $\forall x \in \mathbb{R}$ et $F \in \mathcal{F}_0$ où

$$\mathcal{F}_0 = \{F; F \in \mathcal{F}, F(0) = 1/2\}$$

est l'ensemble des lois absolument continues de médiane nulle et $\mathcal{F}$ l'ensemble des lois définies sur $\mathbb{R}$ absolument continues.
On veut tester :

$$H_0 : \delta_1 = \delta_2 \text{ contre } H_1 : \delta_1 \neq \delta_2. \tag{1.1}$$

Ce modèle est appelé *un modèle de localisation.*
Puisque nous cherchons à savoir si δ_1 est différente de δ_2, il sera plus facile de réécrire ce modèle en notant $\delta = \delta_2 - \delta_1$. Le paramètre δ est appelé *un paramètre de translation.* Par conséquent, on considère $X_1, \ldots, X_m$ un échantillon de loi F et $Y_1, \ldots, Y_n$ un échantillon, indépendant du premier, de loi F_δ définie par : $\forall x \in \mathbb{R}, F_\delta(x) = F(x - \delta)$. Dans ce cas, on s'intéresse à tester :

$$H_0 : \delta = 0 \text{ contre } H_1 : \delta \neq 0. \tag{1.2}$$

L'hypothèse alternative H_1 peut être aussi $\delta > 0$ ou bien $\delta < 0$.
Après la construction des deux hypothèses, il nous faut une statistique de test.

Puisqu'on ne va se concentrer que sur les tests de rangs cela nous mène à utiliser des statistiques de tests qui sont des fonctions seulement des rangs des observations.

1.1.1 Les statistiques linéaires de rang

Les statistiques linéaires de rangs sont de la forme :

$$T_N = \sum_{i=1}^{N} c_{Ni} f(R_i),$$

où $N = m + n$, c_{Ni} est une variable indicatrice, appelée coefficient, qui vaut 1 si l'individu appartient au deuxième échantillon et 0 sinon, R_i est le rang de Y_i dans la réunion des échantillons $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ et f est une fonction score.

1.1.1.1 Exemples

- Statistique de Wilcoxon : $W_N = \sum_{i=m+1}^{N} R_i$. Dans ce cas, on a :

$$f(R_i) = R_i \text{ et } c_{Ni} = \begin{cases} 0 & \text{si } i = 1, \ldots, m, \\ 1 & \text{si } i = m+1, \ldots, N \end{cases}.$$

- Statistique de la médiane : $M_N = \sum_{i=m+1}^{N} \mathbb{1}_{\{R_i > \frac{N+1}{2}\}}$. Dans ce cas, on a :

$$f(R_i) = \mathbb{1}_{\{\frac{R_i}{N+1} > \frac{1}{2}\}} \text{ et } c_{Ni} = \begin{cases} 0 & \text{si } i = 1, \ldots, m, \\ 1 & \text{si } i = m+1, \ldots, N \end{cases}.$$

1.1.1.2 Régle de décision

Considérons le modèle de localisation décrit dans le début de cette section et une statistique linéaire de rang T_N dont la fonction score f est croissante. Par conséquent,

1. Si $H_1 : \delta > 0$, on rejette H_0 lorsque $T_N \geqslant c_\alpha$, où c_α est déterminée par $\mathbb{P}_0(T_N \geqslant c_\alpha) = \alpha$.
2. Si $H_1 : \delta < 0$, on rejette H_0 lorsque $T_N \leqslant c_\alpha$, où c_α est déterminée par $\mathbb{P}_0(T_N \leqslant c_\alpha) = \alpha$.
3. Si $H_1 : \delta \neq 0$, on construit une région critique bilatérale à partir de 1 et 2.

Nous remarquons que les deux exemples de statistiques linéaires de rang données précédemment ont des fonctions scores croissantes donc pour trouver leurs régions critiques il suffit d'identifier leurs lois. Pour cela, nous nous intéressons dans la suite à ces deux statistiques de tests pour des variables aléatoires réelles.

1.1.2 Test de Wilcoxon-Mann-Whitney

Dans cette partie, nous nous concentrerons sur le plus populaire des tests non paramétriques qui est le test de Wilcoxon-Mann-Whitney. En réalité, il y a deux formulations différentes de la statistique de ce test mais elles sont totalement équivalentes. La première est celle de Wilcoxon (1945) qui est une statistique linéaire de rang et la deuxième est définie par Mann and Whitney (1947).

- La statistique de Wilcoxon (1945) est définie par :

$$\mathrm{W}_N = \sum_{i=1}^{n} R_i,$$

où R_i est le rang de Y_i dans la réunion des échantillons de X et de Y de taille $N = m + n$.

- La statistique de Mann and Whitney (1947) est définie par :

$$\mathrm{MW}_N = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{\{Y_i > X_j\}}.$$

- Une statistique équivalente à celle de Mann and Whitney (1947) est donnée par :

$$\mathrm{MW}_N^* = \sum_{i=1}^{n}\sum_{j=1}^{m} sign(Y_i - X_j). \tag{1.3}$$

où, pour tout $x \in \mathbb{R}$, on a

$$sign(x) = \begin{cases} -1 & \text{si} \quad x < 0 \\ 0 & \text{si} \quad x = 0 \\ 1 & \text{si} \quad x > 0 \end{cases}, \tag{1.4}$$

et la relation entre MW_N et MW_N^* est donnée par :

$$\mathrm{MW}_N^* = 2\mathrm{MW}_N - mn.$$

En effet,

$$\begin{aligned} \mathrm{MW}_N^* &= \sum_{i=1}^{n}\sum_{j=1}^{m} sign(Y_i - X_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{\{Y_i > X_j\}} - \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{\{Y_i < X_j\}} \\ &= \mathrm{MW}_N - mn + \mathrm{MW}_N \\ &= 2\mathrm{MW}_N - mn. \end{aligned}$$

Une extension de la statistique MW_N^* pour des données fonctionnelles a été introduite par Chakraborty and Chaudhuri (2014a). Nous y reviendrons dans la Section 1.3.

- La relation entre les statistiques W_N et MW_N est la suivante :

$$\mathrm{W}_N = \frac{n(n+1)}{2} + \mathrm{MW}_N.$$

En effet,
Soit S_i le rang de Y_i dans l'échantillon $(Y_1, \ldots, Y_n)$, pour $i = 1, \ldots, n$. Alors,

$$R_i = S_i + \sharp\{X_j \text{ tel que } Y_i > X_j, j = 1, \ldots, m\},$$

Par suite,

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{\{Y_i > X_j\}},$$

Ce qui nous donne

$$\mathrm{W}_N = \frac{n(n+1)}{2} + \mathrm{MW}_N.$$

Remarque 1.1. *Pour mieux comprendre les hypothèses du test, nous allons dans un premier temps intérpréter la statistique de Mann-Whitney. Remarquons que W_N est un estimateur sans biais de $\mathbb{P}(X \leq Y)$ qui peut s'intérpréter de la manière suivante :*

$$\begin{aligned}
\mathbb{P}(X \leq Y) &= \int_{\mathbb{R}} \mathbb{P}[X \leq y | Y = y]\, dF_\delta(y) \\
&= \int_{\mathbb{R}} F(y) dF_\delta(y) \\
&= \int_0^1 F \circ F_\delta^{-1}(u) du.
\end{aligned}$$

- *Sous H_0, $\mathbb{P}(X \leq Y) = \frac{1}{2}$ ce qui équivaut à dire que Méd$(X - Y) \leq 0$.*
- *Sous H_1, $\mathbb{P}(X \leq Y) > \frac{1}{2}$ (ou $< \frac{1}{2}$) ce qui équivaut à dire que Méd$(X-Y) > 0$ (ou < 0).*

Afin de décider si on rejette l'hypothèse H_0 ou non, nous avons besoin d'étudier les lois exactes et asymptotiques de W_N et de MW_N.

1.1.2.1 Lois exactes des statistiques $\mathbf{MW}_N$ et $\mathbf{W}_N$ sous H_0

- **Loi exacte de $\mathbf{W}_N$:** D'après le Théorème III.1.2 de Capéraà and Cutsem (1988), la loi de la statistique de Wilcoxon W_N est, sous H_0, symétrique par rapport à sa moyenne $\mathbb{E}_0(\mathrm{W}_N)$ et on a les moments suivants (voir Capéraà and Cutsem , 1988) :

$$\mathbb{E}_0(\mathrm{W}_N) = \frac{n(N+1)}{2} \quad \text{et} \quad \mathbb{V}_0(\mathrm{W}_N) = \frac{mn(N+1)}{12}. \tag{1.5}$$

- **Loi exacte de $\mathbf{MW}_N$:** La loi exacte de la statistique de Mann-Whitney MW_N est obtenue à partir d'une formule de récurrence (voir Théorème IV.2.2 de Capéraà and Cutsem , 1988). Les moments de cette statistique sont définis par :

$$\mathbb{E}_0(\mathrm{MW}_N) = \frac{mn}{2} \quad \text{et} \quad \mathbb{V}_0(\mathrm{MW}_N) = \frac{mn(N+1)}{12}.$$

  Pour plus de détails, vous pouvez voir aussi Hàjek et al. (1999).

1.1.2.2 Normalité asymptotique de $\mathbf{MW}_N$ sous H_0

Lorsqu'on a de grandes valeurs de m et n, on fait souvent appel aux lois asymptotiques au lieu des lois exactes. Pour cela, nous traitons dans ce paragraphe l'étude de la loi asymptotique de la statistique MW_N sous H_0 pour mieux comprendre celle de la statistique de test de la médiane étudiée à la sous-section 1.1.3.

Théorème 1.1. *(Loi asymptotique sous H_0 de la statistique de Mann-Whitney) Sous H_0 et si $\frac{m}{N} \to \lambda \in]0,1[$ lorsque $m, n \to \infty$, on a :*

$$\frac{MW_N - \mathbb{E}_0(MW_N)}{(\mathbb{V}_0(MW_N))^{1/2}} \xrightarrow{loi} \mathcal{N}(0,1).$$

Pour prouver la loi asymptotique de la statistique de Wilcoxon-Mann-Whitney, nous allons utiliser le théorème de projection suivant :

Théorème 1.2. *(Théorème de projection : Capéraà and Cutsem , 1988) Soient $X_1, \ldots, X_N$ N variables aléatoires indépendantes de lois $F_1, \ldots, F_N$ appartenant à $\mathcal{F}$. Considérons $T_N = t(X_1, \ldots, X_N)$ une statistique ayant $\mathbb{E}(T_N) = 0$ et $\mathbb{V}(T_N) < +\infty$. Désignons par $\mathcal{L}$ l'ensemble des variables aléatoires de la forme :*

$$L_N = \sum_{i=1}^{N} t_i(X_i),$$

où les t_i sont des fonctions quelconques. Alors, il existe une statistique $L^* \in \mathcal{L}$ telle que :

$$\mathbb{E}[(T_N - L^*)^2] = \inf\{\mathbb{E}[(T_N - L_N)^2]; L_N \in \mathcal{L}\}$$

et dont la forme est donnée par les fonctions t_i^* suivantes

$$\forall i \in \{1, \ldots, N\}; \quad \forall x \in \mathbb{R}; \quad t_i^*(x) = \mathbb{E}[T_N | X_i = x]\,.$$

De plus, on a

$$\mathbb{E}(t_i^*(x)) = 0 \ \text{ et } \ \mathbb{E}[(T_N - L^*)^2] = \mathbb{V}(T_N) - \mathbb{V}(L^*).$$

En utilisant ce théorème, nous obtenons la preuve suivante.

Preuve du Théorème 1.1. On considère

$$\mathrm{MW}_N = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{]0,+\infty[}(Y_i - X_j)$$

de moyenne $\mathbb{E}_0(\mathrm{MW}_N) = \frac{mn}{2}$. En effet,

$$\mathbb{E}_0(\mathrm{MW}_N) = mn\mathbb{E}(\mathbb{1}_{]0,+\infty[}(Y_i - X_j)) = mn\mathbb{P}_0(Y_1 > X_1) = \frac{mn}{2}$$

On note $T_{ij} = \mathbb{1}_{]0,+\infty[}(Y_i - X_j)$ et $T_N = \sum_{i=1}^{n}\sum_{j=1}^{m}(T_{ij} - \frac{1}{2})$. D'après le Théorème 1.2, la projection de T_N est donnée par :

$$L_N^* = n\sum_{j=1}^{m}(\frac{1}{2} - F(X_j)) + m\sum_{i=1}^{n}(F(Y_i) - \frac{1}{2}).$$

On a :

$$\frac{\sqrt{N}}{mn}L_N^* = \frac{\sqrt{N}}{\sqrt{m}}\underbrace{\sum_{j=1}^{m}\frac{(\frac{1}{2} - F(X_j))}{\sqrt{m}}}_{(*)} + \frac{\sqrt{N}}{\sqrt{n}}\underbrace{\sum_{i=1}^{n}\frac{(F(Y_i) - \frac{1}{2})}{\sqrt{n}}}_{(**)}.$$

Ceci permet de voir que $\frac{\sqrt{N}}{mn}L_N^*$ s'écrit comme étant la somme de deux sommes de variables aléatoires indépendantes. Or, on a $\frac{1}{2} - F(X_j) \sim \mathcal{U}[\frac{-1}{2}, \frac{1}{2}]$ alors, en appliquant le théorème de la limite centrale, on trouve que (*) converge en loi vers une variable aléatoire de distribution $\mathcal{N}(0, \frac{1}{12})$. De même, (**) converge en loi vers une autre variable aléatoire de loi $\mathcal{N}(0, \frac{1}{12})$. Par suite,

$$\frac{\sqrt{N}}{mn}L_N^* \xrightarrow{loi} \mathcal{N}(0, \frac{1}{12\lambda(1-\lambda)}). \tag{1.6}$$

Or, on a la décomposition suivante :

$$\frac{\sqrt{N}}{mn}T_N = \frac{\sqrt{N}}{mn}L_N^* + \frac{\sqrt{N}}{mn}(T_N - L_N^*).$$

Il nous reste à étudier la convergence de $\frac{\sqrt{N}}{mn}(T_N - L_N^*)$. Pour cela, on écrit :

$$\mathbb{E}_0[(\frac{\sqrt{N}}{mn}(T_N - L_N^*))^2] = \mathbb{V}_0(\frac{\sqrt{N}}{mn}T_N) - \mathbb{V}_0(\frac{\sqrt{N}}{mn}L_N^*).$$

- Commençons par le calcul de $\mathbb{V}_0(\frac{\sqrt{N}}{mn}T_N)$.

  On a : $T_N = \mathrm{MW}_N - \frac{mn}{2}$. Alors,

  $$\lim_{N\to+\infty}\mathbb{V}_0(\frac{\sqrt{N}}{mn}T_N) = \lim_{N\to+\infty}\mathbb{V}_0(\frac{\sqrt{N}}{mn}\mathrm{MW}_N) = \frac{1}{12\lambda(1-\lambda)}.$$

  Cette dernière égalité vient du fait que $\mathbb{V}_0(\mathrm{MW}_N) = \frac{mn(N+1)}{12}$ (d'après la loi exacte de la statistique de MW_N sous H_0).

- D'après (1.6), on a :

$$\lim_{N\to+\infty} \mathbb{V}_0(\frac{\sqrt{N}}{mn}L_N^*) = \frac{1}{12\lambda(1-\lambda)}.$$

Ainsi

$$\lim_{N\to+\infty} \mathbb{E}_0[(\frac{\sqrt{N}}{mn}(T_N - L_N^*))^2] = 0.$$

Par suite, on obtient

$$\begin{aligned}
\frac{\mathrm{MW}_N - \mathbb{E}_0(\mathrm{MW}_N)}{(\mathbb{V}_0(\mathrm{MW}_N))^{1/2}} &= \frac{T_N}{(\frac{mn(N+1)}{12})^{1/2}} \\
&= (12 \times \frac{m}{N} \times \frac{n}{N} \times \frac{N}{N+1})^{1/2} \times \frac{\sqrt{N}}{mn} \times T_N \\
&\xrightarrow{loi} (12\lambda(1-\lambda))^{1/2} \times \mathcal{N}(0, \frac{1}{12\lambda(1-\lambda)}) = \mathcal{N}(0,1).
\end{aligned}$$

Par conséquent, nous pouvons déterminer la région critique à partir de cette loi asymptotique en utilisant les quantiles de la loi gaussienne centrée réduite. □

Dans ce paragraphe, nous avons prouvé la normalité asymptotique de MW_N puisque les statistiques MW_N et W_N sont équivalentes. Une manière de prouver la normalité asymptotique de la statistique W_N serait d'utiliser directement Théorème 13.5 de Van Der Vaart (1998) qui sert à faire une décomposition de la statistique de rang linéaire pour prouver la loi asymptotique. Nous allons utiliser ce théorème pour prouver la normalité asymptotique de la statistique de la médiane M_N qui sera présentée dans le paragraphe suivant.

1.1.3 Test de la médiane

L'idée de la construction de ce test est : si on suppose que les X_j sont stochastiquement inférieurs aux Y_i alors les rangs de X_j dans l'échantillon complet de taille $N = m + n$ sont dans l'ensemble inférieurs aux rangs des Y_i. Plus précisement, les rangs des Y_i sont dans l'ensemble supérieurs à $\frac{N+1}{2}$ le rang médian de l'échantillon complet.

- On appelle test de la médiane le test de rang introduit initialement par Mood (1950). Il est défini à partir de la statistique M_N suivante :

$$M_N = \sum_{i=1}^{n} \mathbb{1}_{]0,+\infty[}(R_i - \frac{N+1}{2})$$

- Lorsque N est pair, une statistique équivalente à celle de la médiane M_N est donnée par :

$$M_N^* = \sum_{i=1}^{n} sign(R_i - \frac{N+1}{2}), \tag{1.7}$$

où la fonction $x \mapsto sign(x)$ est définie par (1.4), pour tout $x \in \mathbb{R}$. La relation entre M_N et M_N^* est donnée par :

$$M_N^* = 2M_N - 1.$$

En effet,

$$\begin{aligned} M_N^* &= \sum_{i=1}^{n} sign(R_i - \frac{N+1}{2}) \\ &= \sum_{i=1}^{n} \mathbb{1}_{\{R_i > \frac{N+1}{2}\}} - \sum_{i=1}^{n} \mathbb{1}_{\{R_i < \frac{N+1}{2}\}} \\ &= 2M_N - 1 \end{aligned}$$

- Lorsque N est impair, puisque l'une des observations peut être égale à la médiane, certains auteurs utilisent la statistique M'_N suivante :

$$M'_N = \sum_{i=1}^{n} \mathbb{1}_{]0,+\infty[}(R_i - \frac{N+1}{2}) + \frac{1}{2}\mathbb{1}_{\{0\}}(R_i - \frac{N+1}{2}).$$

Remarquons que, si N est pair, $M_N = M'_N$ puisqu'on ne peut pas avoir $i = \frac{N+1}{2}$.

1.1.3.1 Loi exacte de la statistique $\mathbf{M}_N$ sous H_0

La statistique de test de la médiane M_N suit, sous l'hypothèse H_0, une loi hypergéométrique dont les paramètres dépendent de la parité de $N = m+n$ (pour plus de détails sur la loi, voir Théorème IV.1.3 de Capéraà and Cutsem , 1988) et les moments sont définis comme suit :

$$\mathbb{E}_0(\mathrm{M}_N) = \begin{cases} \frac{n}{2} & \text{si } N \text{ est pair} \\ \frac{n(N-1)}{2N} & \text{si } N \text{ est impair} \end{cases}$$

et

$$\mathbb{V}_0(\mathrm{M}_N) = \begin{cases} \frac{mn}{4(N-1)} & \text{si } N \text{ est pair} \\ \frac{mn(N+1)}{4N^2} & \text{si } N \text{ est impair.} \end{cases}$$

Pour plus de détails, vous pouvez voir aussi Hàjek et al. (1999).

1.1.3.2 Normalité asymptotique de $\mathbf{M}_N$ sous H_0

Nous traitons dans ce paragraphe l'étude de la loi asymptotique de la statistique de la médiane M_N sous H_0. Ce paragraphe aide le lecteur à mieux comprendre le résultat théorique prouvé dans le chapitre 2 qui concerne la loi asymptotique pour le test de la médiane proposé pour des données fonctionnelles.

Théorème 1.3. *(Loi asymptotique sous H_0 de la statistique de la médiane) Sous H_0 et si $\frac{m}{N} \to \lambda \in]0,1[$ lorsque $m, n \to \infty$, on a :*

$$\frac{M_N - \mathbb{E}_0(M_N)}{(\mathbb{V}_0(M_N))^{1/2}} \xrightarrow{loi} \mathcal{N}(0,1).$$

Preuve du Théorème 1.3. D'après le Théorème 13.5 de Van Der Vaart (1998), la statistique de la médiane M_N est asymptotiquement équivalente à

$$\tilde{M}_N = -\frac{n}{N}\underbrace{\sum_{j=1}^{m}\mathbb{1}\{F(X_j)\leq 1/2\}}_{(*)'} + \frac{m}{N}\underbrace{\sum_{i=1}^{n}\mathbb{1}\{F(Y_i)\leq 1/2\}}_{(**)'} + \frac{n}{2}.$$

On va considérer la décomposition suivante :

$$\frac{1}{\sqrt{N}}M_N = \frac{1}{\sqrt{N}}\tilde{M}_N + \frac{1}{\sqrt{N}}(M_N - \tilde{M}_N).$$

- Commençons par détérminer la loi de $(*)'$ et $(**)'$.

  On a : $\mathbb{1}_{\{F(X_j)\leq 1/2\}} \sim \mathcal{B}(1/2)$. En effet,

  $$\begin{aligned}\mathbb{E}_0(\mathbb{1}_{\{F(X_j)\leq 1/2\}}) &= \mathbb{P}_0((F(X_j)-\frac{1}{2})\in]-\infty,0]) \\ &= \mathbb{P}_0(Z\leq 0) \quad \text{où} \quad Z\sim\mathcal{U}[-1/2,1/2] \\ &= \frac{1}{2}.\end{aligned}$$

  En utilisant le théorème de la limite centrale, on prouve que $\frac{1}{\sqrt{m}}(*)'$ converge en loi vers une variable aléatoire de loi $\mathcal{N}(\frac{1}{2},\frac{1}{4})$. De même, $\frac{1}{\sqrt{n}}(**)'$ converge en loi vers une autre variable aléatoire de même loi $\mathcal{N}(\frac{1}{2},\frac{1}{4})$.

- Il nous reste à démontrer la convergence de $\frac{1}{\sqrt{N}}(M_N - \tilde{M}_N)$ en moyenne quadratique vers 0.

  On a :

  $$\mathbb{E}_0((\frac{1}{\sqrt{N}}(M_N-\tilde{M}_N))^2) = \mathbb{V}_0(\frac{1}{\sqrt{N}}M_N) - \mathbb{V}_0(\frac{1}{\sqrt{N}}\tilde{M}_N).$$

  Or,

  $$\mathbb{V}_0(\frac{1}{\sqrt{N}}M_N) = \frac{1}{N}\mathbb{V}_0(M_N) = \frac{1}{N}\mathbb{V}_0(\sum_{i=1}^{n}\mathbb{1}\{R_i\leq\frac{N+1}{2}\}).$$

  Si N est pair, ceci nous donne

  $$\lim_{N\to+\infty}\mathbb{V}_0(\frac{1}{\sqrt{N}}M_N) = \lim_{N\to+\infty}\frac{mn}{4N(N-1)} = \frac{1}{4}\lambda(1-\lambda).$$

  Cette dernière variance vient de la loi exacte de la statistique de la médiane. Par suite, on peut conclure que

  $$\lim_{N\to+\infty}\mathbb{E}_0((\frac{1}{\sqrt{N}}(\tilde{M}_N - M_N))^2) = 0.$$

Ce qui nous donne

$$\begin{aligned}\frac{M_N - \mathbb{E}_0(M_N)}{(\mathbb{V}_0(M_N))^{1/2}} &= \left(\frac{4N(N-1)}{mn}\right)^{1/2} \frac{1}{N}\left(M_N - \frac{n}{2}\right) \\ &\xrightarrow{loi} \left(\frac{4}{\lambda(1-\lambda)}\right)^{1/2} \times \mathcal{N}\left(0, \frac{\lambda(1-\lambda)}{4}\right) = \mathcal{N}(0,1).\end{aligned}$$

Pareillement, le calcul est similaire si N est impair.

□

1.1.4 Comparaison entre le test de Wilcoxon-Mann-Whitney et le test de la médiane

- Remarquons tout d'abord que la statistique de test de Wilcoxon nW_N est une approximation de test de rang LMP (localement le plus puissant, voir le Théorème III.2.1 de Capéraà and Cutsem , 1988) pour la loi logistique $\mathcal{L}(0,1)$ puisque en prenant F de loi $\mathcal{L}(0,1)$ on a :

$$\frac{-f'(x)}{f(x)} = 2F(x) - 1, \forall x \in \mathbb{R},$$

où f est la densité de la loi $\mathcal{L}(0,1)$.
A partir de cette dernière égalité, la statistique de rang LMP, est donnée par :

$$T_N = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2R_i}{N+1} - 1\right).$$

Ceci nous permet de dire que le test de Wilcoxon est asymptotiquement optimal pour une distribution de type logistique (Hàjek et al. , 1999).

- Le test basé sur la statistique de la médiane M_N est asymptotiquement optimal en utilisant des lois de type double exponentielle (Hàjek et al. , 1999). En effet, nous remarquons que la statistique nM_N^* est en réalité la statistique du test de rang LMP pour la loi double exponentielle puisque en prenant F de loi $\mathcal{D}exp(0,1)$, on a :

$$\frac{-f'(x)}{f(x)} = -\mathbb{1}_{]-\infty,0[} + \mathbb{1}_{]0,+\infty[} = sign(x), \forall x \in \mathbb{R},$$

où f est la densité de la loi $\mathcal{D}exp(0,1)$ dans ce cas.
Puisque les statistiques nM_N et nM_N^* sont équivalentes, alors la statistique de la médiane nM_N est une approximation du test de rang LMP pour la loi double exponentielle.

- Nous avons étudié en sous-section 1.1.2 le test de Wilcoxon-Mann-Whitney. Bien que ce dernier soit un test non paramétrique, il est capable de concurrencer de nombreux tests paramétriques. Prenons l'exemple de test le plus populaire dans le cadre paramétrique qui est le test de Student pour deux échantillons indépendants : celui-ci n'est plus efficace que le test de Wilcoxon-Mann-Whitney uniquement si les données associées à chaque échantillon suivent une distribution gaussienne $\mathcal{N}(0,1)$ (voir la page 154 de Capéraà and Cutsem , 1988). Par contre le test de Wilcoxon-Mann-Whitney est plus efficace en présence d'autres distributions telles que la loi double exponentielle $\mathcal{D}exp(0,1)$, de Cauchy $\mathcal{C}(0,1)$, logistic $\mathcal{L}(0,1)$, etc.

- Nous avons étudié en sous-section 1.1.3 le test de la médiane. Ce test est moins souvent utilisé que celui de Wilcoxon-Mann-Whitney dans le cadre non paramétrique et celui de Student dans le cadre paramétrique. Pourtant les simulations faites dans la littérature montrent que celui-ci est avantageux dans le contexte des distributions à queues lourdes. Ce test est très efficace lorsque les données prennent des valeurs extrêmes par rapport à celui de Wilcoxon-Mann-Whitney. Nous voyons (sur la page 154 Capéraà and Cutsem , 1988) que l'efficacité relative asymptotique (voir Théorème V.2.6 de Capéraà and Cutsem , 1988) de la statistique de la médiane par rapport à celle de Wilcoxon-Mann-Whitney est supérieure à 1 en présence des distributions $\mathcal{L}(0,1)$ et $\mathcal{D}exp(0,1)$. De plus, nous observons que l'efficacité relative asymptotique du test de la médiane par rapport à celui de Student est supérieure à 1 lorsque les données sont issues des lois $\mathcal{L}(0,1)$, $\mathcal{D}exp(0,1)$ et $\mathcal{C}(0,1)$.
  Ainsi, il nous paraît intéressant de construire une extension de ce test de la médiane aux données fonctionnelles et nous détaillerons cela ultérieurement dans le chapitre 3.

1.2 Statistiques de balayage spatial pour données réelles

Dans de nombreuses applications il nous faut décider si un certain groupement d'événements dans le temps ou/et dans l'espace est significatif ou non. Pour cela, nous nous intéressons dans cette section aux statistiques de balayage qui sont utilisées pour statuer sur le caractère anormal (soit trop élévé, soit trop faible) d'une certaine accumulation d'événements. Nous verrons comment construire des statistiques de balayage basées sur des tests paramétriques existants et sur d'autre non paramétriques tels que la statistique de Wilcoxon-Mann-Whitney définie dans la section précédente. Enfin, nous étudierons la significativité de ces statistiques.

1.2.1 Données et méthodologie

Tout le long de cette partie, nous considérons une variable aléatoire réelle X continue. La nature de la variable peut être aussi discrète, multivariée, fonctionnelle, etc. Pour simplifier les choses, nous nous intéressons durant ce travail uniquement à des données associées à des localisations spatiales (dans $\mathbb{R}^2$), et non temporelles ou spatio-temporelles. Nous considérons $\{s_i, i = 1, \ldots, n\}$ l'ensemble des localisations spatiales, $s_i \in D$ étant une localisation du $i^{ème}$ événement et $D \subset \mathbb{R}^2$ le domaine d'observation. Les observations x_i de X sont associées aux localisations $s_i, i = 1, \ldots, n$. Par la suite, le jeu de données que nous observons est $\{(s_i, x_i), i = 1, \ldots, n\}$. En utilisant le vocabulaire des processus ponctuels, nous dirons que les x_i sont les marques associées aux localisations s_i.
Nous cherchons à identifier l'agrégat le plus probable (appelé aussi l'agrégat le plus significatif) qui est un sous ensemble $Z \subset D$ dans lequel les observations se comportent d'une manière exceptionnelle par rapport à celles dans Z^c, le complémentaire de Z dans D. Nous calculons dans un premier temps une statistique de balayage pour la détection de l'agrégrat le plus probable puis nous évaluons la significativité de cet agrégat par rapport à l'hypothèse nulle dans laquelle la loi de X est la même quelque soit la zone géographique.
Plus précisement, nous considérons un ensemble d'agrégats potentiels $\mathcal{S}$ (nous reviendrons sur sa construction en sous-section 1.2.2) et, pour chacun des agrégats potentiels $Z \in \mathcal{S}$, nous considérons le test de :

$$H_0 : \text{"Absence d'un agrégat"} \text{ contre } H_{1,Z} : \text{"Présence d'un agrégat Z"}.$$

Remarquons que, sous l'hypothèse nulle H_0, les marques sont distribuées de la même manière alors que sous l'hypothèse alternative $H_{1,Z}$ les marques à l'intérieur de Z et celles qui se trouvent à l'extérieur de Z sont distribuées différemment. On remarque que ces hypothèses de test sont les mêmes que celles utilisées pour la comparaison de deux populations indépendantes dans le paragraphe précédent.
A partir de ces hypothèses de tests, nous allons construire des statistiques de balayage qui vont nous permettre dans un premier temps d'identifier l'agrégat le plus probable puis de tester

$$H_0 : \text{"Absence d'un agrégat"} \text{ contre } H_1 : \text{"Présence d'un agrégat quelque part"}.$$

Ces statistiques de balayage sont définies comme étant le maximum d'un indice de concentration observé I sur l'ensemble des agrégats potentiels $\mathcal{S}$. Nous reviendrons sur la construction de l'indice de concentration en sous-section 1.2.3. Par conséquent, la statistique de balayage est définie de la manière suivante :

$$\lambda = \max_{Z \in \mathcal{S}} I(Z)$$

et l'agrégat le plus significatif est donné par :

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

Après la construction de test, il nous reste à évaluer la significativité de λ observée et donc à tester H_0 contre H_1. Pour cela, nous faisons appel à des méthodes de simulations de type Monte-Carlo que l'on va décrire à la sous-section 1.2.4.

1.2.2 Ensemble des agrégats potentiels

Initialement, les statistiques de balayage ont été constuites à partir d'une fenêtre de taille et de forme fixées (Naus , 1965), nous citons par exemple la forme rectangulaire, circulaire, triangulaire, etc. Cette fenêtre balaie l'ensemble du domaine d'observation et le but est d'identifier la zone où la concentration en événements est maximale. Cependant, il semble clair que ces restrictions ont une influence sur la détection du vrai agrégat dans les données surtout si ce dernier n'a pas la même taille que la fenêtre choisie.
Puisque nous ne travaillons qu'avec des localisations spatiales, la forme ciculaire semble la forme la plus naturelle pour modéliser un phénomène de détection d'agrégats. Le choix de cette forme le plus utilisé dans les monographies de statistiques de balayage est celui de Kulldorff (1997). Plus précisement, Kulldorff (1997) a introduit une méthode permettant d'utiliser des fenêtres de tailles variables (voir aussi, Kulldorff and Nagarwalla , 1995). Il s'est d'abord focalisé sur des fenêtres circulaires puis il a généralisé sa méthode à des formes elliptiques (Kulldorff et al. , 2006). Sans perte de généralité, nous n'expliquons dans ce paragraphe que la construction de fenêtres cirulaires que nous réutiliserons dans le chapitre 3.
Nous construisons tous les disques qui sont centrés en une localisation et dont la frontière passe par une autre. Plus précisement, nous considérons l'ensemble

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

où $D_{i,j}$ désigne le disque centré en s_i et de rayon $d(s_i, s_j)$, où $d(.,.)$ désigne la distance euclidienne.

1.2.3 Indice de concentration

L'idée originale de Kulldorff (1997) pour pouvoir comparer des agrégats potentiels de tailles différentes est d'introduire un indice de concentration basé sur un rapport de vraisemblance. Généralement, pour calculer un rapport de vraisemblance, il nous faut modéliser notre variable X à partir d'une loi de probabilité connue. Le choix de cette loi donnera naissance à une nouvelle statistique de balayage. Puisqu'on se place dans le cadre où nos variables aléatoires sont continues, le modèle le plus utilisé dans la littérature est le modèle Gaussien de Kulldorff et al. (2009).
Cependant, lorsqu'on ne connait pas la distribution de nos marques aléatoires, il n'est pas toujours préférable d'utiliser cette méthode paramétrique basée sur le rapport de vraisemblance. Pour cela, certains auteurs (Cucala , 2016; Jung and Cho , 2015) ont utilisé des indices de concentration non paramétriques qui

sont construits à partir de la statistique de Wilcoxon présentée en sous-section 1.1.2 dans le cadre univarié. Dans le cadre multivarié (lorsque X est un vecteur aléatoire réel), Cucala et al. (2019) ont proposé une statistique de balayage basée sur l'extension du test de Wilcoxon (introduite par Oja and Randles , 2004).
Nous présenterons dans le chapitre 3 une extension de ces deux dernières méthodes dans le cas où les données sont de type fonctionnelles. Mais avant tout, nous allons décrire en détails la construction de ces indices de concentration dans le cadre univarié.
Dans un premier temps, nous revenons sur l'indice de concentration introduit par Kulldorff (1997) basé sur le modèle Gaussien. Ensuite, nous ferons la correspondance entre ce dernier et la statistique de test de Student. Puis, nous détaillerons la construction de l'indice de concentration de Cucala (2016) dans le cadre non paramétrique en utilisant la statistique de Wilcoxon.

1.2.3.1 Indice de concentration paramétrique

Dans ce paragraphe, nous expliquons la méthodologie de construction d'un indice de concentration paramétrique. Ensuite, nous donnons un exemple (Kulldorff et al. , 2009). Enfin, nous présentons la relation entre l'indice de concentration utilisé par Kulldorff et al. (2009) et la statistique de test de Student.

- **Méthodologie :** l'idée générale pour la construction d'une statistique de balayage paramétrique est la suivante :

  1. On considère un modèle paramétrique $\mathcal{M}_0$ considérant des marques indépendantes et identiquement distribuées dans D.
  2. On introduit, pour chaque agrégat potentiel $Z \in \mathcal{S}$, un modèle paramétrique $\mathcal{M}_{1,Z}$ où les marques sont indépendantes et différemment distribuées dans Z et dans Z^c, ce qui implique la présence d'un agrégat significatif dans Z.
  3. On calcule le rapport de vraisemblance entre les deux modèles :

  $$RV(Z) = \frac{L^*_{1,Z}}{L^*_0},$$

  où $L^*_{1,Z}$ est la vraisemblance des observations sous le modèle $\mathcal{M}_{1,Z}$ et L^*_0 la vraisemblance des observations sous le modèle $\mathcal{M}_0$.
  4. L'indice de concentration dans Z, $I(Z)$, est construit à partir du rapport de vraisemblance $RV(Z)$.
  5. La statistique de balayage est donnée par l'indice de concentration maximum

  $$\lambda = \max_{Z \in \mathcal{S}} I(Z).$$

  Pour la construction d'un indice de concentration basé sur le rapport de vraisemblance, il nous faut préciser le choix des modèles $\mathcal{M}_0$ et $\mathcal{M}_{1,Z}$

qui dépend de la nature de la variable X. Il existe plusieurs modèles paramétriques proposées dans la littérature tel que le modèle de Bernouilli (Kulldorff , 1997), le modèle Gaussien (Kulldorff et al. , 2009), le modèle exponentiel (Kulldorff et al. , 2007), etc. Nous nous concentrerons dans cet paragraphe sur l'indice de concentration Gaussien de Kulldorff et al. (2009) et sa relation avec le test statistique T^2 de Student.

- **L'indice de concentration de Kulldorff et al. (2009) :**
On considère :

$$\mathcal{M}_0 : X_1, \cdots, X_n \text{ i.i.d } \sim \mathcal{N}(m, \sigma^2)$$

et la log-vraisemblance sous $\mathcal{M}_0$ est donnée par

$$\log\Big(L_0\left((s_1,x_1),\cdots,(s_n,x_n);m,\sigma^2\right)\Big)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{\sum_{i=1}^n(x_i-m)^2}{2\sigma^2}.$$

Notons, $\forall Z \subseteq D$,

$$\overline{X^2}(Z) = \frac{\sum_{i=1}^n X_i^2 \mathbb{1}_Z(s_i)}{n_Z}$$

la moyenne du carré des marques dans Z qui est de taille $n_Z = \sum_{i=1}^n \mathbb{1}_{\{s_i \in Z\}}$.
La log-vraisemblance maximale est obtenue lorsque $m = \bar{x}(D)$ et $\sigma^2 = \overline{x^2}(D) - \Big(\bar{x}(D)\Big)^2 := \sigma^{2*}$ et par suite elle est égale à

$$\log(L_0^*) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^{2*}) - \frac{n}{2}.$$

Il nous reste à calculer la log-vraisemblance maximale sous $H_{1,Z}$. Pour cela, nous utilisons le modèle considérant un agrégat dans Z défini par

$$\mathcal{M}_{1,Z} : X_1, \cdots, X_n \text{ indépendantes et } \begin{cases} X_i \sim \mathcal{N}(m_Z, \sigma^2_{Z,Z^c}) & \text{si } s_i \in Z, \\ X_i \sim \mathcal{N}(m_{Z^c}, \sigma^2_{Z,Z^c}) & \text{si } s_i \in Z^c, \end{cases}$$

Par suite, la log-vraisemblance sous $\mathcal{M}_{1,Z}$ vaut

$$\begin{aligned}\log\Big(L_{1,Z}\Big((s_1,x_1),\cdots,(s_n,x_n);m_Z,m_{Z^c},\sigma^2_{Z,Z^c}\Big)\Big) \;=\; & -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2_{Z,Z^c}) \\ & - \frac{1}{2\sigma^2_{Z,Z^c}}\Big(\sum_{i=1}^n \Big((x_i - m_Z)^2 \mathbb{1}_Z(s_i) \\ & + (x_i - m_{Z^c})^2 \mathbb{1}_{Z^c}(s_i)\Big)\Big).\end{aligned}$$

Considérons $m_Z = \bar{x}(Z)$, $m_{Z^c} = \bar{x}(Z^c)$ et

$$\sigma^2_{Z,Z^c} = \frac{n_Z\Big(\overline{x^2}(Z) - \Big(\bar{x}(Z)\Big)^2\Big) + n_{Z^c}\Big(\overline{x^2}(Z^c) - \Big(\bar{x}(Z^c)\Big)^2\Big)}{n} := \sigma^{2*}_{Z,Z^c}.$$

Par conséquent, la log-vraisemblance maximale vaut

$$\log(L^*_{1,Z}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^{2*}_{Z,Z^c}) - \frac{n}{2}.$$

Le logarithme du rapport de vraisemblance entre les deux modèles est donné par :

$$\log\Big(RV_G(Z)\Big) = \log(L^*_{1,Z}) - \log(L^*_0) = -\frac{n}{2}\Big(\log(\sigma^{2*}_{Z,Z^c}) - \log(\sigma^{2*})\Big).$$

Nous utiliserons donc l'indice de concentration

$$I_G(Z) = \log\Big(RV_G(Z)\Big)$$

pour la construction d'une statistique de balayage paramétrique en utilisant la méthode de vraisemblance.

- **Relation entre le rapport de vraisemblance RV_G et le test T^2 de Student :**
  Nous remarquons qu'il existe une relation entre le rapport de vraisemblance RV_G calculé dans le paragraphe précédent et la statistique T^2 de Student pour la comparaison de deux échantillons indépendants et cette relation est donnée par :

$$\{RV_G\}^{2/n} = 1 + \frac{T^2}{n-2},$$

où

$$T^2 = \frac{(\overline{x}(Z) - \overline{x}(Z^c))^2}{S^2(\frac{1}{n_Z} + \frac{1}{n_{Z^c}})} = \frac{(\overline{x}(Z) - \overline{x}(Z^c))^2}{\frac{n}{n_Z n_{Z^c}}S^2}$$

est la statistique T^2 de Student,

$$S^2 = \frac{1}{n-2}(n_Z s^2_Z + n_{Z^c} s^2_{Z^c}),$$

et

$$s^2_Z = \overline{x^2}(Z) - \Big(\bar{x}(Z)\Big)^2 \text{ et } s^2_{Z^c} = \overline{x^2}(Z^c) - \Big(\bar{x}(Z^c)\Big)^2.$$

En effet, on a

$$\begin{aligned}\log\Big(RV_G(Z)\Big) &= -\frac{n}{2}\Big(\log(\sigma^{2*}_{Z,Z^c}) - \log(\sigma^{2*})\Big) \\ &= \log\left(\left(\frac{\overline{x^2}(D) - \Big(\bar{x}(D)\Big)^2}{\frac{n-2}{n}S^2}\right)^{n/2}\right).\end{aligned}$$

Par suite, on obtient

$$\{RV_G(Z)\}^{2/n} = \frac{\overline{x^2}(D) - \Big(\bar{x}(D)\Big)^2}{\frac{n-2}{n}S^2}.$$

Or, on a

$$\overline{x^2}(D) = \frac{1}{n}\sum_{i=1}^{n} x_i^2(D) = \frac{1}{n}\Big(n_Z\overline{x^2}(Z) + n_{Z^c}\overline{x^2}(Z^c)\Big)$$

et

$$\bar{x}(D) = \frac{1}{n}\sum_{i=1}^{n} x_i(D) = \frac{1}{n}\Big(n_Z\bar{x}(Z) + n_{Z^c}\bar{x}(Z^c)\Big)$$

alors, en calculant $\overline{x^2}(D) - \Big(\bar{x}(D)\Big)^2$, on obtient

$$\overline{x^2}(D) - \Big(\bar{x}(D)\Big)^2 = \frac{n-2}{n}S^2 + B,$$

où

$$B = \frac{n_Z}{n}\Big(\bar{x}(Z)\Big)^2 - \frac{n_Z^2}{n^2}\Big(\bar{x}(Z)\Big)^2 + \frac{n_{Z^c}}{n}\Big(\bar{x}(Z^c)\Big)^2 - \frac{n_{Z^c}^2}{n^2}\Big(\bar{x}(Z^c)\Big)^2 - \frac{2n_Z n_{Z^c}\bar{x}(Z)\bar{x}(Z^c)}{n^2}.$$

En simplifiant l'expression B, nous obtenons

$$B = \frac{n_Z n_{Z^c}}{n^2}\Big(\bar{x}(Z) - \bar{x}(Z^c)\Big)^2.$$

D'où

$$\{RV_G(Z)\}^{2/n} = \frac{\frac{n-2}{n}S^2 + B}{\frac{n-2}{n}S^2} = 1 + \frac{B}{\frac{n-2}{n}S^2} = 1 + \frac{T^2}{n-2} \tag{1.8}$$

Remarquons que, dans le cadre multivarié (lorsqu'on utilise des vecteurs aléatoires réels dans $\mathbb{R}^p$), le test T^2 de Hotelling (1931) est une généralisation du test T^2 de Student. Dans ce cadre, la relation entre la statistique de test de Hotelling et le rapport de vraisemblance RV_G basé sur le modèle Gaussien multivarié est donnée par Anderson (2003).

1.2.3.2 Indice de concentration non paramétrique

D'une manière non paramétrique, pour comparer deux populations, nous utilisons souvent des tests basés sur les rangs qui servent à comparer leurs distributions. Le test le plus utilisé comme on l'a déjà mentionné en sous-section 1.1.2, est celui de Wilcoxon-Mann-Whitney. Par conséquent, pour détecter des agrégats, Cucala (2016) a construit un indice de concentration non paramétrique à partir de la statistique linéaire de rang de Wilcoxon (1945). Nous détaillons ci-dessous sa construction. Tout d'abord, nous commençons par ordonner les marques par ordre croissant :

$$X_{(1)} \leq \ldots \leq X_{(n)}.$$

Notons $R_1, \ldots, R_n$ les rangs de $X_1, \ldots, X_n$. La statistique de Wilcoxon (1945), pour la comparaison des marques dans Z et dans Z^c, est :

$$SR(Z) = \sum_{i=1}^{n} R_i \mathbb{1}_Z(s_i).$$

Un indice de concentration pour détecter des agrégats positifs n'est autre que la statistique $SR(Z)$ centrée et réduite. Plus précisement, il est défini par :

$$I_R^+(Z) = \frac{SR(Z) - \mathbb{E}_0\Big(SR(Z)\Big)}{\sqrt{\mathbb{V}_0\Big(SR(Z)\Big)}},$$

où $\mathbb{E}_0\Big(SR(Z)\Big)$ et $\mathbb{V}_0\Big(SR(Z)\Big)$ sont la moyenne et la variance de la statistique $SR(Z)$ calculées sous l'hypothèse nulle H_0 et elles sont définies par (1.5) en remplaçant respectivement m et n par n_Z et n_{Z^c}. Il est à noter que, sous H_0 l'espérance et la variance de cet indice de concentration ne dépend pas de la taille de l'agrégat n_Z.
La maximisation de cet indice correspond à la recherche des agrégats posititfs. Généralement, si on cherche à détecter les agrégats les plus significatifs (positifs ou négatifs), un indice de concentration pour la construction d'une statistique de balayage non paramétrique est donné par :

$$I_R(Z) = |I_R^+(Z)|.$$

1.2.4 Significativité

La dernière étape de la méthode de balayage consiste à évaluer la significativité de la valeur λ (sous-section 1.2.1) observée c'est à dire que nous nous intéressons à la probabilité, sous l'hypothèse nulle, d'obtenir une concentration maximale aussi élevée. Pour calculer cette dernière, il nous faut identifier la loi de probabilité sous l'hypothèse nulle de la statistique de balayage. Ceci est très compliqué en raison de la dépendance entre tous les agrégats potentiels $Z \in \mathcal{S}$. Pour cela, certains auteurs ont proposé des procédures basées sur des simulations de Monte-Carlo pour estimer la distribution de la statistique de test.

- **Dans le cadre paramétrique :** cette méthode consiste à simuler T échantillons sous l'hypothèse nulle suivant la loi de la variable X (les localisations spatiales $s_1, \ldots, s_n$ restent identiques), puis à calculer les statistiques de balayage correspondantes, notées par $\lambda_{sim}^{(1)}, \ldots, \lambda_{sim}^{(T)}$. D'après Dwass (1957), la p-valeur de la statistique de balayage observée, notée par p_{value}^{sim}, est donnée par le rapport entre le nombre de statistiques de test simulées, $\lambda_{sim}^{(i)}, 1 \leq i \leq T$ supérieures ou égales à la valeur observée de la statistique de test λ et le nombre total de simulations plus un, $T + 1$:

$$p_{value}^{sim} = \frac{1 + \sum_{i=1}^{T} \mathbb{1}_{\{\lambda_{sim}^{(i)} > \lambda\}}}{T + 1},$$

  qui n'est autre que le rang de la statistique de balayage observée λ sur l'ensemble des $T + 1$ statistiques de balayage $(\lambda^{(1)}, \ldots, \lambda^{(T)}, \lambda)$. Par conséquent, si la p_{value} est inférieure au risque de première espèce α, alors on décide de rejeter l'hypothèse H_0.

- **Dans le cadre non paramétrique :** puisqu'on ne connait pas la loi de la variable X, il est impossible de générer des données pour calculer des statistiques de balayage simulées. Une alternative à cette méthode dans ce cadre est d'utiliser des permutations aléatoires c'est à dire que nous effectuons T permutations aléatoires de nos marques puis nous évaluons les statistiques de balayage correspondantes, notées par $\lambda_{perm}^{(1)}, \ldots, \lambda_{perm}^{(T)}$. Ensuite, nous calculons la p-valeur qui sera notée par p_{value}^{perm} par analogie à la méthode paramétrique, définie par

$$p_{value}^{perm} = \frac{1 + \sum_{i=1}^{T} \mathbb{1}_{\{\lambda_{perm}^{(i)} > \lambda\}}}{T+1}.$$

  De même, si p_{value}^{perm} est inférieure au risque de première espèce α, alors H_0 sera rejetée.

1.3 Généralités sur les données fonctionnelles

Dans cette section, nous donnerons quelques généralités sur les données fonctionnelles pour aider le lecteur à comprendre les constructions faites dans les chapitres suivants. Nous commençons par introduire quelques notions utiles, pour des données fonctionnelles, qui donnent naissance à une extension du test de Wilcoxon-Mann-Whitney introduite par Chakraborty and Chaudhuri (2015) et celle du test de la médiane que nous introduirons dans le chapitre 2. Enfin, nous présenterons quelques détails sur la statistique de Wilcoxon-Mann-Whitney qui nous ont inspirés pour construire la statistique de test de la médiane et la statistique de balayage spatial non paramétrique pour des données fonctionnelles.

1.3.1 Fonction signe et fonction de distribution spatiale pour données fonctionnelles

Nous voyons bien, dans la première section de ce chapitre, que les tests statistiques non paramétriques dans le cadre univarié peuvent s'écrire grâce à la fonction signe. Nous avons vu que cette fonction permet de construire les principales statistiques de tests non paramétriques (voir par exemple les statistiques (1.3) et (1.7)). Cette dernière a été considérée comme un outil non paramétrique important dans l'analyse statistique univariée (voir Section 1.1), multivariée (voir par exemple, Oja , 2010; Oja and Randles , 2004) et fonctionnelle (voir par exemple, Chakraborty and Chaudhuri , 2014b). De plus, elle sert à définir la fonction de distribution spatiale pour des données fonctionnelles qui sera utilisée dans le chapitre 2. Par conséquent, nous allons introduire ces fonctions pour mieux comprendre la construction de la statistique de Wilcoxon-Mann-Whitney et celle de la médiane pour des données fonctionnelles.

1.3.1.1 Fonction signe

- **Dans la cadre univarié :** la fonction signe est définie sous la forme (1.4), pour tout $x \in \mathbb{R}$. D'une manière équivalente, cette dernière s'écrit sous la forme suivante :

$$sign(x) = \begin{cases} \frac{x}{|x|} & \text{si} \quad x \neq 0 \\ 0 & \text{si} \quad x = 0 \end{cases} .$$

  Pour tout $x \neq 0$, $sign(x)$ est la derivée de la fonction $x \mapsto |x|$.

- **Dans le cadre fonctionnel :** nous considérons $x \in \chi$, où χ est un espace de Banach muni de la norme $\|.\|_\chi$ et χ^* son espace dual (espace des formes linéaires continues sur χ). Nous supposons de plus que χ est lisse, c'est-à-dire que la norme $\|.\|_\chi$ est Gateaux différentiable et sa derivée de Gateaux est la fonction signe notée par $\mathbf{SGN}_x \in \chi^*$. Plus précisement, cette fonction signe est définie par :

$$\mathbf{SGN}_x(h) = \begin{cases} \lim_{t\to 0} \frac{\|x+th\|_\chi - \|x\|_\chi}{t} & \text{si} \quad x \neq 0 \\ 0 & \text{si} \quad x = 0 \end{cases} , \tag{1.9}$$

  où $h \in \chi$ est la direction de cette dérivée.

  - **Exemple 1 :** si χ est un espace de Hilbert muni d'un produit scalaire $\langle ., . \rangle_{\chi,\chi}$, la fonction signe (1.9) est égale à :

$$\mathbf{SGN}_x = \begin{cases} \frac{x}{\|x\|_\chi} & \text{si} \quad x \neq 0 \\ 0 & \text{si} \quad x = 0 \end{cases} . \tag{1.10}$$

    On peut trouver ce résultat en utilisant directement le théorème de répresentation de Riesz ou bien en faisant le calcul suivant : soit $h \in \chi$, pour tout $x \neq 0$, on a :

$$\begin{aligned} \mathbf{SGN}_x(h) &= \lim_{t\to 0} \frac{\|x+th\|_\chi - \|x\|_\chi}{t} \\ &= \lim_{t\to 0} \frac{\|x+th\|_\chi^2 - \|x\|_\chi^2}{t(\|x+th\|_\chi + \|x\|_\chi)} \\ &= \lim_{t\to 0} \frac{\langle x+th, x+th\rangle_{\chi,\chi} - \langle x, x\rangle_{\chi,\chi}}{t(\|x+th\|_\chi + \|x\|_\chi)} \\ &= \lim_{t\to 0} \frac{2t\langle h, x\rangle_{\chi,\chi} + t^2\|h\|_\chi^2}{t(\|x+th\|_\chi + \|x\|_\chi)} \\ &= \lim_{t\to 0} \frac{2\langle h, x\rangle_{\chi,\chi} + t\|h\|_\chi^2}{\|x+th\|_\chi + \|x\|_\chi} \\ &= \frac{\langle h, x\rangle_{\chi,\chi}}{\|x\|_\chi}. \end{aligned}$$

– **Exemple 2 :** si $\chi = L^p([a,b],\mathbb{R})$ pour $-\infty < a < b < +\infty$ et $p \in (1,\infty)$, autrement dit χ est l'espace de Banach des fonctions $x : [a,b] \to \mathbb{R}$ ayant $\|x\|_\chi^p = \int_a^b |x(s)|^p ds < \infty$. Soit $h \in \chi$, la fonction signe (1.9), pour tout $x \in \chi, x \neq 0$, vaut :

$$\mathbf{SGN}_x(h) = \int_a^b \frac{sign(x(s)|x(s)|^{p-1}h(s)}{\|x\|_\chi^{p-1}} ds,$$

En effet, on considère la fonction $g(p) = \|x\|_\chi^p$. Elle est dérivable et sa dérivée est égale à $g'(p) = p\mathbf{SGN}_x\|x\|_\chi^{p-1}$, d'où

$$\mathbf{SGN}_x = \frac{g'(p)}{p\|x\|_\chi^{p-1}}.$$

Soit $h \in \chi$, on a :

$$\begin{aligned}
\{g'(p)\}(h) &= \lim_{t\to 0} \frac{\|x+th\|_\chi^p - \|x\|_\chi^p}{t} \\
&= \lim_{t\to 0} \frac{1}{t}\int_a^b (|x(s)+th(s)|^p - |x(s)|^p) ds \\
&= \lim_{t\to 0} \frac{1}{t}\int_a^b \left(\int_0^1 \frac{d}{dy}|x(s)+th(s)y|^p dy\right) ds \\
&= \lim_{t\to 0} p\int_a^b h(s)\Big(\int_0^1 sign\Big(x(s)+th(s)y\Big)\,|x(s)+th(s)y|^{p-1} dy\Big) ds \\
&= p\int_a^b sign\,(x(s))\,|x(s)|^{p-1}\,h(s) ds.
\end{aligned}$$

Par suite,

$$\begin{aligned}
\mathbf{SGN}_x(h) &= \frac{p\int_a^b sign\,(x(s))\,|x(s)|^{p-1}\,h(s) ds}{p\|x\|_\chi^{p-1}} \\
&= \int_a^b \frac{sign(x(s)|x(s)|^{p-1}h(s)}{\|x\|_\chi^{p-1}} ds.
\end{aligned}$$

1.3.1.2 Fonction de distribution spatiale

- **Dans le cadre univarié :** si on considère $X_1, \ldots, X_n$ un échantillon i.i.d de X qui est une variable aléatoire réelle, la fonction de distribution spatiale au point x, par rapport à la distribution de probabilité de X, est définie par :

$$S(x) = \mathbb{E}(sign(x - X)), \tag{1.11}$$

où $x \in \mathbb{R}$. Une extension de cette dernière a été étudiée, dans le cadre multivarié (dans $\mathbb{R}^d$), par Koltchinskii (1997). Une version empirique de

cette fonction est donnée par :

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} sign(x - X_i).$$

Nous remarquons que la fonction de distribution spatiale (1.11) satisfait

$$\mathbb{E}\Big(S(x)\Big) = 2F(x) - 1,$$

où F est la fonction de répartition de X. En effet,

$$\begin{aligned}\mathbb{E}\Big(S(x)\Big) &= \mathbb{E}\Big(sign(x - X)\Big) \\ &= \mathbb{E}(\mathbb{1}_{\{x>X\}}) - \mathbb{E}(\mathbb{1}_{\{x<X\}}) \\ &= \mathbb{P}(X < x) - 1 + \mathbb{P}(X < x) \\ &= 2F(x) - 1. \end{aligned} \tag{1.12}$$

On peut remarquer que les valeurs de x qui annulent $\mathbb{E}(S(x))$ et $S_n(x)$ sont respectivement la médiane et la médiane empirique.

- **Dans le cadre fonctionnel :** la distribution spatiale au point $x \in \chi$ (Chakraborty and Chaudhuri , 2014b) est définie à partir de la fonction signe (1.9), comme suit :

$$\mathbf{S}_x = \mathbb{E}(\mathbf{SGN}_{x-X}), \tag{1.13}$$

où X est un élement aléatoire (courbe) à valeurs dans l'espace χ définie précédemment. Cette dernière est égale, dans le cadre univarié, à (1.12). En prenant maintenant $X_1, \ldots, X_n$ des observations indépendantes de X, alors la distribution spatiale empirique (Chakraborty and Chaudhuri , 2014b) est donnée par :

$$\mathbf{S}_{n,x} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{SGN}_{x-X_i}, \tag{1.14}$$

où $x \in \chi$.

Remarque 1.2.

1. *Si χ est un espace de Hilbert, en utilisant la fonction (1.10), la fonction (1.13) est égale à :*

$$\boldsymbol{S}_x = \mathbb{E}\Big(\frac{x - X}{\|x - X\|_\chi}\Big).$$

et la fonction (1.14) vaut :

$$\boldsymbol{S}_{n,x} = \frac{1}{n}\sum_{i=1}^{n} \frac{x - X_i}{\|x - X_i\|_\chi}.$$

2. *Lorsque* $\boldsymbol{S}_x = \boldsymbol{0}$ *ceci implique que* x *est la médiane spatiale de* X *dans le cadre fonctionnel. Un estimateur empirique de la médiane spatiale de* X *vérifie l'équation suivante :*

$$\boldsymbol{S}_{n,x} = \boldsymbol{0}.$$

3. *Lorsque* $\boldsymbol{S}_x = \boldsymbol{u}$ *ceci implique que* x *est le* $\boldsymbol{u}$*-quantile spatial de* X *(voir, Chakraborty and Chaudhuri , 2014b), où* $\boldsymbol{u} \in \mathcal{B}^*(0,1)$ *qui répresente la boule unité ouverte de* χ^*.

4. *Il est bien connu, dans le cadre multivarié, qu'il y a une dépendance entre la profondeur spatiale et la distribution spatiale, dont les définitions sont données par Vardi and Zhang (2000) et Serfling (2002). Cette relation est définie dans le cadre fonctionnel, en utilisant un espace de Banach lisse* χ*, par :*

$$\boldsymbol{SD}(x) = 1 - \|\boldsymbol{S}_x\|_{\chi^*},$$

où $\boldsymbol{SD}(x)$ *est la profondeur spatiale pour des données fonctionnelles introduite dans la monographie de Chakraborty and Chaudhuri (2014b). Remarquons bien que, lorsque* $\boldsymbol{SD}(x) = 1$, x *devient la médiane spatiale de* X*. Une version empirique de cette profondeur spatiale est définie par :*

$$\boldsymbol{SD}_n(x) = 1 - \|\boldsymbol{S}_{n,x}\|_{\chi^*}.$$

1.3.2 Test de Wilcoxon-Mann-Whitney pour des données fonctionnelles

Dans ce paragraphe, nous présentons la construction de test de Wilcoxon-Mann-Whitney pour des données fonctionnelles introduit par Chakraborty and Chaudhuri (2015) et sa loi asymptotique sous l'hypothèse nulle définie ci-dessous. La statistique de ce test, basée sur la fonction signe définie précédemment, sera comparée à la statistique de test de la médiane pour des données fonctionnelles dans le chapitre 2 (voir l'étude de simulation). Elle sera aussi nécessaire, dans les chapitres 3 et 4, pour la construction d'une statistique de balayage spatial non paramétrique pour des données fonctionnelles.
Pour cela, nous considérons X et Y deux éléments aléatoires d'un espace de Banach lisse χ. Notons χ^* son espace dual. Soient $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ des observations indépendantes de X et Y venant de deux mesures de probabilité P et Q dans χ. On suppose que P et Q diffèrent par un paramètre de translation Δ. Dans ce cas, la statistique de Wilcoxon-Mann-Whitney (Chakraborty and Chaudhuri , 2014a), pour tester

$$H_0 : \Delta = 0 \text{ contre } H_1 : \Delta \neq 0,$$

est définie par

$$\text{WMW} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{SGN}_{Y_i - X_j}, \tag{1.15}$$

qui est à valeurs dans χ^*. On rejette l'hypothèse H_0 pour des grandes valeurs de $\|\text{WMW}\|_{\chi^*}$. Nous voyons que la statistique (1.15) est égale, à une constante près, à la statistique (1.3) dans le cadre univarié.

Remarque 1.3.

1. *La statistique (1.15) est égale, lorsque χ est un espace de Hilbert lisse, à*

$$WMW = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_i - X_j}{\|Y_i - X_j\|_\chi}.$$

   Pour plus de détails, voir Chakraborty and Chaudhuri (2015).

2. *Remarquons que la statistique WMW s'écrit comme étant une U-statistique (Borovskikh , 1996) ce qui facilite l'obtention de sa loi asymptotique en utilisant la décomposition de Hoeffding pour les U-statistiques à valeurs dans un espace de Banach (Borovskikh , 1996).*

3. *Nous voyons que WMW est un estimateur sans biais de*

$$\mu = \mathbb{E}(\boldsymbol{SGN}_{Y-X}). \tag{1.16}$$

   Sous H_0, on a $\mu = 0$ ce qui implique que la médiane spatiale de $Y - X$ est égale à $\boldsymbol{0}$.

Nous présentons maintenant la loi asymptotique de la statistique de Wilcoxon-Mann-Whitney définie par (1.15) sous H_0.

1.3.2.1 Loi asymptotique de la statistique WMW

Avant de commencer la détermination de la loi asympotitque, nous avons besoin de quelques notations et la définition d'un espace de Banach de type 2 suivantes :

Notations :

- On note $\mathbf{G}(m, C)$ la distribution gaussienne d'un élement aléatoire dans un espace de Banach séparable χ avec $m \in \chi$ est la moyenne et C est la covariance, où $C : \chi^* \times \chi^* \longrightarrow \mathbb{R}$ est une fonction bilinéaire continue définie positive et symétrique.

- On note $\Gamma_1, \Gamma_2 : \chi^{**} \times \chi^{**} \to \mathbb{R}$ deux fonctions bilinéaires continues définies positives et symétriques données par :

$$\Gamma_1(f, g) = \mathbb{E}\left[f(\mathbb{E}[\mathbf{SGN}_{Y-X}|X])g(\mathbb{E}[\mathbf{SGN}_{Y-X}|X])\right] - f(\mu)g(\mu), \tag{1.17}$$

   et

$$\Gamma_2(f, g) = \mathbb{E}\left[f(\mathbb{E}[\mathbf{SGN}_{Y-X}|Y])g(\mathbb{E}[\mathbf{SGN}_{Y-X}|Y])\right] - f(\mu)g(\mu), \tag{1.18}$$

   où, $f, g \in \chi^{**}$ et μ est définie par (1.16). Remarquons que, sous H_0, $\Gamma_1 = \Gamma_2$ et $\mu = 0$.

Definition 1.1. *(Espace de Banach de type 2)*
Un espace de Banach χ est dit de type 2 s'il existe une constante $b > 0$ telle que, pour tout $n \geq 1$ et $U_1, \ldots, U_n$ des éléments aléatoires de χ indépendants, de moyennes nulles et qui satisfont $\mathbb{E}(\|U_i\|_\chi^2) < \infty, \forall i = 1, \ldots, n$, on a :

$$\mathbb{E}(\|\sum_{i=1}^{n} U_i\|_\chi^2) \leq b \sum_{i=1}^{n} E(\|U_i\|_\chi^2).$$

Par suite, la normalité asymptotique de la statistique de test de Wilcoxon-Mann-Whitney pour des données fonctionnelles est donnée par :

Théorème 1.4. *(Loi asymptotique de WMW)*
Soient $N = m+n$ et $\frac{m}{N} \to \gamma \in (0,1)$, lorsque $m, n \to \infty$. Supposons que χ^* est un espace de Banach de type 2 séparable. Alors, pour toutes mesures de probabilités P et Q dans χ, on a :

$$\Big(\frac{mn}{N}\Big)^{1/2}(WMW - \mu) \quad \textit{converge faiblement vers} \quad \boldsymbol{G}\Big(0, (1-\gamma)\Gamma_1 + \gamma\Gamma_2\Big),$$

lorsque $m, n \to \infty$, où μ, Γ_1, Γ_2 sont définies respectivement par (1.16), (1.17) et (1.18).

Nous détaillerons ci-dessous la preuve de ce théorème car c'est en se basant sur cette démonstration que nous avons montré la normalité asymptotique de la statistique de la médiane introduite dans le chapitre 2.

Preuve du Théorème 1.4. La preuve de ce théorème est basée sur la décomposition de Hoeffding pour les U-statistiques à valeurs dans un espace de Banach (Borovskikh , 1996), la définition 1.1 et le théorème de la limite centrale pour des éléments aléatoires à valeurs dans un espace de Banach de type 2 (Araujo and Giné , 1980). Tout d'abord, considérons la fonction

$$u(x, y) = \mathbf{SGN}_{y-x} - \mu,$$

où μ est définie par (1.16). On a $\mathbb{E}(u(X, Y)) = 0$. On peut centrer la statistique de Wilcoxon-Mann-Whitney de la manière suivante :

$$\mathrm{WMW} - \mu = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} u(X_j, Y_i).$$

En faisant la décomposition de Hoeffding (Section 1.2 dans Borovskikh , 1996), on obtient

$$\mathrm{WMW} - \mu = \underbrace{\frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[u(X_j, Y)|X_j]}_{(a)} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[u(X, Y_i)|Y_i]}_{(b)} + R_{m,n}, \tag{1.19}$$

où

$$R_{m,n} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \tilde{u}(X_j, Y_i),$$

avec $\tilde{u}(x,y) = u(x,y) - \mathbb{E}[u(X,Y)|X=x] - \mathbb{E}[u(X,Y)|Y=y]$. Nous remarquons que la statistique WMW $-\ \mu$ se décompose comme étant la somme de trois sommes de variables aléatoires. Par conséquent, la démonstration de la normalité asymptotique de WMW $-\ \mu$ sera divisée en quatre étapes.

1. Montrer la convergence en probabilité du reste $R_{m,n}$ vers 0.
2. Montrer la convergence en loi de (a) vers un élément Gaussien.
3. Montrer la convergence en loi de (b) vers un élément Gaussien.
4. Conclure le résultat de la normalité asymptotique de la statistique de test WMW.

- **Etape 1.** La preuve ici est basée sur le fait que χ^* est un espace de Banach de type 2. Pour montrer la convergence en probabilité de $R_{m,n}$ vers 0, il suffit de prouver sa convergence en L^2 vers 0. Pour cela, on considère la fonction

$$\phi(X_j) = \sum_{i=1}^{n} \tilde{u}(X_j, Y_i).$$

Alors,

$$R_{m,n} = \frac{1}{mn} \sum_{j=1}^{m} \phi(X_j).$$

Remarquons que, sachant $Y_i, i = 1, \ldots, n$, $\tilde{u}(X_j, Y_i)$ est une suite de variables aléatoires indépendantes. De plus, on a : $\mathbb{E}[\tilde{u}(X,Y)|Y=y] = 0$ pour tout $y \in \chi$, en effet,

$$\begin{aligned}
\mathbb{E}[\tilde{u}(X,Y)|Y=y] &= \mathbb{E}[u(X,Y)|Y=y] - \mathbb{E}[\mathbb{E}[u(X,Y)|X=x]|Y=y] \\
&- \mathbb{E}[\mathbb{E}[u(X,Y)|Y=y]|Y=y] \\
&= \mathbb{E}[u(X,Y)|Y=y] - \mathbb{E}[u(X,Y)] - \mathbb{E}[u(X,Y)|Y=y] \\
&= 0.
\end{aligned}$$

Par conséquent, en utilisant le fait que χ^* est un espace de Banach de type 2 (définition (1.1)), on obtient

$$\begin{aligned}
\mathbb{E}\Big[\|R_{m,n}\|_{\chi^*}^2 \Big| Y_i, i = 1, \ldots, n\Big] &= \frac{1}{m^2 n^2} \mathbb{E}\left[\left\|\sum_{j=1}^{m} \phi(X_j)\right\|_{\chi^*}^2 \middle| Y_i, i = 1, \ldots, n\right] \\
&\leq \frac{b}{m^2 n^2} \sum_{j=1}^{m} \mathbb{E}\Big[\|\phi(X_j\|_{\chi*}^2 \Big| Y_i, i = 1, \ldots, n\Big].
\end{aligned}$$

En faisant la moyenne des deux côtés, on obtient

$$\mathbb{E}\left(\|R_{m,n}\|_{\chi^*}^2\right) \leq \frac{b}{mn^2} \mathbb{E}\left(\|\phi(X_1)\|_{\chi^*}^2\right) = \frac{b}{mn^2} \mathbb{E}\left(\left\|\sum_{i=1}^{n} \tilde{u}(X_1, Y_i)\right\|_{\chi^*}^2\right).$$

Or, sachant X_1, $\{\tilde{u}(X_1, Y_i); i = 1, \ldots, n\}$ est une suite de variables aléatoires indépendantes. De plus, $\mathbb{E}[\tilde{u}(X, Y)|X = x] = 0$ pour tout $x \in \chi$. Par conséquent, en utilisant le fait que χ^* est un espace de Banach de type 2, on obtient :

$$\begin{aligned}\mathbb{E}\left(\left\|\sum_{i=1}^{n}\tilde{u}(X_1, Y_i)\right\|_{\chi^*}^2\right) &= \mathbb{E}\left(\mathbb{E}\left[\left\|\sum_{i=1}^{n}\tilde{u}(X_1, Y_i)\right\|_{\chi^*}^2\middle|X_1\right]\right)\\ &\leq b\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{E}\left[\|\tilde{u}(X_1, Y_i)\|_{\chi^*}^2\middle|X_1\right]\right)\\ &\leq bn\mathbb{E}\left(\|\tilde{u}(X_1, Y_1)\|_{\chi^*}^2\right).\end{aligned}$$

Or, en utilisant le fait que $\|\mathbf{SGN}_{Y_1 - X_1}\|_{\chi^*} \leq 1$, on obtient $\|\tilde{u}(X_1, Y_1)\|_{\chi^*} \leq 4$. Par conséquent,

$$\mathbb{E}\left(\|R_{m,n}\|_{\chi^*}^2\right) \leq \frac{16b^2}{mn}$$

donc

$$\mathbb{E}\left(\left\|\left(\frac{mn}{N}\right)^{1/2} R_{m,n}\right\|_{\chi^*}^2\right) \leq \frac{16b^2}{N} \underset{m,n\to\infty}{\longrightarrow} 0. \tag{1.20}$$

Finalement, d'après (1.20), on obtient

$$\left(\frac{mn}{N}\right)^{1/2} R_{m,n} \quad \text{converge en probabilité vers} \quad 0, \tag{1.21}$$

lorsque $m, n \to \infty$.

- **Etape 2.** Pour montrer la convergence en loi de (a) vers un élément Gaussien, il nous faut vérifier les trois conditions du Corollaire 7.8 de Araujo and Giné (1980). Pour cela, on considère

$$\psi_N(X_j)) = m^{-1/2}\mathbb{E}[u(X_j, Y)|X_j], \quad \forall j = 1, \ldots, m$$

une suite d'éléments aléatoires i.i.d et centrés à valeurs dans l'espace χ^* qui est un espace de Banach de type 2. Alors, pour appliquer le Corollaire 7.8 de Araujo and Giné (1980), il suffit de vérifier les trois conditions que nous traiterons ci-dessous.

  - **Condition 1.** Montrons que, pour tout $\epsilon > 0$, on a

$$\lim_{m\to\infty}\sum_{j=1}^{m}\mathbb{P}\left(\|\psi_N(X_j)\|_{\chi^*} > \epsilon\right) = 0.$$

    En utilisant l'inégalité de Markov, on obtient

$$\begin{aligned}\sum_{j=1}^{m}\mathbb{P}\left(\|\psi_N(X_j)\|_{\chi^*} > \epsilon\right) &\leqslant \sum_{j=1}^{m}\frac{\mathbb{E}\left(\left\|\mathbb{E}\left[\mathbf{SGN}_{Y-X_j}\middle|X_j\right] - \mu\right\|_{\chi^*}^3\right)}{m^{3/2}}\\ &\leqslant \frac{2^3 m}{m^{3/2}} = 8m^{-1/2} \underset{m\to+\infty}{\longrightarrow} 0.\end{aligned}$$

- **Condition 2.** Montrons que, pour tout $f \in \chi^{**}$,

$$\lim_{m\to\infty} \sum_{j=1}^{m} \mathbb{E}\left(f^2\Big(\psi_N(X_j) - \mathbb{E}\left(\psi_N(X_j)\right)\Big)\right) = \Gamma_1(f,f) < \infty,$$

où Γ_1 est une fonction à déterminer.
Pour cela, on considère $f \in \chi^{**}$, où χ^{**} est l'espace bidual de χ. En utilisant la linéarité de f, on obtient :

$$\begin{aligned}
\mathbb{E}\Big(f^2\Big(\psi_N(X_j) - \mathbb{E}\left(\psi_N(X_j)\right)\Big)\Big) &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\Big(f^2\Big(\mathbb{E}[u(X_j,Y)|X_j]\Big)\Big) \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\Big(\Big(f\Big(\mathbb{E}\Big[\mathbf{SGN}_{Y-X_j}\Big|X_j\Big]\Big) - f(\mu)\Big)^2\Big) \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left(\Big(Z_{N,j} - \mathbb{E}(Z_{N,j})\Big)^2\right),
\end{aligned}$$

où $Z_{N,j} = f\Big(\mathbb{E}\Big[\mathbf{SGN}_{Y-X_j}\Big|X_j\Big]\Big)$. En utilisant le fait que les X_j pour $j = 1, \ldots, m$ sont identiquement distribuées, on obtient :

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\Big(\Big(Z_{N,j} - \mathbb{E}(Z_{N,j})\Big)^2\Big) &= \mathbb{E}\left(\Big(Z_{N,1} - E(Z_{N,1})\Big)^2\right) \\
&= \mathbb{E}(Z_{N,1}^2) - \mathbb{E}^2(Z_{N,1}) \\
&= \mathbb{E}\Big(f^2\Big(\mathbb{E}[\mathbf{SGN}_{Y-X_1}|X_1]\Big)\Big) - f^2(\mu) \\
&= \Gamma_1(f,f) < \infty,
\end{aligned}$$

où Γ_1 est définie par (1.17).

- **Condition 3.** Pour vérifier cette condition, nous considérons $\{\mathcal{F}_k\}_{k\geq 1}$ une suite de sous espaces de dimensions finies de χ^* telles que $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ et l'adhérence de $\bigcup_{k=1}^{\infty} \mathcal{F}_k$ est égale à χ^*. Pour tout $x \in \chi^*$ et $k \geq 1$, on définit la fonction $x \mapsto d(x, \mathcal{F}_k) = \inf\{\|x - y\| : y \in \mathcal{F}_k\}$. Montrons que

$$\lim_{k\to+\infty} \overline{\lim_{m\to\infty}} \sum_{j=1}^{m} \mathbb{E}\left(d^2\Big(\psi_N(X_j) - \mathbb{E}(\psi_N(X_j); \mathcal{F}_k\Big)\right) = 0.$$

L'existence de la suite $\{\mathcal{F}_k\}_{k\geq 1}$ vient du fait que l'espace χ^* est supposé séparable. On a $\mathbb{E}(\psi_N(X_j)) = 0$, alors

$$\begin{aligned}
\sum_{j=1}^{m} \mathbb{E}\left(d^2\Big(\psi_N(X_j), \mathcal{F}_k\Big)\right) &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left(d^2\Big(\mathbb{E}\Big[\mathbf{SGN}_{Y-X_j}\Big|X_j\Big] - \mu, \mathcal{F}_k\Big)\right) \\
&= \mathbb{E}\left(d^2\Big(\mathbb{E}[\mathbf{SGN}_{Y-X_1}|X_1] - \mu, \mathcal{F}_k\Big)\right).
\end{aligned}$$

Or, d'après la construction de la suite $\mathcal{F}_k$, on a, pour tout $x \in \chi^*$, $d(x, \mathcal{F}_k) \to 0$ lorsque $k \to \infty$. Par suite,

$$\lim_{k\to+\infty} \overline{\lim_{m\to\infty}} \sum_{j=1}^{m} \mathbb{E}\left(d^2\Big(\psi_N(X_j), \mathcal{F}_k\Big)\right) = 0.$$

Par conséquent, les trois conditions du Corollaire 7.8 de Araujo and Giné (1980) sont vérifiées, ce qui implique que

$$m^{1/2}(a) = \sum_{j=1}^{m} \psi_N(X_j) \quad \text{converge faiblement en loi vers} \quad \mathbf{G}(0, \Gamma_1)$$

lorsque $m, n \to \infty$. Par suite,

$$\left(\frac{mn}{N}\right)^{1/2}(a) \quad \text{converge faiblement en loi vers} \quad \mathbf{G}(0, (1-\gamma)\Gamma_1) \tag{1.22}$$

lorsque $m, n \to \infty$.

- **Etape 3.** Pour montrer la convergence en loi de (b) vers un élément Gaussien, il suffit de suivre la même procédure que celle utilisée dans l'étape 2, en considérant dans cette étape

$$\varphi_N(Y_i) = n^{-1/2}\mathbb{E}[u(X, Y_i)|Y_i], \quad \forall i = 1, \ldots, n,$$

une suite d'éléments aléatoires i.i.d et centrés à valeurs dans l'espace χ^*. Donc, il suffit juste de vérifier les trois conditions du Corollaire 7.8 de Araujo and Giné (1980) comme précédemment.

  - **Condition 1.** Montrons que, pour tout $\epsilon > 0$, on a

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}\left(\|\varphi_N(Y_i)\|_{\chi^*} > \epsilon\right) = 0.$$

  En utilisant l'inégalité de Markov, on obtient

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{P}\left(\|\varphi_N(Y_i)\|_{\chi^*} > \epsilon\right) &\leqslant \sum_{i=1}^{n} \frac{\mathbb{E}\left(\|\mathbb{E}[\mathbf{SGN}_{Y_i - X}|Y_i] - \mu\|_{\chi^*}^3\right)}{m^{3/2}} \\ &\leqslant \frac{2^3 n}{n^{3/2}} = 8n^{-1/2} \underset{n\to+\infty}{\longrightarrow} 0.\end{aligned}$$

  - **Condition 2.** Montrons que, pour tout $f \in \chi^{**}$,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left(f^2\Big(\varphi_N(Y_i) - \mathbb{E}\left(\varphi_N(Y_i)\right)\Big)\right) = \Gamma_2(f, f) < \infty,$$

où Γ_2 est une fonction à déterminer.

En utilisant la linéarité de f, on obtient :

$$\mathbb{E}\left(f^2\Big(\varphi_N(Y_i) - \mathbb{E}\left(\varphi_N(Y_i)\right)\Big)\right) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left(\Big(W_{N,i} - \mathbb{E}(W_{N,i})\Big)^2\right),$$

où $W_{N,i} = f\Big(\mathbb{E}[\mathbf{SGN}_{Y_i - X}|Y_i]\Big)$. En utilisant le fait que les Y_i pour $i = 1, \ldots, n$ sont identiquement distribuées, on obtient :

$$\frac{1}{m}\sum_{i=1}^{n}\mathbb{E}\Big(\Big(W_{N,i} - \mathbb{E}(W_{N,i})\Big)^2\Big) = \mathbb{E}\left(\Big(W_{N,1} - E(W_{N,1})\Big)^2\right) = \Gamma_2(f,f) < \infty,$$

où Γ_2 est définie par (1.18).

- **Condition 3.** Considérons $\{\mathcal{F}_k\}_{k\geq 1}$ une suite de sous espaces de dimensions finies de χ^* telles que $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ et l'adhérence de $\bigcup_{k=1}^{\infty}\mathcal{F}_k$ est égale à χ^*. Pour tout $x \in \chi^*$ et $k \geq 1$, on définit la fonction $x \mapsto d(x, \mathcal{F}_k) = \inf\{\|x - y\| : y \in \mathcal{F}_k\}$. Montrons que

$$\lim_{k\to+\infty}\overline{\lim_{n\to\infty}}\sum_{i=1}^{n}\mathbb{E}\left(d^2\Big(\varphi_N(Y_i) - \mathbb{E}(\varphi_N(Y_i); \mathcal{F}_k\Big)\right) = 0.$$

On a $\mathbb{E}(\varphi_N(Y_i)) = 0$. Par suite,

$$\sum_{i=1}^{n}\mathbb{E}\left(d^2\Big(\varphi_N(Y_i), \mathcal{F}_k\Big)\right) \;=\; \mathbb{E}\left(d^2\Big(\mathbb{E}[\mathbf{SGN}_{X-Y_1}|Y_1] - \mu, \mathcal{F}_k\Big)\right)$$

or, d'après la construction de la suite $\mathcal{F}_k$, on trouve le résultat d'une manière similaire à la condition 3 de la deuxième étape.

Par conséquent, en utilisant le Corollaire 7.8 de Araujo and Giné (1980), on obtient

$$n^{1/2}(b) = \sum_{i=1}^{n}\varphi_N(Y_i) \quad \text{converge faiblement en loi vers} \quad \mathbf{G}(0, \Gamma_2)$$

lorsque $m, n \to \infty$. Par suite,

$$\left(\frac{mn}{N}\right)^{1/2}(b) \quad \text{converge faiblement en loi vers} \quad \mathbf{G}(0, \gamma\Gamma_2) \tag{1.23}$$

lorsque $m, n \to \infty$.

- **Etape 4.** En utilisant (1.19) et en multipliant par $(\frac{mn}{N})^{1/2}$, nous obtenons la décomposition suivante :

$$\left(\frac{mn}{N}\right)^{1/2}(\mathrm{WMW} - \mu) = \left(\frac{mn}{N}\right)^{1/2}(a) + \left(\frac{mn}{N}\right)^{1/2}(b) + \left(\frac{mn}{N}\right)^{1/2}R_{m,n}.$$

En utilisant (1.21), (1.22), (1.23), l'indépendence des deux échantillons $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$, et le Lemme de Slutsky, nous obtenons le résultat escompté, c'est-à-dire

$$\left(\frac{mn}{N}\right)^{1/2} (\mathrm{WMW} - \mu) \quad \text{converge faiblement en loi vers} \quad \mathbf{G}\Big(0, (1-\gamma)\Gamma_1 + \gamma\Gamma_2\Big).$$

□

Chapter 2

A median test for functional data

"Beautiful light is born of darkness, so the faith that springs from conflict is often the strongest and the best"

— Ross Turnbull

Chapter contents

Abstract

The median test is more powerful than the Student t-test and the Wilcoxon-Mann-Whitney test in heavy-tailed cases for univariate data. When dealing with multidimensional data, the multivariate extension of the median test, called the sign test, is more efficient than the Hotelling T^2 and the Wilcoxon-Mann-Whitney tests for high dimensions and in very heavy-tailed cases. In this paper, we construct a median type test based on spatial ranks for functional data, i.e in infinite dimensional space, and

we obtain asymptotic results. Then, we compare the proposed functional median test with the existing ones using simulated and real functional data.

2.1 Introduction

Statistical hypothesis testing plays an essential role in statistics (Lehmann , 1986; Lehmann and Romano , 2005). In nonparametric statistics, tests of hypotheses are known as nonparametric or distribution-free tests. It is not necessary to assume hypotheses on the shape of the distribution and estimate its parameters. These tests can be used to verify that two or more datasets come from identical populations.
Here, we will focus on this type of tests to solve the two-sample location problem which received a considerable attention in the past. More specifically, we consider the known two-sample problem with independent observations

$$\begin{aligned} X_1, \ldots, X_m &\sim F \\ Y_1, \ldots, Y_n &\sim G, \end{aligned}$$

where F and G are continuous distributions functions. Then, we only focus on the situation where the distribution function G is considered as a shifted version of F, i.e. $G(\cdot) = F(\cdot - \Delta)$. In this case, the null hypothesis of equality of F and G can be expressed as $H_0 : \Delta = 0$ against the alternative one which can be $\Delta \neq 0$.
For univariate data, Wilcoxon (1945) and Mann and Whitney (1947) proposed nonparametric tests based on ranks. Each of them defined their own test statistic which leads to the same test named Wilcoxon-Mann-Whitney. This test is more powerful than the Student's t-test for various non-Gaussian distributions (Blair and Higgins , 1980) and it is also asymptotically optimum in case of a logistic type density (Hàjek et al. , 1999). Another test of hypothesis of the location problem is assigned to Mood (1950) and it is called the median test. Another version of this test based on ranks was presented in Van Der Vaart (1998). This version of test is an asymptotically optimum in the case of a double exponential distribution (Capéraà and Cutsem , 1988; Hàjek et al. , 1999). In fact, it is based on a statistic which counts the number of individuals from the second sample exceeding the median of the pooled sample unlike the Wilcoxon-Mann-Whitney test statistic which uses the sum of ranks of the second sample in the pooled sample. Nowadays, the median test is not often used because it is less powerful than the Wilcoxon-Mann-Whitney test when applied to Gaussian distributions (Mood , 1954). However, this test is more efficient, when dealing with symmetrical distributions with heavy-tails, than the Wilcoxon-Mann-Whitney one (Capéraà and Cutsem , 1988).
For multivariate data, several versions of the Hotelling, Wilcoxon-Mann-Whitney and median tests have been studied. See, for example, Puri and Sen (1971), Chakraborty and Chaudhuri (1999), Oja and Randles (2004), Oja (2010). The extension of univariate two-sample Mood test is called the sign test and it has

the best efficiency in very heavy-tailed cases and for high dimensions : it also outperforms the Hotelling test in heavy-tailed cases (Oja and Randles , 2004).
Currently, the development of the sensoring and computing tools allows us to work with huge datasets. So, we have more and more access to data of functional type, for example the functional chemometric data and the electricity consumption of different regions (Ferraty and Vieu , 2006; Kokoszka and Reimherr , 2017; Ramsay and Silverman , 2005). These kinds of data are not real random variables or vectors but they are a collection of random elements like curves, surfaces, images, etc, and each sample variable is usually considered as a function. The main particularity of such data is the infinite dimension of the data space such as Banach and Hilbert spaces. Appropriate statistical tools are necessary to handle these types of data. More details on the treatment of functional data, suitable mathematical backgrounds and definitions of centrality and dispersion parameters and their specific estimations in this context are introduced in several monographs such as Cuevas (2014), Chakraborty and Chaudhuri (2014b), Goia and Vieu (2016), Aneiros et al. (2019) in the parametric case. Nonparametric techniques have been also handled in the literature (Geenens , 2015; Ling and Vieu , 2018), mainly for modelling and regressing these specific data. Moreover, as in the univariate and multivariate cases, two-sample tests of hypotheses are also adapted to functional data using parametric and nonparametric methods.
In the parametric case, to decide whether two samples of curves are issued from the same distribution, Horváth et al. (2013) proposed two test statistics for testing the equality of mean functions. These two test statistics are based on the orthogonal projections on the space generated by the eigenfunctions of an L^2-consistent estimator. This estimator is obtained from the asymptotic covariance operator of the difference between the two-sample mean functions. Among these two tests, one is the same as the Hotelling statistic in finite dimensional space. Still using a parametric approach, Cuevas et al. (2004) introduced an analog of the classical one-way analysis of variance (ANOVA) problem for functional data. In a nonparametric setting, Chakraborty and Chaudhuri (2015) (see, also Chakraborty and Chaudhuri , 2014a) proposed a Wilcoxon-Mann-Whitney test based on spatial ranks. Their statistic is an extension of the one defined for example in Hàjek et al. (1999) and Van Der Vaart (1998) for real valued random variables.
Our goal here is to construct an extension of the median test for processes valued in infinite dimensional Banach spaces. The rest of this paper is organised as follows: in Section 2.2, we propose a median test statistic based on spatial ranks in Banach space and especially in separable Hilbert space. We also introduce a modified and more simple version of this statistic. Then, we study the asymptotic behavior of the latter one under the null hypothesis and we proceed a random permutation method to implement the test. To illustrate our theoretical results, we compare in Section 2.3 the performance of the proposed test with various other tests, either parametric or nonparametric, using simulated and real functional datasets. We conclude with a discussion of the methods and results.

2.2 The construction of the test

2.2.1 The introduction of the median statistics

First we recall what median tests look like in the univariate case. Let X and Y be two $\mathbb{R}$-valued random variables. We consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ two independent random samples of X and Y with distribution functions F and F_θ respectively, such that $\forall x \in \mathbb{R}; F_\theta(x) = F(x - \theta)$. The constant θ is called *the translation parameter*.
The median test statistic based on ranks (Capéraà and Cutsem , 1988) for testing the hypothesis

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0$$

is defined as

$$T_{\text{Mo}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{R_i > 0\}},$$

where $R_i = 1 + (\sum_{j=1}^{m} \mathbb{1}_{\{Y_i > X_j\}} + \sum_{k=1}^{n} \mathbb{1}_{\{Y_i > Y_k\}} - \frac{N+1}{2})$ is the centered rank of Y_i when $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are ordered together in the same sample of size $N = n + m$.
This test is based on the number of observations of $Y_1, \ldots, Y_n$ that is strictly greater than the global median of the N observations.
We consider also the following test statistic:

$$T'_{\text{Mo}} = \frac{1}{n} \sum_{i=1}^{n} \text{sign}(R_i), \tag{2.1}$$

where the sign function is $x \longmapsto \text{sign}(x) = \begin{cases} \frac{x}{|x|} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$.
These two test statistics are equivalent and related to the same test since $T'_{\text{Mo}} = 2\text{T}_{\text{Mo}} - 1$.
To make things easier afterwards, we introduce a test statistic which counts the number of the observations $Y_1, \ldots, Y_n$ that are greater than the median of the observations $X_1, \ldots, X_m$ instead of the global median. In other words, in the univariate case it is equal to

$$T_{\text{MED}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\hat{F}_m(Y_i) > \frac{1}{2}\}},$$

where $\hat{F}_m(x) = 1/m \sum_{j=1}^{m} \mathbb{1}_{\{X_j \leq x\}}$ is the empirical distribution function of $X_1, \ldots, X_m$.
This statistic is inspired from the work of Koul and Staudte (1972).
Our goal here is to construct an extension of T_{MED} in infinite dimensional space.

2.2.1.1 T_{MED} in the functional case

We shall now consider X and Y two independant random elements in a Banach space χ. We denote by χ^* its dual space, i.e., the space of the linear continuous

functions on χ with values in $\mathbb{R}$, and χ^{**} its bidual space, i.e., the space of the linear continuous functions on χ^* with values in $\mathbb{R}$. Now, we suppose that:

- The space χ is smooth, i.e., the norm function $\|.\|_\chi$ is Gateaux differentiable at each $x \neq 0, x \in \chi$. We denote by $\mathbf{SGN}_x \in \chi^*$ its Gateaux derivative. This sign function is defined, for all $h \in \chi$, as

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

- The space χ^* is a smooth, i.e, the norm function $\|.\|_{\chi^*}$ is Gateaux differentiable at each $y \neq 0, \in \chi^*$. We denote by $\mathbf{SGN}_y^* \in \chi^{**}$ its Gateaux derivative. This sign function is defined, for all $H \in \chi^*$, as

$$\mathbf{SGN}_y^*(H) = \begin{cases} \lim_{t\to 0} \frac{\|y+tH\|_{\chi^*} - \|y\|_{\chi^*}}{t} & \text{if } \ y \neq 0 \\ 0 & \text{if } \ y = 0 \end{cases}.$$

Also, we consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ independent random samples of X and Y from two probability measures P and Q on χ. We suppose that P and Q differ by a shift $\Delta \in \chi$.
Then, for testing

$$H_0 : \Delta = 0 \ \text{ against } \ H_1 : \Delta \neq 0,$$

the statistic T_{MED} becomes

$$\begin{aligned} \text{MED} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{SGN}^*\left\{\tfrac{1}{m}\sum_{j=1}^{m} \mathbf{SGN}_{\{Y_i - X_j\}}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \phi(F_m(Y_i)), \end{aligned} \tag{2.2}$$

where

* $\phi : u \mapsto \phi(u) = \mathbf{SGN}_u^*$ is a map on $\chi^* \backslash \{0\}$ (see, e.g., Corollary 4.2.12 in Borwein and Vanderwerff , 2010) that is continuous if the norm on χ^* is twice Gateaux differentiable. This sign function was used to proof asymptotic properties of the spatial depth in infinite dimensional spaces. See Theorem 4.1 in Chakraborty and Chaudhuri (2014b) for more details.

* $F_m : y \mapsto F_m(y) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{SGN}_{\{y - X_j\}}$ is the empirical spatial distribution associated to the i.i.d observations $X_1, \ldots, X_m$ (see, Chakraborty and Chaudhuri , 2014b). This empirical spatial distribution has been used to develop the Wilcoxon-Mann-Whitney type test for two-sample problems in infinite dimensional spaces (Chakraborty and Chaudhuri , 2015). Remark that, in the univariate case, the empirical spatial distribution is equal to $2\hat{F}_m(y) - 1$ where $\hat{F}_m$ is the empirical distribution function of $X_1, \ldots, X_m$.

∗ We will also denote by $F(y)$ the spatial distribution of X at $y \in \chi$ which is equal to $\mathbb{E}\left[\mathbf{SGN}_{\{y-X\}}\right]$. For more details, see Chakraborty and Chaudhuri (2014b). Note that, in the univariate case, the spatial distribution is equal to $2\tilde{F}(y) - 1$ where $\tilde{F}$ is the distribution function of X.

As stated by Chakraborty and Chaudhuri (2014a), when the space χ is assumed to be an Hilbert one, sign functions become simpler so that $\mathbf{SGN}_x = \frac{x}{\|x\|_\chi}$ and $\mathbf{SGN}^*_y = \frac{y}{\|y\|_{\chi^*}}$. Thus, we might rewrite the statistic MED as follows:

$$\text{MED} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{m} \frac{Y_i - X_j}{\|Y_i - X_j\|_\chi}}{\left\| \sum_{j=1}^{m} \frac{Y_i - X_j}{\|Y_i - X_j\|} \right\|_\chi}. \tag{2.3}$$

2.2.1.2 T'_{Mo} in the functional case

An extension of T'_{Mo} defined as (2.1) in the functional case can be written as

$$\text{Mo} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{SGN}^* \left\{ \sum_{k=1}^{n} \mathbf{SGN}_{Y_i - Y_k} + \sum_{j=1}^{m} \mathbf{SGN}_{Y_i - X_j} \right\}.$$

When χ is assumed to be an Hilbert space, the statistic Mo becomes

$$\text{Mo} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=1, k\neq i}^{n} \frac{Y_i - Y_k}{\|Y_i - Y_k\|_\chi} + \sum_{j=1}^{m} \frac{Y_i - X_j}{\|Y_i - X_j\|_\chi}}{\left\| \sum_{k=1, k\neq i}^{n} \frac{Y_i - Y_k}{\|Y_i - Y_k\|_\chi} + \sum_{j=1}^{m} \frac{Y_i - X_j}{\|Y_i - X_j\|_\chi} \right\|_\chi}. \tag{2.4}$$

2.2.2 Asymptotic distribution of MED

In this section, we study the asymptotic normality of MED.
First, we introduce the following notations which will be used later:

- We denote by $\mathbf{G} := G(\mathbf{m}, \mathbf{C})$ the distribution of a Gaussian random element (say, $\mathbf{G}$) in a separable Banach space χ with mean $\mathbf{m} \in \chi$ and covariance $\mathbf{C}$, where $\mathbf{C} : \chi^* \times \chi^* \to \mathbb{R}$ is a symmetric nonnegative definite continuous bilinear function. For all $\mathbf{l} \in \chi^*$, $\mathbf{l}(\mathbf{G})$ has a Gaussian distribution on $\mathbb{R}$ with mean $\mathbf{l}(\mathbf{m})$ and variance $\mathbf{C}(\mathbf{l}, \mathbf{l})$.
- For all $x, y \in \chi$, define

$$F_X(y) = \mathbb{E}\left[\mathbf{SGN}_{\{y-X\}} \middle| y\right], \tag{2.5}$$

$$F_Y(x) = \mathbb{E}\left[\mathbf{SGN}_{\{Y-x\}} \middle| x\right]. \tag{2.6}$$

These two functions are used to prove the theorem of the asymptotic normality of the Wilcoxon-Mann-Whitney test statistic under finite and shrinking locations shifts (Chakraborty and Chaudhuri , 2015). Moreover, we denote

$$\mu = \mathbb{E}\left[\mathbf{SGN}^*_{\{F_X(Y)\}}\right]$$

and

$$\tilde{\mu} = \mathbb{E}\left[\mathbf{SGN}^*_{\{F_Y(X)\}}\right].$$

Remark that, under H_0, we have $\mu = \tilde{\mu}$.

- Let $\Gamma_1, \Gamma_2 : \chi^{***} \times \chi^{***} \to \mathbb{R}$ be the symmetric positive definite continuous bilinear operators defined as :

$$\Gamma_1(f,g) = \mathbb{E}\left[f\left(\mathbf{SGN}^*_{\{F_X(Y)\}}\right) g\left(\mathbf{SGN}^*_{\{F_X(Y)\}}\right)\right] - f(\mu)g(\mu) \tag{2.7}$$

and

$$\Gamma_2(f,g) = \mathbb{E}\left[f\left(\mathbf{SGN}^*_{\{F_Y(X)\}}\right) g\left(\mathbf{SGN}^*_{\{F_Y(X)\}}\right)\right] - f(\tilde{\mu})g(\tilde{\mu}), \tag{2.8}$$

where $f, g \in \chi^{***}$.

For our next theorem, we shall also consider the following assumptions and definition:

Assumption 2.1. *We assume that the norm in χ^* is twice Gateaux differentiable at every $x \neq 0$.*

From Assumption 2.1 (see also, e.g., Chapter 4, Section 6 in Borwein and Vanderwerff , 2010), let $\mathbf{J}_x : \chi^* \to \chi^{**}$ denote, when it exists, the Hessian of the function $g : x \mapsto \mathbb{E}\Big[\|F_X(Y) + x\|_{\chi^*}\Big|X_1, \ldots, X_m\Big]$, $x \in \chi^*$. In particular, if we assume that χ is an Hilbert space, then χ^* is also an Hilbert one. Since the norms in Hilbert spaces are twice Gateaux differentiable (page 6 in Chakraborty and Chaudhuri , 2014a), and if $Z = F_X(Y)$, the derivative of the map g is defined as :

$$\nabla_x g = \mathbb{E}\Big[\mathbf{SGN}^*_{\{Z+x\}}\Big|X_1, \ldots, X_m\Big] = \mathbb{E}\left[\frac{Z+x}{\|Z+x\|_{\chi^*}}\middle|X_1, \ldots, X_m\right]$$

and its Hessian is:

$$\begin{array}{rcccccc} \mathbf{J}_x & : & \chi^* & \to & \chi^{**} & & \\ & & h & \mapsto & \mathbf{J}_x(h) & : \chi^* \to & \mathbb{R} \\ & & & & & v \mapsto & \{\mathbf{J}_x(h)\}(v) := \langle \mathbf{J}_x(h), v\rangle. \end{array}$$

Then, we have

$$\mathbf{J}_x = \mathbb{E}\left[\frac{1}{\|Z+x\|_{\chi^*}}\left(\mathbf{I}_{\chi^*} - \frac{(Z+x)\otimes(Z+x)}{\|Z+x\|^2_{\chi^*}}\right)\middle|X_1, \ldots, X_m\right],$$

where $\mathbf{I}_{\chi^*}$ is the identity operator in χ^* and $u \otimes v(h) = \langle u, h\rangle.v$ for all $h, v \in \chi^*$. Thus,

$$\mathbf{J}_x(h) = \mathbb{E}\left[\frac{h}{\|Z+x\|_{\chi^*}} - \frac{\langle Z+x, h\rangle\,(Z+x)}{\|Z+x\|^3_{\chi^*}}\middle| X_1, \ldots, X_m\right],$$

for all $h \in \chi^*$. More explicitly, $\mathbf{J}_x$ is given by

$$\{\mathbf{J}_x(h)\}(v) = \langle \mathbf{J}_x(h), v\rangle = \mathbb{E}\left[\frac{1}{\|Z+x\|_{\chi^*}}\left(\langle h, v\rangle - \frac{\langle Z+x, h\rangle\langle Z+x, v\rangle}{\|Z+x\|^2_{\chi^*}}\right)\middle| X_1, \ldots, X_m\right],$$

for all $h, v \in \chi^*$.

Assumption 2.2. *The Hessian operator $\boldsymbol{J}_x$ defined as above exists for all $x \in \chi^*$ and there is a constant $c > 0$ such that*

$$\|\boldsymbol{J}_0\| \leq c.$$

Definition 2.1. *(Banach space of type 2)*
A Banach space χ is said to be of type 2 if there is a constant $b > 0$ such that for any $n \geq 1$ and independent zero mean random elements $V_1, \ldots, V_n$ in χ satisfying $\mathbb{E}(\|V_i\|^2) < \infty$, for all $i = 1, \ldots, n$, we have

$$\mathbb{E}(\|\sum_{i=1}^n V_i\|^2) \leq b\sum_{i=1}^n \mathbb{E}(\|V_i\|^2).$$

Then, the asymptotic normality of MED is given by the following theorem.

Theorem 2.1. *(Asymptotic Gaussianity of MED)*
Let $N = m + n$ and $m/N \to \lambda \in (0, 1)$ as $m, n \to \infty$. Assume that the bidual χ^{**} space is a separable and type 2 Banach space. Then, under Assumptions 2.1 and 2.2, for any two probability measures P and Q on χ,

$$(mn/N)^{1/2}(MED - \mu) \quad \textit{converges weakly to} \quad G\left(0, \lambda\Gamma_1 + (1-\lambda)\Gamma_2\right)$$

as $m, n \to \infty$.
The proof of Theorem 2.1 is available in Appendix 2.5.

Remark. *For easier understanding, we develop μ in the univariate case. As defined before,*

$$\mu = \mathbb{E}\left[\boldsymbol{SGN}^*_{\{F_X(Y)\}}\right] = \mathbb{E}\left[\boldsymbol{SGN}^*_{\left\{\mathbb{E}\left[\boldsymbol{SGN}_{\{Y-X\}}|Y\right]\right\}}\right].$$

Thus,

$$\begin{aligned} F_X(Y) &= \mathbb{E}\Big[\boldsymbol{SGN}_{\{Y-X\}}\Big|Y\Big] \\ &= \mathbb{E}\Big[\mathbb{1}_{\{Y>X\}}\Big|Y\Big] - \mathbb{E}\Big[\mathbb{1}_{\{Y<X\}}\Big|Y\Big] \\ &= 2\mathbb{E}\Big[\mathbb{1}_{\{Y>X\}}\Big|Y\Big] - 1 \\ &= 2\tilde{F}_X(Y) - 1, \end{aligned}$$

where $\tilde{F}_X(\cdot)$ is the conditional distribution function of X given Y i.e. the projection of $\mathbb{1}_{\{X<\cdot\}}$ onto the subspace spanned by Y. Then, we obtain

$$\begin{aligned}
\mu &= \mathbb{E}\left[sign(2\tilde{F}_X(Y)-1)\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\left\{2\tilde{F}_X(Y)-1>0\right\}}\right] - \mathbb{E}\left[\mathbb{1}_{\left\{2\tilde{F}_X(Y)-1<0\right\}}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\left\{F_X(Y)>\frac{1}{2}\right\}}\right] - \mathbb{E}\left[\mathbb{1}_{\left\{F_X(Y)<\frac{1}{2}\right\}}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\left\{Y>\tilde{F}_X^{-1}\left(\frac{1}{2}\right)\right\}}\right] - \mathbb{E}\left[\mathbb{1}_{\left\{Y<\tilde{F}_X^{-1}\left(\frac{1}{2}\right)\right\}}\right] \\
&= 1 - 2\mathbb{E}\left[\mathbb{1}_{\left\{Y<\tilde{F}_X^{-1}\left(\frac{1}{2}\right)\right\}}\right] \\
&= 1 - 2\,G\left(\tilde{F}_X^{-1}\left(1/2\right)\right),
\end{aligned}$$

where G is the distribution function of Y. Under H_0, we have $\mu = 0$.

2.2.3 Computing the significance

The implementation of the test based on the statistic introduced in the previous section can be done using the asymptotic distribution of the statistic under the null hypothesis. Since $\mu = 0$ under H_0, we decide to reject the null hypothesis if $\left\|(mn/N)^{1/2}\mathrm{MED}\right\| > s_\alpha$, where s_α denote the $(1-\alpha)$ quantile of the limiting distribution $\|G\left(0, \lambda\Gamma_1 + (1-\lambda)\Gamma_2\right)\|$. However, using this theoretical quantile suffers two limitations: the need to estimate the covariance operators Γ_1 and Γ_2 and the distance to the asymptotic distribution when m and n are quite small, which is often the case when comparing two samples of functions.
Consequently, we decided to use a test procedure based on Monte-Carlo simulations allowing to give an approximation of the null distribution (Dwass , 1957). The procedure is based on the following steps:

1. Let $X_{\text{obs}} = (X_1, \ldots, X_m)$ and $Y_{\text{obs}} = (Y_1, \ldots, Y_n)$. Among the $m+n$ observations of $(X_{\text{obs}}, Y_{\text{obs}})$, m of them are randomly chosen to create X_{perm} and the n others to create Y_{perm}.

2. The simulated median statistic S_{perm} is then computed using X_{perm} and Y_{perm} instead of X and Y.

3. Based on n_{perm} random permutations, the *p-value* of the median statistic is given by
$$p_{value} = \frac{1 + \sum_{l=1}^{n_{\text{perm}}} \mathbb{1}_{\left\{S_{\text{perm}}^{(l)} > S\right\}}}{n_{\text{perm}} + 1},$$
where S is the value of the statistic computed using the observed data and $S_{perm}^{(l)}$ the value computed using the l^{th} random permutation.

4. We will reject H_0 when the *p-value* is below the level of the test.

2.3 Applications

2.3.1 A simulation study

In this section, we aim to compare the power of the two median statistics proposed in the precedent section with those of the tests available in Chakraborty and Chaudhuri (2015), Cuevas et al. (2004) and Horváth et al. (2013).
We set $\chi = L^2[0,1]$. The test studied by Chakraborty and Chaudhuri (2015) is based on the Wilcoxon-Mann-Whitney statistic, which is defined as a U-statistic (Borovskikh , 1996) like in the univariate case and can be rewrited in χ as follows

$$\text{WMW} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{SGN}_{Y_i - X_j} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_i - X_j}{\|Y_i - X_j\|_\chi}.$$

The test statistic studied by Cuevas et al. (2004) is defined by

$$\text{CFF} = \frac{m\|\bar{X} - \mu_g\|_\chi^2 + n\|\bar{Y} - \mu_g\|_\chi^2}{\frac{1}{N-2}\left(\sum_{j=1}^{m}\|X_j - \bar{X}\|_\chi^2 + \sum_{i=1}^{n}\|Y_i - \bar{Y}\|_\chi^2\right)},$$

where $\bar{X}$ and $\bar{Y}$ are the empirical mean of the X_j's and Y_i's respectively for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$ and μ_g is the empirical mean of the pooled sample of the X_j's and Y_i's. Horváth et al. (2013) studied the test statistics defined as follows

$$\text{HKR1} = \frac{mn}{N}\sum_{l=1}^{p}\frac{\langle \bar{X} - \bar{Y}, \hat{\varphi}_l\rangle}{\hat{\lambda}_l},$$

and

$$\text{HKR2} = \frac{mn}{N}\sum_{l=1}^{p}\langle \bar{X} - \bar{Y}, \hat{\varphi}_l\rangle.$$

Here, $\langle ., . \rangle$ denotes the inner product in χ, the $\hat{\lambda}_l$'s are the eigenvalues of the empirical pooled covariance of the X_j's and the Y_i's in descending order of magnitude and the $\hat{\varphi}_l$'s are the corresponding empiricial eigenvectors. In this section, we have used the usual empirical pooled covariance for the two tests HKR1 and HKR2 of Horváth et al. (2013) and the numbers of projection directions p are chosen using the cumulative variance method described in their paper. In the multivariate case, when $\chi = \mathbb{R}^d$ and $p = d$, HKR2 reduces to Hotelling's T^2 statistic.
Now, let consider the decomposition

$$X = \sum_{k=1}^{\infty} Z_k e_k,$$

where for all $k \geqslant 0$, $e_k = \sqrt{2}sin(t/\sigma_k)$ is an orthonormal basis of χ, $\sigma_k = ((k - 0.5)\pi)^{-1}$ and Z_k's are independant random variables which correspond to the projection of X on the Karhunen-Loève basis (Karhunen , 1947; Lévy and Loève , 1948). We have considered four scenarios:

(i) A standard Brownian motion (sBm) i.e., Z_k/σ_k follows a $\mathcal{N}(0,1)$ distribution.

(ii) A centered t process on $[0,1]$ with 5 degrees of freedom i.e., $Z_k/\sigma_k \backsim t(5)$.

(iii) A Cauchy distribution with parameters 0 and 1 i.e. $Z_k/\sigma_k \backsim \mathcal{C}(0,1)$.

(iv) A double exponential distribution with parameters 0 and 1 i.e., $Z_k/\sigma_k \backsim \mathcal{D}exp(0,1)$.

The scenarios (i) and (ii) are studied in Chakraborty and Chaudhuri (2015) and we have chosen the scenarios (iii) and (iv) to study the performance of the differents tests using more heavy-tailed distributions.
Assume that Y is distributed as $X + \Delta$ and under the alternative hypotheses $H_1 : \Delta \neq 0$. We have chosen $m = n = 10$ and each sample of curve is observed at 100 equispaced points in $[0,1]$. Three choices are considered, namely : $\Delta_1(t) = c$, $\Delta_2(t) = ct$ and $\Delta_3(t) = ct(1-t)$ where $c > 0$ for all $t \in [0,1]$. Figure 2.1 shows examples of simulated data.

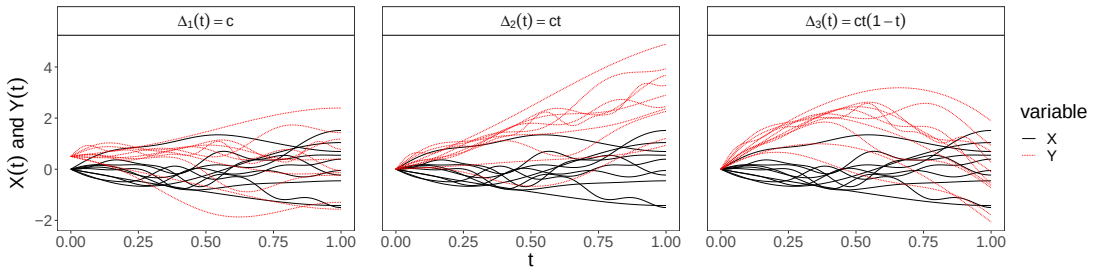

Figure 2.1: Examples of generated data using the scenario (i) with $c = 0.5$ (left panel), $c = 3$ (middle panel) and $c = 8$ (right panel). In black: 10 samples of X. In red: 10 samples of Y.

For each simulated dataset, all the test statistics and their critical values are derived in the same way as in subsection 2.2.3. The power of the statistics is estimated using $n_{sim} = 1000$ random simulations of (X, Y) for each simulation. The hypothesis H_0 is rejected if $p_{value} < \alpha$, where α is the significance level which is chosen equal to 0.05.
Before computing the powers of the tests previously introduced, we have derived the size of each of these tests, i.e. the probability of rejecting the null hypothesis if it is true. Then, we obtain the results gathered in Table 2.1.

Sizes \ Distributions	$N(0,1)$	$t(5)$	$\mathcal{C}(0,1)$	$\mathcal{D}exp(0,1)$
Size(Mo)	0.045	0.045	0.041	0.051
Size(MED)	0.045	0.05	0.045	0.053
Size(WMW)	0.046	0.049	0.048	0.052
Size(HKR1)	0.046	0.056	0.043	0.061
Size(HKR2)	0.047	0.052	0.043	0.052
Size(CFF)	0.044	0.052	0.048	0.048

Table 2.1: Sizes of all the tests using the different scenarios (i), (ii), (iv) and (v).

We see in Table 2.1 that the sizes of all the tests (when $c = 0$) considered as above are close to the nominal 5% level for all the distributions. Figure 2.2, Figure 2.3 and Figure 2.4 show the corresponding power results.

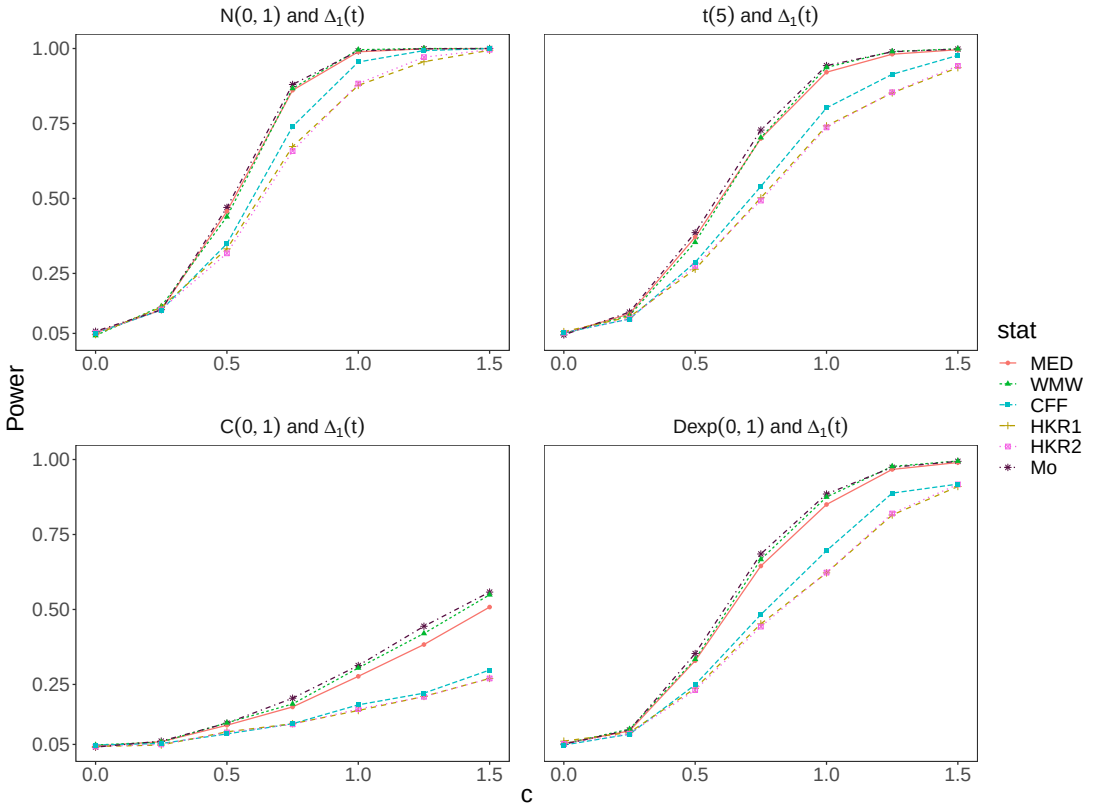

Figure 2.2: The power results for MED, Mo, WMW, CFF, HKR1 and HKR2 when $\Delta_1(t) = c$, $n_{\text{perm}} = 999$, $n_{\text{sim}} = 1000$ and $n = m = 10$ in the different scenarios (i), (ii), (iii) and (iv).

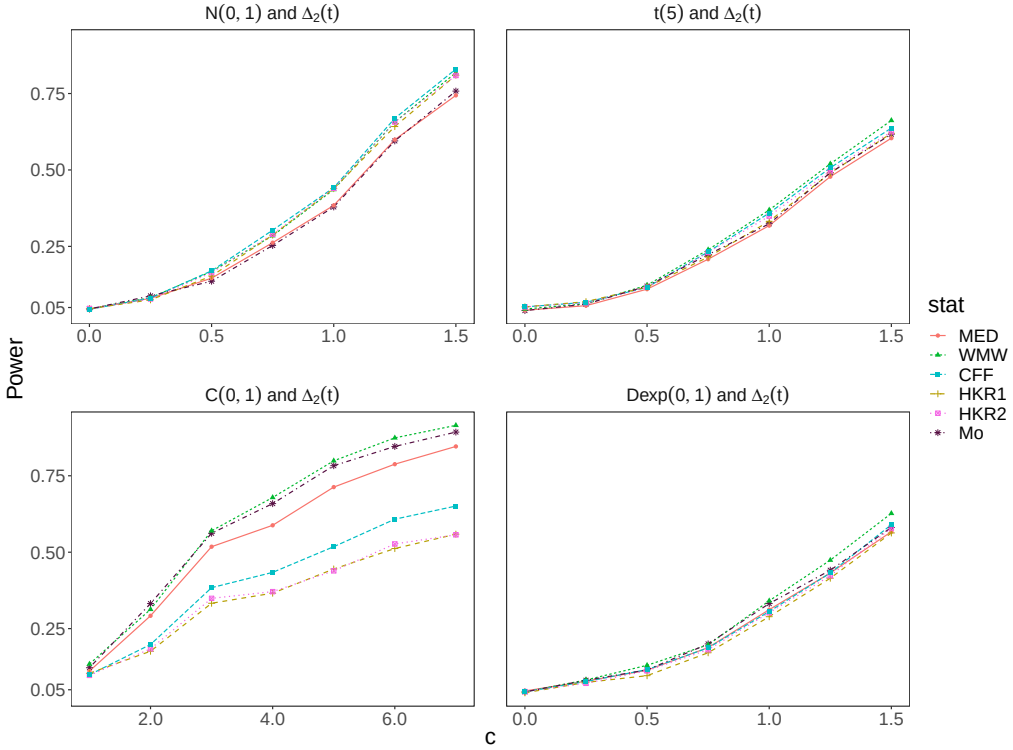

Figure 2.3: The power results for MED, Mo, WMW, CFF, HKR1 and HKR2 when $\Delta_2(t) = ct$, $n_{\text{perm}} = 999$, $n_{\text{sim}} = 1000$ and $n = m = 10$ in the different scenarios (i), (ii), (iii) and (iv).

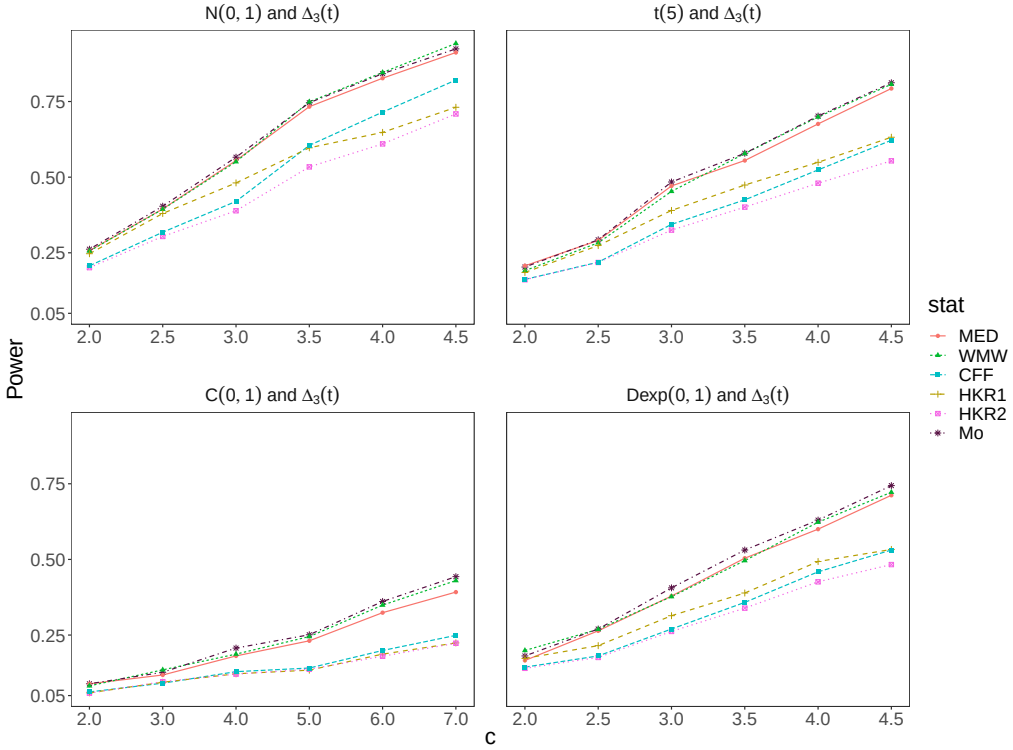

Figure 2.4: The power results for MED, Mo, WMW, CFF, HKR1 and HKR2 when $\Delta_3(t) = ct(1-t)$, $n_{\text{perm}} = 999$, $n_{\text{sim}} = 1000$ and $n = m = 10$ in the different scenarios (i), (ii), (iii) and (iv).

From the previous figures, we can say that:

- For $\Delta_1(t)$, the tests based on Mo, WMW and MED have similar powers under all the distributions considered except the Cauchy one where the test based on MED is less powerful for large values of c. We can see also in Figure 2.2 that the tests using the CFF, HKR1 and HKR2 statistics are less powerful than the ones based on MED, Mo and WMW for large values of c using the different distributions.

- For $\Delta_2(t)$ and under all the distributions except the Cauchy one, all the tests have a similar power for small values of c (see, Figure 2.3). For large values of c using the sBm process, the test based on CFF outperforms all the other tests although using the Student process, the one based on WMW is more powerful than the other ones. We notice that, using the heavy-tailed distributions $\mathcal{C}(0,1)$ and $\mathcal{D}exp(0,1)$, the Mo and the MED outperfoms the parametric tests CFF, HKR1 and HKR2.

- For $\Delta_3(t)$, the test using Mo statistic outperforms the test based on WMW against heavy-tailed distributions $t(5)$, $\mathcal{C}(0,1)$ and $\mathcal{D}exp(0,1)$ for large values of c and it has a similar power against sBm distribution (see, Figure 2.4). We remark also that the power of the test based on MED is nearly in conformity with the power of the tests based on WMW and Mo for small values of the shift c and for all the distributions. However, for large values of c, it has a similar power than the tests based on WMW and Mo for all the distributions except the $\mathcal{C}(0,1)$ one. We can see also in Figure 2.4 that using the four distributions, the nonparametric tests based on MED, Mo and WMW performs better than the parametric ones based on CFF, HKR1 and HKR2.

2.3.2 An application to real data

In this section, we compare the two median tests based on MED and Mo with those based on WMW, CFF, HKR1 and HKR2 which are presented in the previous section using two datasets already analysed by Chakraborty and Chaudhuri (2014a) (for more details, see Ramsay and Silverman , 2005 and Ferraty and Vieu , 2006). In all these datasets, each observation can be observed as an element in the separable Hilbert space $\chi = L^2[a,b]$.

2.3.2.1 Coffee data

This dataset can be dowloaded from http://www.cs.ucr.edu/~eamonn/time_series_data/. It contains the spectroscopy values for 14 samples of each of the two different types of coffee beans (Arabica and Robusta) recorded at 286 wavelengths (see Figure 2.5).

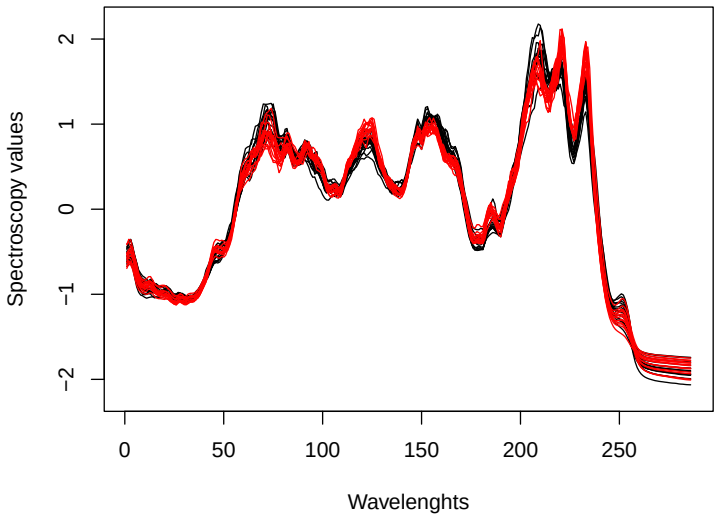

Figure 2.5: In red: Spectroscopy curves of Robusta beans. In black: Spectroscopy curves of Arabica beans.

Based on 999 random permutations, all the p-values obtained on this dataset are equal to 0.001 except the one obtained using the test based on HKR2 which is equal to 0.003. This leads us to reject the null hypothesis. From the Figure 2.5, the spectroscopy curves of the two coffee types are clearly different since the maximum values are not observed in the same wavelengths for Arabica and Robusta. However, in the paper of Chakraborty and Chaudhuri (2014a) the p-values, based on the asymptotic distributions, of the tests based on WMW (0.072), CFF (0.169), HKR1 (0.273) and HKR2 (0.273) fail to reject H_0. We suppose that the small size of this dataset ($n = m = 14$) does not allow the asymptotic results to be relevant in that case.

2.3.2.2 Berkeley growth data

This dataset is available in the R package fda and contains the heights of 39 boys and 54 girls measured at 31 time points from age 1 to 18 (see Figure 2.6).

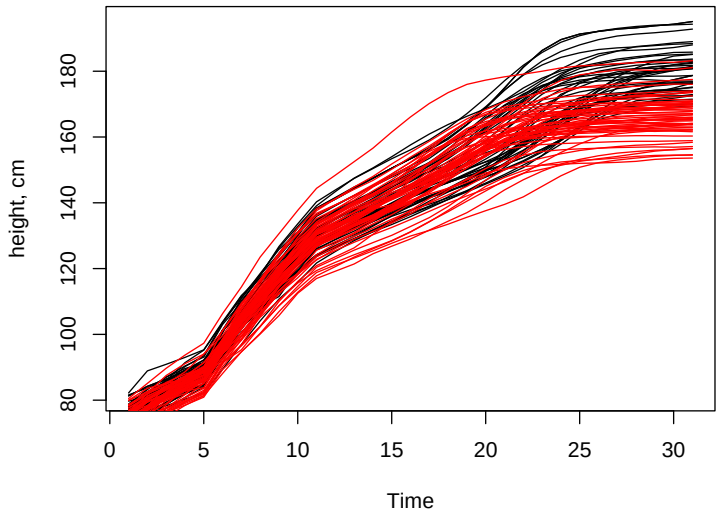

Figure 2.6: In red: Heights of the 54 girls. In black: Heights of the 39 boys.

All the p-values based on 999 random permutations are 0 up to two decimal places which are similar to the results obtained in the article of Chakraborty and Chaudhuri (2014a). These p-values show a strong difference between the two distributions. So, we decided to evaluate the proportion of rejection of the null hypothesis to compare the behaviour of the different statistics when the level α is equal to 0.05. Such as done by Chakraborty and Chaudhuri (2014a), we have chosen randomly 20% subsamples of the 2 classes of the complete dataset and this subsampling was repeated 100 times. Results are given in Table 2.2.

Statistic	MED	Mo	WMW	CFF	HKR1	HKR2
Proportion of rejection	1	1	0.99	0.87	0.28	0.25

Table 2.2: The proportions of rejection of the null hypothesis of the different test statistics MED, Mo, WMW, CFF, HKR1 and HKR2.

As seen in Table 2.2, for this type of data, our tests based on Mo and MED statistics have the highest rate of rejection of the null hypothesis, close to the one obtained using Chakraborty and Chaudhuri (2015) test and much larger than the ones based on tests proposed by Cuevas et al. (2004) and Horváth et al. (2013).

2.4 Discussion

Nowadays and with the development of modern technology, scientists often observe functional datasets instead of multivariate ones. As a consequence, there is a need

for testing procedures adapted to these infinite dimensional data. In this paper, we have proposed an extension of an existing nonparametric test based on ranks in the infinite dimensional spaces to compare two datasets (samples of curves).
We introduce a median test in the functional case, similar to the one rank-based test presented in Capéraà and Cutsem (1988) and Van Der Vaart (1998) in the univariate one. It can be noted that we introduce the notion of ranking functional elements through a sign function. The median test is one way to use this sign function for ranking functional elements but other possibilities have been investigated such as functional depth (Chakraborty and Chaudhuri , 2014b; Estévez-Pérez and Vieu , 2021; Gijbels and Nagy , 2017): these could lead to other types of nonparametric tests for comparing samples of functions that we shall investigate in a future work.
First, we proposed two median statistics in a Banach space then their equivalent in a particular case which is an Hilbert space. Second, we obtained in this paper the asymptotic Gaussianity of one of the median statistics proposed in Section 2.2. Remark that our test statistic is not a U-statistic (Borovskikh , 1996) such as Wilcoxon-Mann-Whitney one and it is based on two sign functions instead of one, which increases the complexity of the proof of the corresponding asymptotic Gaussianity. Then, the application to simulated and real data shows that the median tests have good perfomance compared to the Wilcoxon-Mann-Whitney one proposed by Chakraborty and Chaudhuri (2015), the ANOVA test based on CFF and the mean-based tests introduced by Horváth et al. (2013).
A perspective would be to develop the tests introduced in this paper and the Wilcoxon-Mann-Whitney one to compare more than two datasets in infinite dimensional space. To do so, we may follow the strategy used by Oja (2010) in the multivariate case: this becomes a multiple location test.
Recently, Smida et al. (2021) have proposed a nonparametric spatial scan statistic for detecting spatial clusters using functional data. This scan statistic was constructed using the Wilcoxon-Mann-Whitney two sample test for functional data (Chakraborty and Chaudhuri , 2015). Another perspective would be to develop a new nonparametric scan statistic in the functional case using the statistics introduced in this paper.

2.5 Appendix – Proof of theorem

Proof of theorem 2.1. The median statistic (2.2) is

$$\begin{aligned} \text{MED} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{SGN}^{*}\left\{\frac{1}{m}\sum_{j=1}^{m}\mathbf{SGN}_{\{Y_i-X_j\}}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\phi(F_m(Y_i)). \end{aligned}$$

We remark that the random elements $\mathbf{SGN}_{\{Y_i-X_j\}}$ are not independent for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,m\}$. To ensure the independence, our strategy is to

use the conditional expectations $F_X(y)$ and $F_Y(x)$ defined respectively by (2.5) and (2.6) for all $x, y \in \chi$.
Thus, we can use the following decomposition :

$$\text{MED} - \mathbb{E}[\text{MED}|X_j; j = 1, \ldots, m] = L_n^{'} + L_m^{''} + K_{m,n}^{'},$$

where

$$L_n^{'} = \frac{1}{n} \sum_{i=1}^{n} \left[\phi(F_X(Y_i)) - \mathbb{E}(\phi(F_X(Y_i)))\right], \tag{2.9}$$

$$L_m^{''} = \frac{1}{m} \sum_{j=1}^{m} \left[\phi(F_Y(X_j)) - \mathbb{E}(\phi(F_Y(X_j)))\right] \tag{2.10}$$

and

$$K_{m,n}^{'} = \text{MED} - \mathbb{E}[\text{MED}|X_j; j = 1, \ldots, m] - L_n^{'} - L_m^{''}. \tag{2.11}$$

Remark that the same decomposition is used in Section 14.1.1 of Van Der Vaart (1998) for $\mathbb{R}$-valued random variables. Furthermore, consider

$$K_{m,n}^{''} = \mathbb{E}[\text{MED}|X_j; j = 1, \ldots, m] - \mathbb{E}[\phi(F_X(Y))]. \tag{2.12}$$

Hence, we can write the following decomposition :

$$\text{MED} - \mu = L_n^{\prime} + L_m^{''} + R_{m,n}^{\prime}, \tag{2.13}$$

where

$$R_{m,n}^{\prime} = K_{m,n}^{'} + K_{m,n}^{''}.$$

The proof of the asymptotic distribution of MED can be split into four steps :

- **Step 1:** Show that $L_n^{\prime}$ converges in *law* to a Gaussian element.
- **Step 2:** Show that $L_m^{''}$ converges in *law* to a Gaussian element.
- **Step 3:** Show that $R_{m,n}^{\prime}$ converges in *probability* to 0.
- **Step 4:** Conclude the asymptotic normality of MED.

2.5.1 Step 1 : Asymptotic behavior of $L_n^{\prime}$

Let

$$L_n^{'} = \frac{1}{n} \sum_{i=1}^{n} \left[\phi(F_X(Y_i)) - \mathbb{E}(\phi(F_X(Y_i)))\right].$$

We want to prove here the asymptotic Gaussianity of $L_n^{'}$. For this purpose, for all $i = 1, \ldots, n$, let us write the sequence

$$\begin{aligned} \psi_n(Y_i) &= n^{-1/2} \left[\phi(F_X(Y_i)) - \mathbb{E}(\phi(F_X(Y_i)))\right] \\ &= n^{-1/2} \left[\mathbf{SGN}^*_{\left\{\mathbb{E}\left[\mathbf{SGN}_{\{Y_i - X\}}|Y_i\right]\right\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\left\{\mathbb{E}\left[\mathbf{SGN}_{\{Y_i - X\}}|Y_i\right]\right\}}\right]\right]. \end{aligned}$$

Note that $\mathbb{E}\left[\psi_n(Y_i)\right] = 0$. In order to show the asymptotic Gaussianity of $\sum_{i=1}^n \psi_n(Y_i)$, we check that the triangular array $\{\psi_n(Y_1), \ldots, \psi_n(Y_n)\}_{n=1}^{\infty}$ of row-wise independent and identically distributed random elements satisfies the three conditions of Corollary 7.8 in Araujo and Giné (1980).

- **Condition 1 :** Let us show that

$$\forall \epsilon > 0, \lim_{n\to+\infty} \sum_{i=1}^{n} \mathbb{P}\left(\|\psi_n(Y_i)\|_{\chi^{**}} > \epsilon\right) = 0.$$

Using the Bienaymé-Tchebychev inequality, we obtain : for any $\epsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{P}\left(\|\psi_n(Y_i)\|_{\chi^{**}} > \epsilon\right) &\leq \sum_{i=1}^{n} \frac{\mathbb{E}\left[\left\|\mathbf{SGN}^*_{\{F_X(Y_i)\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right]\right\|^3_{\chi^{**}}\right]}{\epsilon^3 n^{3/2}} \\ &\leq \frac{8}{\epsilon^3 n^{1/2}} \underset{n\to+\infty}{\longrightarrow} 0.\end{aligned}$$

For the last inequality, we have used the fact that $\left\|\mathbf{SGN}^*_{\{x\}}\right\|_{\chi^{**}} \leq 1$ for all $x \in \chi^*$. Thus, the first condition of the Corollary 7.8 in Araujo and Giné (1980) holds.

- **Condition 2 :** Let us show that

$$\forall f \in \chi^{***}, \lim_{n\to+\infty} \sum_{i=1}^{n} \mathbb{E}\left[f^2\left(\psi_n(Y_i) - \mathbb{E}(\psi_n(Y_i))\right]\right] = \Gamma_1(f,f) < \infty.$$

Le us fix $f \in \chi^{***}$. Since f is linear, we may write

$$\begin{aligned}&\sum_{i=1}^{n} \mathbb{E}\left[f^2\left[\psi_n(Y_i) - \mathbb{E}(\psi_n(Y_i))\right]\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[f^2\left[\psi_n(Y_i)\right]\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f^2\left(\mathbf{SGN}^*_{\{F_X(Y_i)\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right]\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(f\left(\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right) - f\left(\mathbb{E}\left[\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right]\right)\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(f\left(\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right) - \mathbb{E}\left[f\left(\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right)\right]\right)^2\right].\end{aligned}$$

We consider now, for all $i = 1, \ldots, n$,

$$W_i := f\left(\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right) = f\left(\mathbf{SGN}^*_{\{\mathbb{E}\left[\mathbf{SGN}_{\{Y_i - X\}}\middle| Y_i\right]\}}\right).$$

Hence, we get

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[f^2\left[\psi_n(Y_i)\right]\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left[W_i-\mathbb{E}\left(W_i\right)\right]^2\right]\\ &= \mathbb{E}\left[\left[W_1-\mathbb{E}\left(W_1\right)\right]^2\right]\\ &= \mathbb{E}\left[W_1^2\right]-\mathbb{E}^2\left[W_1\right]\\ &= \Gamma_1(f,f)<\infty,\end{aligned}$$

where Γ_1 is defined as (2.7). Thus, the second condition of the Corollary 7.8 in Araujo and Giné (1980) holds.

- **Condition 3 :** Let us show that

$$\lim_{k\to+\infty}\overline{\lim_{n\to+\infty}}\sum_{i=1}^{n}\mathbb{E}\left[d^2\left(\psi_n(Y_i)-\mathbb{E}\left[\psi_n(Y_i)\right],\mathcal{F}_k\right)\right]=0,$$

where $\{\mathcal{F}_k\}_{k\geq 1}$ is a sequence of finite dimensional subspaces of χ^{**} such that $\mathcal{F}_k\subseteq\mathcal{F}_{k+1}$ for all $k\geq 1$ and the closure of $\cup_{k=1}^{\infty}\mathcal{F}_k$ is equal to χ^{**}. This sequence exists because of the separability of χ^{**}. Also, for any $x\in\chi^{**}$ and any $k\geq 1$, we define $d(x,\mathcal{F}_k)=\inf\{\|x-y\|_{\chi^{**}}:y\in\mathcal{F}_k\}$. It is easy to prove that for all $k\geq 1$, the map $x\longmapsto d(x,\mathcal{F}_k)$ is continuous and bounded on any closed ball in χ^{**}.
First, since $\mathbb{E}\left[\psi_n(Y_i)\right]=0$, we have

$$\begin{aligned}&\sum_{i=1}^{n}\mathbb{E}\left[d^2\left(\psi_n(Y_i),\mathcal{F}_k\right)\right]\\ =\;&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[d^2\left(\mathbf{SGN}^*_{\{F_X(Y_i)\}}-\mathbb{E}\left[\mathbf{SGN}^*_{\{F_X(Y_i)\}}\right],\mathcal{F}_k\right)\right]\\ =\;&\mathbb{E}\left[d^2\left(\mathbf{SGN}^*_{\{F_X(Y_1)\}}-\mathbb{E}\left[\mathbf{SGN}^*_{\{F_X(Y_1)\}}\right],\mathcal{F}_k\right)\right].\end{aligned}$$

From the choice of the sequence $\{\mathcal{F}_k\}_{k\leq 1}$, we obtain $d(x,\mathcal{F}_k)\longrightarrow 0$ as $k\to\infty$ for any $x\in\chi^{**}$. Thus,

$$\lim_{k\to+\infty}\overline{\lim_{n\to+\infty}}\sum_{i=1}^{n}\mathbb{E}\left[d^2\left(\psi_n(Y_i),\mathcal{F}_k\right)\right]=0$$

and the third condition of the Corollary 7.8 in Araujo and Giné (1980) holds.

Therefore, using Corollary 7.8 in Araujo and Giné (1980), $\sum_{i=1}^{n}\psi_n(Y_i)$ converges weakly to a centered Gaussian random element in χ^{**} as $m,n\to\infty$. Moreover, the asymptotic covariance is Γ_1 which was obtained while checking the second condition presented as above. Finally,

$$\sqrt{n}L'_n=n^{-1/2}\sum_{i=1}^{n}\left[\phi(F_X(Y_i))-\mathbb{E}\left[\phi(F_X(Y_i))\right]\right]\xrightarrow{\mathcal{L}}\mathbf{G}(0,\Gamma_1)\tag{2.14}$$

weakly as $m,n\to\infty$.

2.5.2 Step 2 : Asymptotic behavior of $L_m^{''}$

Let

$$L_m^{''} = \frac{1}{m}\sum_{j=1}^{m}\left[\phi(F_Y(X_j)) - \mathbb{E}(\phi(F_Y(X_j)))\right]. \tag{2.15}$$

Similarly to the previous step, our goal in this part is to prove the asymptotic Gaussianity of $L_m^{''}$. To do that, for all $j = 1, \ldots, m$, we consider

$$\begin{aligned}\tilde{\psi}_m(X_j) &= m^{-1/2}\left[\phi(F_Y(X_j)) - \mathbb{E}(\phi(F_Y(X_j)))\right] \\ &= m^{-1/2}\left[\mathbf{SGN}^*_{\left\{\mathbb{E}\left[\mathbf{SGN}_{\{Y-X_j\}}\middle|X_j\right]\right\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\left\{\mathbb{E}\left[\mathbf{SGN}_{\{Y-X_j\}}\middle|X_j\right]\right\}}\right]\right].\end{aligned}$$

To show the asymptotic Gaussianity of $\sum_{j=1}^{m}\tilde{\psi}_m(X_j)$, we will use the same procedure as in step 1, replacing the array $\{\psi_n(Y_1),\ldots,\psi_n(Y_n)\}_{n=1}^{\infty}$ by $\{\tilde{\psi}_m(X_1),\ldots,\tilde{\psi}_m(X_m)\}_{m=1}^{\infty}$. Note that $\mathbb{E}\left[\tilde{\psi}_m(X_j)\right] = 0$. Now,we shall check that the triangular array $\{\tilde{\psi}_m(X_1),\ldots,\tilde{\psi}_m(X_m)\}_{m=1}^{\infty}$ of rowwise independent and identically distributed random elements also satisfies the three conditions of Corollary 7.8 in Araujo and Giné (1980).

- **Condition 1 :** Let us show that

$$\forall\epsilon>0, \lim_{m\to+\infty}\sum_{j=1}^{m}\mathbb{P}\left(\left\|\tilde{\psi}_m(X_j)\right\|_{\chi^{**}} > \epsilon\right) = 0.$$

  Observe that for any $\epsilon > 0$,

$$\begin{aligned}\sum_{j=1}^{m}\mathbb{P}\left(\left\|\tilde{\psi}_m(X_j)\right\|_{\chi^{**}} > \epsilon\right) &\leq \sum_{j=1}^{m}\frac{\mathbb{E}\left[\left\|\mathbf{SGN}^*_{\{F_Y(X_j)\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right]\right\|^3_{\chi^{**}}\right]}{\epsilon^3 m^{3/2}} \\ &\leq \frac{8}{\epsilon^3 m^{1/2}} \underset{m\to+\infty}{\longrightarrow} 0.\end{aligned}$$

  For the last inequality, we have used the fact that $\left\|\mathbf{SGN}^*_{\{x\}}\right\|_{\chi^{**}} \leq 1$ for all $x \in \chi^*$. Thus, the first condition of the Corollary 7.8 in Araujo and Giné (1980) holds.

- **Condition 2 :** Let us show that

$$\forall f\in\chi^{***}, \lim_{m\to+\infty}\sum_{j=1}^{m}\mathbb{E}\left[f^2\left(\tilde{\psi}_m(X_j) - \mathbb{E}(\tilde{\psi}_m(X_j)\right)\right] = \Gamma_2(f,f) < \infty.$$

Le us fix $f \in \chi^{***}$. Since f is linear and $\mathbb{E}(\tilde{\psi}_m(X_j)) = 0$, for all $j = 1, \ldots, m$, we can write

$$\begin{aligned}
\sum_{j=1}^{m} \mathbb{E}\left[f^2\left[\tilde{\psi}_m(X_j)\right]\right] &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left[f^2\left(\mathbf{SGN}^*_{\{F_Y(X_j)\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right]\right)\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left[\left(f\left(\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right) - f\left(\mathbb{E}\left[\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right]\right)\right)^2\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left[\left(f\left(\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right) - \mathbb{E}\left[f\left(\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right)\right]\right)^2\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left[\left[v_j - \mathbb{E}\left(v_j\right)\right]^2\right] \\
&= \mathbb{E}\left[\left[v_1 - \mathbb{E}\left(v_1\right)\right]^2\right] \\
&= \mathbb{E}\left[v_1^2\right] - \mathbb{E}^2\left[v_1\right] \\
&= \Gamma_2(f, f) < \infty,
\end{aligned}$$

where $v_j := \mathbf{SGN}^*_{\{F_Y(X_j)\}} = \mathbf{SGN}^*_{\{\mathbb{E}\left[\mathbf{SGN}_{\{Y - X_j\}}\middle| X_j\right]\}}$, for all $j = 1, \ldots, m$ and Γ_2 is defined as (2.8). Thus, the second condition of the Corollary 7.8 in Araujo and Giné (1980) holds.

- **Condition 3 :** Let us show that

$$\lim_{k \to +\infty} \overline{\lim_{m \to +\infty}} \sum_{j=1}^{m} \mathbb{E}\left[d^2\left(\tilde{\psi}_m(X_j) - \mathbb{E}\left[\tilde{\psi}_m(X_j)\right], \mathcal{F}_k\right)\right] = 0,$$

where $\{\mathcal{F}_k\}_{k \geq 1}$ is a sequence of finite dimensional subspaces of χ^{**} such that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$ and the closure of $\cup_{k=1}^{\infty} \mathcal{F}_k$ is equal to χ^{**}. This sequence exists because of the separability of χ^{**}. Also, for any $x \in \chi^{**}$ and any $k \geq 1$, we define $d(x, \mathcal{F}_k) = \inf\{\|x - y\|_{\chi^{**}} : y \in \mathcal{F}_k\}$. It is easy to prove that for all $k \geq 1$, the map $x \longmapsto d(x, \mathcal{F}_k)$ is continuous and bounded on any closed ball in χ^{**}.

We have $\mathbb{E}\left[\tilde{\psi}_m(X_j)\right] = 0$, for all $j = 1, \ldots, m$. Hence,

$$\begin{aligned}
\sum_{j=1}^{m} \mathbb{E}\left[d^2\left(\tilde{\psi}_m(X_j), \mathcal{F}_k\right)\right] &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left[d^2\left(\mathbf{SGN}^*_{\{F_Y(X_j)\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\{F_Y(X_j)\}}\right], \mathcal{F}_k\right)\right] \\
&= \mathbb{E}\left[d^2\left(\mathbf{SGN}^*_{\{F_Y(X_1)\}} - \mathbb{E}\left[\mathbf{SGN}^*_{\{F_Y(X_1)\}}\right], \mathcal{F}_k\right)\right]
\end{aligned}$$

So, we have

$$\lim_{k \to +\infty} \overline{\lim_{m \to +\infty}} \sum_{j=1}^{m} \mathbb{E}\left[d^2\left(\tilde{\psi}_m(X_j), \mathcal{F}_k\right)\right] = 0.$$

The last equality is derived from the choice of the $\mathcal{F}_k$'s which implies that $d(x, \mathcal{F}_k) \to 0$ as $k \to \infty$ for all $x \in \chi^{**}$. Thus, the third condition of the Corollary 7.8 in Araujo and Giné (1980) holds.

Consequently, $\sum_{j=1}^{m} \tilde{\psi}_m(X_j)$ converges weakly to a centered Gaussian element in χ^{**} as $m, n \to \infty$. Moreover, its covariance asymptotic covariance is Γ_2 which was obtained while checking the second condition as above. Finally,

$$\sqrt{m}L_m^{''} = m^{-1/2} \sum_{j=1}^{m} \left[\phi(F_Y(X_j)) - \mathbb{E}\left[\phi(F_Y(X_j))\right]\right] \xrightarrow{\mathcal{L}} \mathbf{G}(0, \Gamma_2), \tag{2.16}$$

weakly as $m, n \to \infty$.

2.5.3 Step 3 : Asymptotic behavior of $R_{m,n}^{'}$

Since

$$R'_{m,n} = K_{m,n}^{'} + K_{m,n}^{''},$$

the convergence of $K_{m,n}^{'}$ and $K_{m,n}^{''}$ to $\mathbf{0}$ ensures that $R_{m,n}^{'}$ converges to $\mathbf{0}$. We have

$$\begin{aligned}
K_{m,n}^{'} &= \text{MED} - \mathbb{E}[\text{MED}|X_j; j = 1, \ldots, m] - L_n^{'} - L_m^{''} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\phi(F_m(Y_i)) - \mathbb{E}[\phi(F_m(Y_i))|X_1, \ldots, X_m] - \phi(F_X(Y_i)) + \mathbb{E}(\phi(F_X(Y_i)))\right] \\
&- \frac{1}{m} \sum_{j=1}^{m} \left[\phi(F_Y(X_j)) - \mathbb{E}(\phi(F_Y(X_j)))\right] \\
&= \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} h(X_j, Y_i),
\end{aligned}$$

where $h(X_j, Y_i) = \phi(F_m(Y_i)) - \mathbb{E}[\phi(F_m(Y_i))|X_1, \ldots, X_m] - \phi(F_X(Y_i)) + \mathbb{E}(\phi(F_X(Y_i))) - \phi(F_Y(X_j)) + \mathbb{E}(\phi(F_Y(X_j)))$. Hence,

$$K_{m,n}^{'} = \frac{1}{n} \sum_{i=1}^{n} U(Y_i),$$

where $U(Y_i) = \frac{1}{m} \sum_{j=1}^{m} h(X_j, Y_i)$ for all $i = 1, \ldots, n$. Since the X_j's and the Y_i's are independent, conditionally to $X_j, j = 1, \ldots, m$, the $U(Y_i)$'s for $i = 1, \ldots, n$ are independent and zero mean random elements. Using the fact that χ^{**} is a Banach space of type 2 (see, Definition 2.1), there is $b > 0$ such that

$$\mathbb{E}\Big[\Big\|K_{m,n}^{'}\Big\|_{\chi^{**}}^2 \Big| X_j, j = 1, \ldots, m\Big] \leq \frac{b}{n^2 m^2} \sum_{i=1}^{n} \mathbb{E}\Big[\|U(Y_i)\|_{\chi^{**}}^2 \Big| X_j, j = 1, \ldots, m\Big] \tag{2.17}$$

Computing the expectations of both sides of (2.17) leads to

$$\mathbb{E}\left[\left\|K_{m,n}^{'}\right\|_{\chi^{**}}^{2}\right] \leq \frac{b}{nm^{2}}\mathbb{E}\left[\left\|\sum_{j=1}^{m} h(X_j, Y_1)\right\|_{\chi^{**}}^{2}\right].$$

Now, we are willing to find an upper bound for $\mathbb{E}\left[\left\|\sum_{j=1}^{m} h(X_j, Y_1)\right\|_{\chi^{**}}^{2}\right]$. Since the X_j's and the Y_i's are independent, conditionally to Y_1, the $h(X_j, Y_1)$'s for $j = 1, \ldots, m$, are independent and zero mean random elements. Consequently, using the definition of Banach space of type 2 (see, Definition 2.1), there is $b > 0$ such that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^{m} h(X_j, Y_1)\right\|_{\chi^{**}}^{2}\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{j=1}^{m} h(X_j, Y_1)\right\|_{\chi^{**}}^{2}\middle| Y_1\right]\right] \\
&\leq b\sum_{j=1}^{m}\mathbb{E}\left[\mathbb{E}\left[\|h(X_j, Y_1)\|_{\chi^{**}}^{2}\middle| Y_1\right]\right] \\
&= bm\mathbb{E}\left[\|h(X_1, Y_1)\|_{\chi^{**}}^{2}\right].
\end{aligned}$$

Consequently, using the the fact that $\left\|\mathbf{SGN}_{\{x\}}^{*}\right\|_{\chi^{**}} \leq 1$ for all $x \in \chi^{*}$, we have

$$\mathbb{E}\left[\left\|K_{m,n}^{'}\right\|_{\chi^{**}}^{2}\right] \leq \frac{36b^{2}}{mn}. \tag{2.18}$$

Now, we want to find an upper bound for $K_{m,n}^{''}$.

Using Assumption 2.1, the map $g : x \mapsto \mathbb{E}\left[\|F_X(Y) + x\|_{\chi^{*}}\middle| X_1, \ldots, X_m\right]$, for all $x \in \chi^{*}$, is twice Gateaux differentiable and since $\mathbf{J}_x$ exists (see Assumption 2.2), then, for all $h \in \chi^{*}$,

$$\mathbb{E}\left[\mathbf{SGN}_{\{F_X(Y)+x+th\}}^{*}\middle| X_1, \ldots, X_m\right] = \mathbb{E}\left[\mathbf{SGN}_{\{F_X(Y)+x\}}^{*}\middle| X_1, \ldots, X_m\right] + t\mathbf{J}_x(h) + \mathbf{R}(t),$$

where $\|\mathbf{R}(t)\|_{\chi^{**}}/t \to 0$ when $t \to 0$.

Consequently, when $x = 0$, $t = \frac{1}{m}$ and $h = \sum_{j=1}^{m}\left(\mathbf{SGN}_{\{Y_i - X_j\}} - \mathbb{E}\left[\mathbf{SGN}_{\{Y-X\}}\middle| Y\right]\right)$, we obtain $th = F_m(Y_i) - F_X(Y)$ and

$$\begin{aligned}
\mathbb{E}\left[\mathbf{SGN}_{\{F_m(Y_i)\}}^{*}\middle| X_1, \ldots, X_m\right] &= \mathbb{E}\left[\mathbf{SGN}_{\{F_X(Y)\}}^{*}\middle| X_1, \ldots, X_m\right] \\
&\quad + \frac{1}{m}\mathbf{J}_0\left(\sum_{j=1}^{m}\left(\mathbf{SGN}_{\{Y_i - X_j\}} - \mathbb{E}\left[\mathbf{SGN}_{\{Y-X\}}\middle| Y\right]\right)\right) + \mathbf{R}\left(\frac{1}{m}\right).
\end{aligned}$$

Using the linearity of J_0, for all $i = 1, \ldots, n$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{SGN}_{\{F_m(Y_i)\}}^{*}\middle| X_1, \ldots, X_m\right] &= \mathbb{E}\left[\mathbf{SGN}_{\{F_X(Y)\}}^{*}\middle| X_1, \ldots, X_m\right] + \mathbf{J}_0\left(F_m(Y_i) - F_X(Y)\right) \\
&\quad + \mathbf{R}\left(\frac{1}{m}\right).
\end{aligned}$$

Hence,

$$\begin{aligned} K''_{m,n} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\phi(F_m(Y_i))|X_1,\ldots,X_m] - \mathbb{E}\left[\phi(F_X(Y))\right] \\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\mathbf{J}_0\left(\mathbf{SGN}_{\{Y_i-X_j\}}\right) - \mathbf{J}_0\left(\mathbb{E}\left[\mathbf{SGN}_{\{Y-X\}}\middle|Y\right]\right)\right] + \mathbf{R}\left(\frac{1}{m}\right) \\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{h}(Y_i,X_j) + \mathbf{R}\left(\frac{1}{m}\right), \end{aligned}$$

where $\tilde{h}(Y_i,X_j) = \mathbf{J}_0\left(\mathbf{SGN}_{\{Y_i-X_j\}}\right) - \mathbf{J}_0\left(\mathbb{E}\left[\mathbf{SGN}_{\{Y-X\}}\middle|Y\right]\right)$. Thus, we consider

$$K''_{m,n} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(Y_i) + \mathbf{R}\left(\frac{1}{m}\right),$$

where $\tilde{\phi}(Y_i) = \frac{1}{m}\sum_{j=1}^{m}\tilde{h}(Y_i,X_j)$. From the definition of the operator J_0 and once again from the definition of type 2 Banach spaces, the independence between the samples of X and Y and the Y_i's being identically distributed, we get

$$\begin{aligned} \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(Y_i)\right\|^2_{\chi^{**}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(Y_i)\right\|^2_{\chi^{**}}\middle| X_j, j=1,\ldots,m\right]\right] \\ &\leq \frac{b}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\tilde{\phi}(Y_i)\right\|^2_{\chi^{**}}\middle| X_j, j=1,\ldots,m\right]\right] \\ &= \frac{b}{n}\mathbb{E}\left[\left\|\tilde{\phi}(Y_1)\right\|^2_{\chi^{**}}\right]. \end{aligned}$$

Conditionally to Y_1, using the definition of type 2 Banach spaces and the X_j's being identically distributed, we have

$$\begin{aligned} \mathbb{E}\left[\left\|\tilde{\phi}(Y_1)\right\|^2_{\chi^{**}}\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\tilde{h}(Y_1,X_j)\right\|^2_{\chi^{**}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\tilde{h}(Y_1,X_j)\right\|^2_{\chi^{**}}\middle| Y_1\right]\right] \\ &\leq \frac{b}{m}\mathbb{E}\left[\left\|\tilde{h}(Y_1,X_1)\right\|^2_{\chi^{**}}\right] \\ &= \frac{b}{m}\mathbb{E}\left[\left\|\mathbf{J}_0\left(\mathbf{SGN}_{\{Y_1-X_1\}}\right) - \mathbf{J}_0\left(\mathbb{E}\left[\mathbf{SGN}_{\{Y-X\}}\middle|Y\right]\right)\right\|^2_{\chi^{**}}\right] \\ &\leq \frac{4bc^2}{m}. \end{aligned} \tag{2.19}$$

The inequality (2.19) comes from $\mathbf{J}_0$ being a linear continuous map, the Assumption 2.2 and the fact that $\left\|\mathbf{SGN}_{\{x\}}\right\|_{\chi^*} \leq 1$, for all $x \in \chi$. Thus, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|K_{m,n}^{''}\right\|_{\chi^{**}}^2\right] &\leq 2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(Y_i)\right\|_{\chi^{**}}^2\right] + 2\mathbb{E}\left[\left\|\mathbf{R}\left(\frac{1}{m}\right)\right\|_{\chi^{**}}^2\right] \\ &\leq \frac{8b^2c^2}{mn} + 2\left\|\mathbf{R}\left(\frac{1}{m}\right)\right\|_{\chi^{**}}^2. \end{aligned} \tag{2.20}$$

Therefore, combining (2.18) and (2.20), we get

$$\begin{aligned}\mathbb{E}\left[\|R_{m,n}^{'}\|_{\chi^{**}}^2\right] &\leq 2\mathbb{E}\left[\|K_{m,n}^{'}\|_{\chi^{**}}^2\right] + 2\mathbb{E}\left[\|K_{m,n}^{''}\|_{\chi^{**}}^2\right] \\ &\leq \frac{72b^2}{mn} + \frac{16b^2c^2}{mn} + 4\left\|\mathbf{R}\left(\frac{1}{m}\right)\right\|_{\chi^{**}}^2\end{aligned}$$

Finally,

$$\mathbb{E}\left[\left\|\left(\frac{mn}{N}\right)^{1/2}R_{m,n}^{'}\right\|_{\chi^{**}}^2\right] \leq \frac{72b^2}{N} + \frac{16b^2c^2}{N} + \frac{4n}{mN}\left(\frac{\left\|\mathbf{R}\left(\frac{1}{m}\right)\right\|_{\chi^{**}}}{\frac{1}{m}}\right)^2.$$

Since $\frac{m}{N} \to \lambda \in (0,1)$ as $m, n \to \infty$, we obtain

$$\mathbb{E}\left[\left\|\left(\frac{mn}{N}\right)^{1/2}R_{m,n}^{'}\right\|_{\chi^{**}}^2\right] \underset{m,n\to\infty}{\longrightarrow} 0. \tag{2.21}$$

2.5.4 Step 4 : Asymptotic behavior of MED

Let's take the equation (2.13) again :

$$\text{MED} - \mathbb{E}\left[\phi(F_X(Y))\right] = L_n^{'} + L_m^{''} + R_{m,n}^{'}$$

$\Leftrightarrow$

$$\sqrt{\frac{mn}{N}}\left[\text{MED} - \mathbb{E}\left[\phi(F_X(Y))\right]\right] = \sqrt{\frac{mn}{N}}L_n^{'} + \sqrt{\frac{mn}{N}}L_m^{''} + \sqrt{\frac{mn}{N}}R_{m,n}^{'}.$$

From the convergence results (2.21), (2.14) and (2.16) achieved in steps 1, 2 and 3, we get

$$\sqrt{\frac{mn}{N}}L_n^{'} \textit{ converges weakly to } \mathbf{G}(0, \lambda\Gamma_1), \tag{2.22}$$

$$\sqrt{\frac{mn}{N}}L_m^{''} \textit{ converges weakly to } \mathbf{G}\left(0, (1-\lambda)\Gamma_2\right) \tag{2.23}$$

and

$$\sqrt{\frac{mn}{N}}R_{m,n}^{'} \textit{ converges in probability to } \mathbf{0} \tag{2.24}$$

as $m, n \to \infty$. Hence, using Slutsky lemma and the independence of $L_n^{'}$ and $L_m^{''}$, we obtain

$$\sqrt{\frac{mn}{N}}\,[\mathrm{MED} - \mu] \xrightarrow{\mathcal{L}} \mathbf{G}(0, \lambda\Gamma_1 + (1-\lambda)\Gamma_2), \tag{2.25}$$

weakly as $m, n \to \infty$. This completes the proof of the Theorem 2.1. □

Chapter 3

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

"Happiness can be found even in the darkest of times if one only remembers to turn on the light."

— J.K. Rowling

Chapter contents

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.

3.1 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, 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 3.2, we build a nonparametric 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 3.3, 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.

3.2 A nonparametric spatial scan statistic for functional data

3.2.1 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 (Chen and Glaz , 2009; Loader , 1991), circular (Kulldorff , 1997; Kulldorff and Nagarwalla , 1995), 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 (Assunção et al. , 2006; Demattei et al. , 2007; 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:

$$\mathcal{S} = \{D_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n\}, \tag{3.1}$$

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 3.1. *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 $n_Z, n_{Z^c} \to \infty$,

$$(n_Z n_{Z^c}/n)^{1/2} T_{\mathrm{WMW}}(Z) \text{ converges weakly to } G(0,\Gamma), \tag{3.2}$$

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

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

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

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

This latter is called the most likely cluster.

3.2.2 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

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

  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: WMW (WMW test statistic value)
    Initialization: WMW $= 0$
    for $i \in A$ **do**
        for $j \in B$ **do**
            WMW $=$ WMW $+ R_{i,j}$
    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**
        $O =$ 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**
                $\tilde{c} = c$, $\tilde{i} = i$, $\tilde{j} = j$
    $\tilde{O} =$ ORDER$(\{d_{\tilde{i},j}\}_{j\in\{1,\ldots,n\}})$
    MLC $= \{\tilde{O}_k\}_{k\in\{1,\ldots,\tilde{j}\}}$

3.2.3 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.

3.3 Applications

3.3.1 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 Figure 3.1.

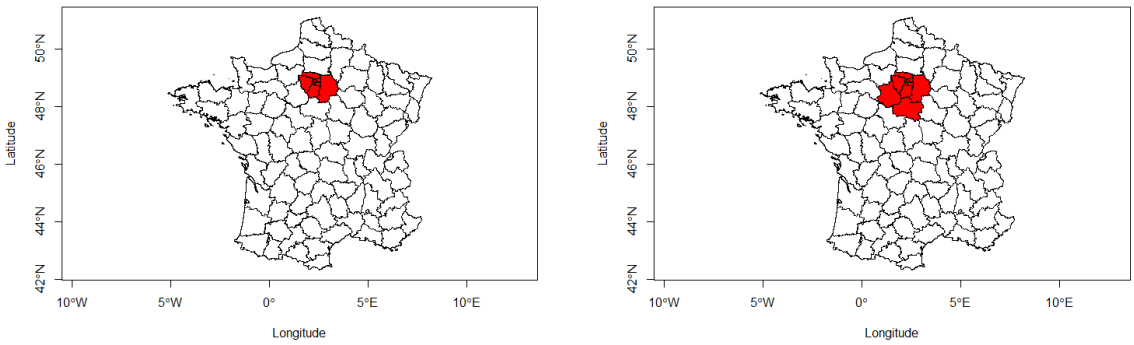

Figure 3.1: The 94 French *départements*. In red: simulated clusters (8 and 10 *départements*).

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 $\hat{\text{C}}$ 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 3.5.2 but they are summarized in Table 3.1 and Table 3.2 below.

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 3.1: 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	**0.060**	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

Table 3.2: 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.

From Table 3.1 and Table 3.2, 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 3.7 and Table 3.10 in Appendix 3.5.2): 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 Λ_{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.

3.3.2 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 Figure 3.2) 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 Figure 3.3). 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.

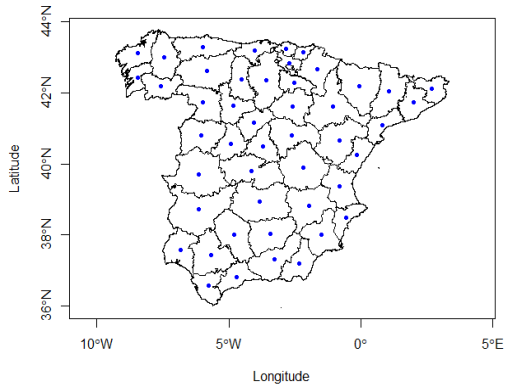

Figure 3.2: The 47 Spanish provinces and their geometrical centres.

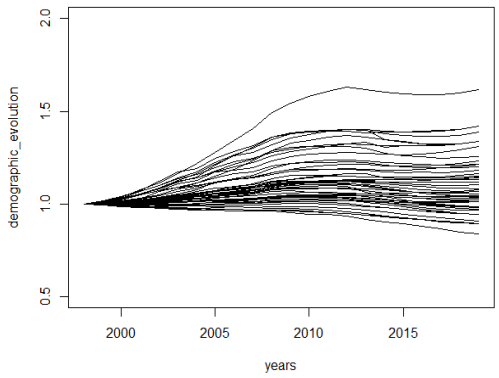

Figure 3.3: Demographic evolution in the 47 provinces from 1998 to 2019.

3.3.2.1 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: $\Lambda_{\text{WMWFSS}} = 2.72025$. 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 Figure 3.4.

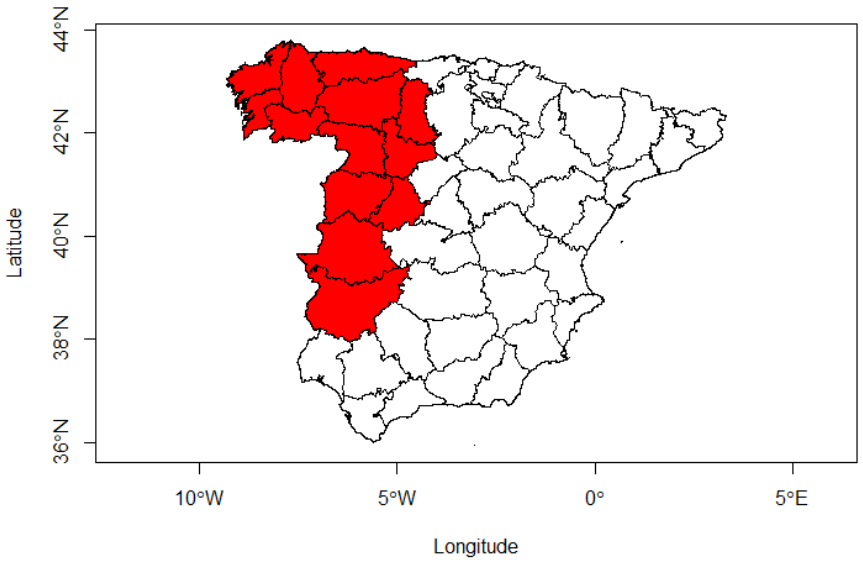

Figure 3.4: The most likely cluster detected by the functional scan statistic Λ_{WMWFSS}.

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 Figure 3.5.

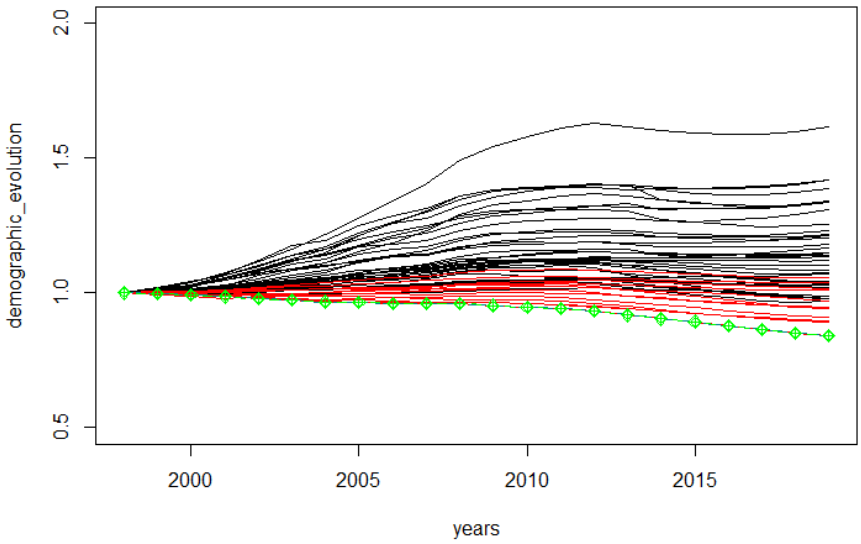

Figure 3.5: The demographic evolution curves (from 1998 to 2019) in each province are presented. Curves in red correspond to provinces inside the cluster, curves in black correspond to provinces outside the cluster and the curve in green corresponds to *Zamora* which is inside the cluster too.

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 Figure 3.5) 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 3.3.

	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

Table 3.3: The p-values and computation time (in seconds) of the different scan methods using different number of permutations.

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 Figure 3.6). 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.

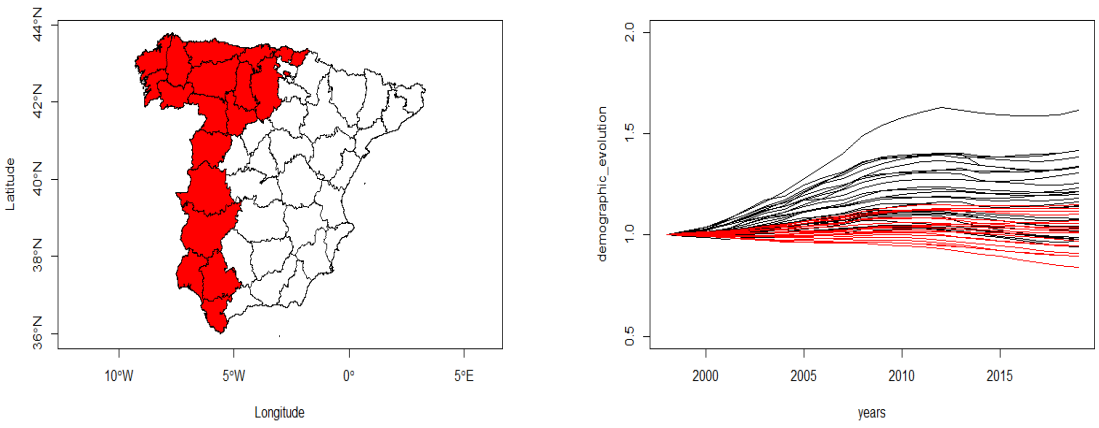

Figure 3.6: The most likely cluster detected by Λ_{MBUSS} and demographic evolution curves associated.

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. (2021c).
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.

3.3.2.2 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

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

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 subsection 3.3.1, Table 3.4 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 Figure 3.4).
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.

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

Table 3.4: 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.

3.4 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 et al. , 2010; Zhang and Chen , 2007) as long as its asymptotic distribution is known. In a preprint, Frévent et al. (2021a) 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 Λ_{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. , 2021c).
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).

3.5 Appendix

3.5.1 Examples of the generated data in subsection 3.3.1

The following Figure 3.7, Figure 3.8 and Figure 3.9 show examples of simulated data using sBm process with different types of shifts.

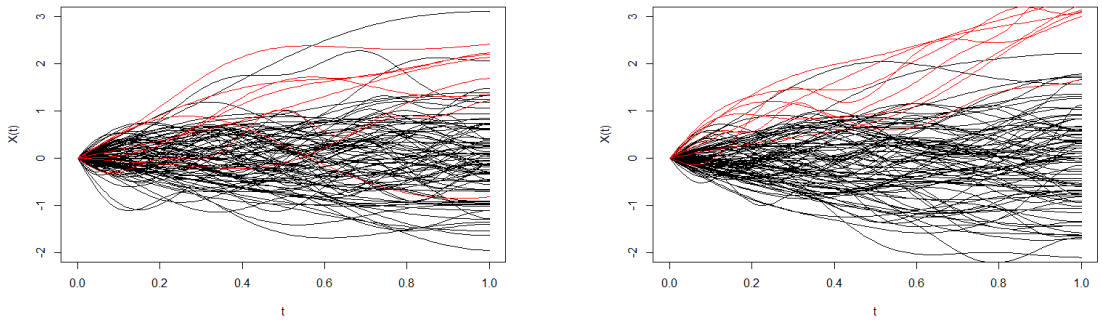

Figure 3.7: An example of the simulated data for the sBm process with $\Delta_1(t) = t$ (left panel) and $\Delta_1(t) = 3t$ (right panel). Curves in red correspond to the observations in the cluster.

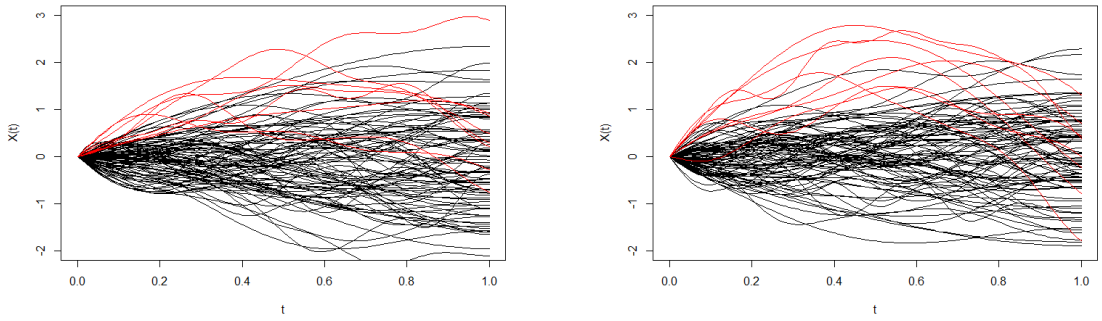

Figure 3.8: An example of the simulated data for the sBm process with $\Delta_2(t) = 4t(1-t)$ (left panel) and $\Delta_2(t) = 7t(1-t)$ (right panel). Curves in red correspond to the observations in the cluster.

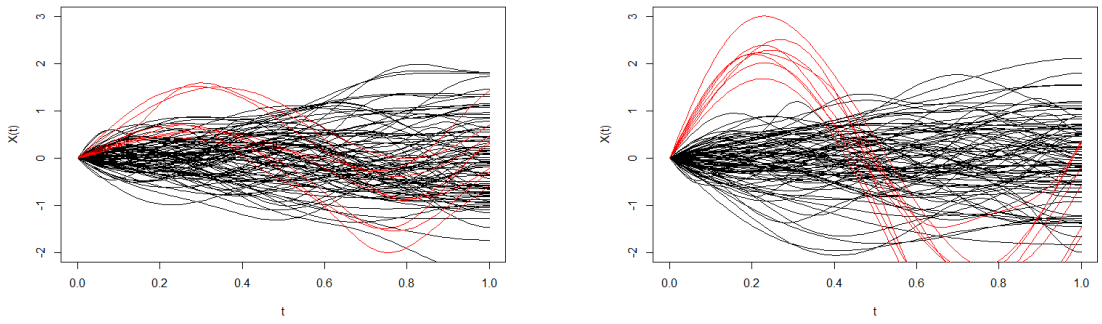

Figure 3.9: An example of the simulated data for the sBm process with $\Delta_3(t) = \sin(2\pi t)$ (left panel) and $\Delta_3(t) = 2.5\sin(2\pi t)$ (right panel). Curves in red correspond to the observations in the cluster.

3.5.2 Results of the simulation study in subsection 3.3.1

- When the true cluster is a set of 8 *départements*:
  The following Table 3.5, Table 3.6 and Table 3.7 give the results obtained in this simulation study. 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
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 3.5: 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}
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 3.6: 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}
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

Table 3.7: 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*.

- When the true cluster is a set of 10 *départements*:
  The following Table 3.8, Table 3.9 and Table 3.10 give the results obtained in this simulation study. 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
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 3.8: 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}
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 3.9: 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}
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

Table 3.10: 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*.

Chapter 4

The R Package HDSpatialScan for Multivariate and Functional Spatial Scan Statistics

"Sometimes the best results come when you are thrown in the deep end."

— Natalie Cook

Chapter contents

Abstract

This paper introduces the R package **HDSpatialScan**. This package allows users to apply easily spatial scan statistics on real-valued multivariate data or both univariate and multivariate functional data. It also permits to plot the detected clusters and to summarize them. In this article the methods are presented and the use of the package is illustrated through examples on environmental data provided in the package.

4.1 Introduction

Spatial cluster detection methods are useful tools for objective detecting aggregation of events indexed in space and determining the latter's statistical significance. Examples of applications of these methods are numerous: in the field of epidemiology, these methods allow epidemiologists to detect spatial clusters of disease cases and to formulate etiological hypotheses; in the environmental sciences, researchers can be led to search for particularly polluted geographical areas, either by one pollutant in particular or by several pollutants simultaneously. In astronomy, researchers may want to identify star clusters from telescope image data.

Several cluster detection methods have been proposed in the literature. In particular, spatial scan statistics (originally proposed by Kulldorff and Nagarwalla (1995) and Kulldorff (1997) for Bernoulli and Poisson models) are powerful methods for detecting statistically significant spatial clusters, which can be defined by an aggregation of sites presenting an abnormal concentration (mean, proportion ...) of an observed variable, with a variable scanning window and in the absence of pre-selection bias. Following on from Kulldorff's initial work, several researchers have adapted spatial scan statistics to other spatial data distributions: exponential (Huang et al. , 2007), ordinal (Jung et al. , 2007), normal (Kulldorff et al. , 2009), Weibull (Bhatt and Tiwari , 2014)... Others use nonparametric approaches such as Jung and Cho (2015) and Cucala (2016) who respectively extend the Wilcoxon-Mann-Whitney test for spatial scan statistics and for temporal or spatial scan statistics. Note that in the case of spatial data the two approaches are equivalent by generalizing the method of Jung and Cho (2015) to detect either high or low clusters.

When multiple variables are observed simultaneously at each spatial location, researchers may be interested in detecting spatial clusters with anomalous values of all measured variables. In this context, Kulldorff et al. (2007) proposed a multivariate spatial scan statistic using a combination of independent univariate scan statistics. However it fails to take into account the correlations between the variables. A first spatial scan statistic for multivariate data taking into account the correlations was proposed by Cucala et al. (2017). Their method is based on a multivariate normal probability model and a likelihood ratio. Then Cucala et al. (2019) proposed a nonparametric spatial scan statistic for multivariate data based on a multivariate Wilcoxon-Mann-Whitney test.

Technological developments in measurement tools and data storage capacity have yielded to the increasing use of sensors, cell phones and more generally connected devices that collect data continuously or almost continuously over time. This has led to the introduction of new analysis methods for functional data (Ramsay and Silverman , 2005), as well as the adaptation of classical statistical methods such as principal component analysis (Berrendero et al. , 2011; Boente and Fraiman , 2000) or regression (Chiou and Müller , 2007; Cuevas et al. , 2002; Ferraty and Vieu , 2002).

In the field of spatial scan statistics, Frévent et al. (2021a) and Smida et al. (2021) proposed new methods for univariate processes. However sometimes such as in environmental surveillance, numerous variables are simultaneously measured which makes a multivariate functional approach necessary to detect environmental black-spots. Although Smida et al. (2021) only studied their approach in the univariate functional framework, they suggest that it could also be adapted for multivariate processes. Frévent et al. (2021b) studied this adaptation and also developed new efficient methods for multivariate functional spatial scan statistics.

In R several packages provide spatial scan statistics implementations. The best known is certainly the **rsatscan** package (Kleinman , 2015) which provides functions to interface R and the SaTScan software (Kulldorff , 2021), allowing the latter to be launched from R. It implements lots of univariate methods (ordinal, Bernoulli, Poisson, ...) but also the space-time spatial scan statistic (Kulldorff et al. , 1998) and the multivariate extensions proposed by Kulldorff et al. (2007). The function `kulldorff()` implemented in the R package **SpatialEpi** (Chen et al. , 2018) also performs the spatial scan statistics based on the Poisson and the Bernoulli models. Other softwares were created to detect clusters such as ClusterSeer (Durbeck et al. , 2012; Greiling et al. , 2012) which performs spatial, temporal and space-time clustering, and TreeScan (Kulldorff , 2018) which implements the tree-based scan statistic (Kulldorff et al. , 2003). The Shiny application SpatialEpiApp (Moraga , 2017) and the R package **SpatialEpiApp** allow the detection and visualization of clusters by using the scan statistics implemented in SaTScan. Finally the software FlexScan (Takahashi et al. , 2010) and the R package **rflexscan** (Otani and Takahashi , 2021) implement the spatial scan statistic using a scan window with a non pre-defined shape, defined by Takahashi and Tango (2005). Other R packages also allow clusters detection such as **graphscan** (Loche et al. , 2016) (the `cluster()` function), **SPATCLUS** (Dematteï et al. , 2006) or **scanstatistics** (Allévius , 2018) for spatial or space-time data. It should be noted that these last two packages are no longer available on the CRAN (The Comprehensive R Archive Network) repository. Although existing packages implement a large number of statistical spatial scan models, none of them propose multivariate scan models taking into account correlation between variables or scan models for functional data. Thus, we have developed the R package **HDSpatialScan** for

high-dimensional spatial scan statistics. The latter allows on the one hand the detection of spatial clusters in multivariate or functional data, and on the other hand, their display on a map and the description of their characteristics.
This paper is organized as follows: Section 4.2 presents the different models implemented in the R package **HDSpatialScan**. Section 4.3 describes the implementation of the methods, and, in Section 4.4, examples of use of the package are given. Finally the paper is concluded in Section 4.5.

4.2 Models

Let $s_1, \ldots, s_n$ be n non-overlapping locations of an observation domain $S \subset \mathbb{R}^2$ and $X_1, \ldots, X_n$ be the observations of a variable X in $s_1, \ldots, s_n$. Hereafter all observations are considered to be independent, which is a classical assumption in scan statistics. Two types of spatial data can be considered: either lattice data (the data are aggregated at the spatial level, e.g.: county) or geostatistical data (each individual measure corresponds to a unique spatial location, e.g.: pollutant concentration measured by sensors over a region). Spatial scan statistics aim at detecting spatial clusters and testing their significance. Hence, one tests a null hypothesis $\mathcal{H}_0$ (the absence of a cluster) against a composite alternative hypothesis $\mathcal{H}_1$ (the presence of at least one cluster $w \subset S$ presenting abnormal values of X). For this purpose we use a set of potential clusters $\mathcal{W}$. The approach most often used and introduced by Kulldorff and Nagarwalla (1995) is to use circular clusters of variable size. An approach often advised is to limit the maximum size to half of the studied region since otherwise it would be like detecting a "negative cluster" in the areas outside the clusters covering almost all the studied region (Kulldorff and Nagarwalla , 1995). Following the latter's the set of potential clusters is defined by

$$\mathcal{W} = \{w_{i,j} \ / \ 1 \leq |w_{i,j}| \leq \frac{n}{2}, \ 1 \leq i, j \leq n\}, \tag{4.1}$$

where $w_{i,j}$ is the disc centered on s_i that passes through s_j and $|w_{i,j}|$ corresponds to the number of sites in $w_{i,j}$.

4.2.1 Multivariate spatial scan statistics

In a few words. In this subsection we consider the case where several continuous variables are simultaneously observed in each spatial location: $X = (X^{(1)}, \ldots, X^{(p)})^\top$ is a p-dimensional variable ($p \geq 2$). In this context the objective is to identify multivariate spatial clusters that is aggregations of sites in which X takes higher or lower values than elsewhere. For example one could observe the average concentrations of several pollutants over a day: a vector can be associated with each site, each element of which corresponds to the average concentration of one pollutant. In this context a spatial cluster corresponds to a set of sites under or overexposed to multiple pollutants. Different approaches will be presented: a parametric method based on a Gaussian model and a nonparametric one. On

Gaussian data both approaches offer high power and high true positive rates. On the other hand, in the case of non-normal data, the non-parametric method has to be preferred.

Cucala et al. (2017) proposed a parametric spatial scan statistic for multivariate data based on a multivariate normal model taking into account the correlations between the variables.

The null hypothesis $\mathcal{H}_0$, corresponding to the absence of any cluster in the data, is the following: $\forall i \in [\![1;n]\!],\ X_i \sim \mathcal{N}_p(\mu, \Sigma)$ and the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as: $\forall i \in [\![1;n]\!],\ X_i \sim \begin{cases} \mathcal{N}_p(\mu_w, \Sigma_{w,w^c}) & \text{if } s_i \in w \\ \mathcal{N}_p(\mu_{w^c}, \Sigma_{w,w^c}) & \text{otherwise} \end{cases}$.

Then we can compute the MLE estimates of $\mu, \mu_w, \mu_{w^c}, \Sigma$ and Σ_{w,w^c}: $\hat{\mu}, \widehat{\mu}_w, \widehat{\mu}_{w^c}, \widehat{\Sigma}$ and $\widehat{\Sigma}_{w,w^c}$, and we can show that the log-likelihood ratio between these two hypotheses is

$$\begin{aligned}\widehat{LLR^w} = &-\frac{n}{2}\ln\left[\det\Big(\sum_{\substack{i\\ s_i\in w}}(X_i-\widehat{\mu}_w)(X_i-\widehat{\mu}_w)^\top + \sum_{\substack{i\\ s_i\in w^c}}(X_i-\widehat{\mu}_{w^c})(X_i-\widehat{\mu}_{w^c})^\top\Big)\right] \\ &+\frac{n}{2}\ln\left[\det\Big(\sum_{i=1}^{n}(X_i-\hat{\mu})(X_i-\hat{\mu})^\top\Big)\right],\end{aligned}$$

where $\hat{\mu}_g = \dfrac{1}{|g|}\displaystyle\sum_{i,s_i\in g} X_i$ for $g \in \{w, w^c\}$ and $\hat{\mu} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.
Finally the log-likelihood ratio is used as a concentration index and maximised over the set of potential clusters $\mathcal{W}$.
Thus we can show that the multivariate Gaussian (MG) scan statistic is

$$\lambda_{\text{MG}} = \min_{w\in\mathcal{W}} \det\Big(\sum_{\substack{i\\ s_i\in w}}(X_i-\widehat{\mu}_w)(X_i-\widehat{\mu}_w)^\top + \sum_{\substack{i\\ s_i\in w^c}}(X_i-\widehat{\mu}_{w^c})(X_i-\widehat{\mu}_{w^c})^\top\Big).$$

This test performs very well against Gaussian alternatives but faces problems when the data is not normal, which is often the case when dealing with environmental data exhibiting extreme values. For that reason Cucala et al. (2019) developed a nonparametric spatial scan statistic for multivariate data based on a multivariate extension of the Wilcoxon-Mann-Whitney test for multivariate data (Oja and Randles , 2004).
In this context the null hypothesis $\mathcal{H}_0$ can be rewritten as $\mathcal{H}_0 : X_1, \ldots, X_n$ are identically distributed, whatever the associated location.
Let

$$\begin{array}{rccl} \text{sgn} & : & \mathbb{R}^p & \to & \mathbb{R}^p \\ & & x & \mapsto & \begin{cases} ||x||_2^{-1}x & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases} \end{array},$$

then the multivariate ranks R_i are defined by $R_i = \frac{1}{n}\sum_{j=1}^{n} \text{sgn}(A_X(X_i - X_j))$ where the matrix A_X makes the ranks such that $\frac{p}{n}\sum_{i=1}^{n} R_i R_i^\top = \frac{1}{n}\sum_{i=1}^{n} R_i^\top R_i I_p$. Note that this matrix can be easily computed using an iterative procedure. Then the multivariate extension of the Wilcoxon-Mann-Whitney statistic proposed by Oja and Randles (2004) is

$$U^2(w) = \frac{p}{c_X^2}[\ |w|\ ||\bar{R}_w||_2^2 + |w^{\mathsf{c}}|\ ||\bar{R}_{w^{\mathsf{c}}}||_2^2\ ],\ \text{where } c_X^2 = \frac{1}{n}\sum_{i=1}^{n} R_i^\top R_i.$$

Cucala et al. (2019) used $U^2(w)$ as a concentration index to build the spatial scan statistic: the multivariate nonparametric (MNP) scan statistic is

$$\lambda_{\text{MNP}} = \max_{w \in \mathcal{W}} U^2(w).$$

It should be noted that in the case $p = 1$, these statistics are respectively equivalent to the ones introduced by Kulldorff et al. (2009) (which is equivalent to the scan statistic developed by Cucala (2014)), and Jung and Cho (2015).

4.2.2 Spatial scan statistics for univariate functional data

In a few words. This subsection considers the case where a continuous variable is observed in each spatial location over time: $\{X(t),\ t \in \mathcal{T}\}$ is a real-valued stochastic process where $\mathcal{T}$ is an interval of $\mathbb{R}$. In this context the objective is to identify functional spatial clusters that is aggregations of sites in which the curves are higher or lower than elsewhere. For example, one can observe the concentration of an air pollutant over time in different geographical areas. Then a cluster corresponds to an aggregation of sites in which the concentration of the air pollutant is higher or lower over the time than in the other spatial units. Several methods will be considered: a parametric method based on a functional ANOVA, a nonparametric approach using a Wilcoxon-Mann-Whitney test for high-dimensional data, a distribution-free approach based on a pointwise Student's t-test and finally a pointwise rank-based method. On Gaussian data, for non localized clusters in time all approaches show high power and high true positive rates. However the performances of the ANOVA-based method strongly decrease on non-normal data. For localized clusters in time (that are aggregations of sites that take higher or lower values for X only in a small interval of time) the pointwise approaches should be favored.

4.2.2.1 The parametric spatial scan statistic for univariate functional data

Frévent et al. (2021a) supposed that the process X takes values in a semi-metric space, in particular in $\mathcal{L}^2(\mathcal{T}, \mathbb{R})$ and proposed a parametric spatial scan statistic for

functional data, based on a functional ANOVA. Here the null hypothesis $\mathcal{H}_0$ can be rewritten: $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^c} = \mu_S$, and the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as follows: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$, where μ_w, μ_{w^c} and μ_S stand for the mean functions in w, outside w and over S, respectively. Cuevas et al. (2004) and Górecki and Smaga (2015) proposed the following ANOVA test statistic:

$$F_n^{(w)} = \frac{|w|\ ||\bar{X}_w - \bar{X}||_2^2 + |w^c|\ ||\bar{X}_{w^c} - \bar{X}||_2^2}{\frac{1}{n-2}\left[\sum_{j,s_j \in w} ||X_j - \bar{X}_w||_2^2 + \sum_{j,s_j \in w^c} ||X_j - \bar{X}_{w^c}||_2^2\right]},$$

where $\bar{X}_g(t) = \frac{1}{|g|} \sum_{i,s_i \in g} X_i(t)$ are empirical estimators of μ_g $(g \in \{w, w^c\})$,
$\bar{X}(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t)$ is the empirical estimator of μ_S and $||x||_2^2 = \int_{\mathcal{T}} x^2(t)\ \mathrm{d}t$.
Thus, Frévent et al. (2021a) proposed to use $F_n^{(w)}$ as a concentration index and the proposed parametric functional spatial scan statistic (PFSS) is

$$\Lambda_{\text{PFSS}} = \max_{w \in \mathcal{W}} F_n^{(w)}.$$

This method gives high powers and F-measures on normal data but as in the multivariate framework the parametric method faces problems when the data is not normal. Smida et al. (2021) proposed a nonparametric spatial scan statistic for functional data based on a functional Wilcoxon-Mann-Whitney test (Chakraborty and Chaudhuri , 2015).

4.2.2.2 A nonparametric spatial scan statistic for functional data

Here X is a process of a smooth Banach space χ, with a Gateaux differentiable norm $||.||_\chi$. Let denote P_w and P_{w^c} the probability measures of X in w and in w^c respectively, then $\mathcal{H}_0$ corresponds to: $\mathcal{H}_0 : \forall w \in \mathcal{W},\ P_w = P_{w^c}$ and the alternative hypothesis associated with a potential cluster w can be rewritten as $\mathcal{H}_1^{(w)} : P_w(X) = P_{w^c}(X - \Delta),\ \Delta \in \chi \backslash \{0\}$.
Chakraborty and Chaudhuri (2015) defined the sign function in the functional framework as

$$\forall h \in \chi,\ \text{Sgn}_X(h) = \begin{cases} \lim\limits_{v \to 0^+} \dfrac{||X + vh||_\chi - ||X||_\chi}{v} & \text{if } X \neq 0 \\ 0 & \text{if } X = 0 \end{cases}.$$

Then they proposed the following test statistic:

$$T_{WMW}(w) = \frac{1}{|w||w^c|} \sum_{i,s_i \in w} \sum_{j,s_j \in w^c} \text{Sgn}_{X_j - X_i}.$$

Under $\mathcal{H}_0$, if $\frac{|w|}{n} \to \gamma \in [0;1]$ as $|w|, |w^c| \to \infty$, $\sqrt{\frac{|w||w^c|}{n}} T_{WMW}(w)$ converges weakly to a distribution that does not depend on $|w|$. Thus Smida et al. (2021)

proposed to use

$$U(w) = \left|\left| \sqrt{\frac{|w||w^{\mathsf{c}}|}{n}} T_{WMW}(w) \right|\right|$$

as a concentration index: the nonparametric functional scan statistic (NPFSS) is

$$\Lambda_{\text{NPFSS}} = \max_{w \in \mathcal{W}} U(w).$$

It should be noticed that although Smida et al. (2021) only studied the performances of the NPFSS in the univariate functional framework, their method is also applicable on multivariate functional data as shown by Frévent et al. (2021b).

4.2.2.3 A distribution-free spatial scan statistic for univariate functional data

Frévent et al. (2021a) also proposed to combine the distribution-free spatial scan statistic for univariate data proposed by Cucala (2014) and the max statistic of Lin et al. (2021). They supposed that for each time t, $\mathbb{V}[X_i(t)] = \sigma^2(t)$ for all $i \in [\![1; n]\!]$. Then for each t, the concentration index proposed by Cucala (2014) to test $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w(t) = \mu_{w^{\mathsf{c}}}(t) = \mu_S(t)$ was

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)]}},$$

where

$$\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)] = \widehat{\sigma^2}(t) \left[\frac{1}{|w|} + \frac{1}{|w^{\mathsf{c}}|} \right]$$

and

$$\widehat{\sigma^2}(t) = \frac{1}{n-2} \left[\sum_{i, s_i \in w} (X_i(t) - \bar{X}_w(t))^2 + \sum_{i, s_i \in w^{\mathsf{c}}} (X_i(t) - \bar{X}_{w^{\mathsf{c}}}(t))^2 \right].$$

Then the idea is to globalize the information by maximizing the previous quantity over the time for each potential cluster w, as suggested by Lin et al. (2021):

$$I^{(w)} = \sup_{t \in \mathcal{T}} I^{(w)}(t).$$

For cluster detection, as for the PFSS, the null hypothesis $\mathcal{H}_0$ (the absence of cluster) is defined as follows: $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^{\mathsf{c}}} = \mu_S$. And the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as follows: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^{\mathsf{c}}}$.

Frévent et al. (2021a) considered $I^{(w)}$ as a concentration index and maximized it over the set of potential clusters $\mathcal{W}$ yielding to the following distribution-free functional spatial scan statistic (DFFSS):

$$\Lambda_{\text{DFFSS}} = \max_{w \in \mathcal{W}} I^{(w)}.$$

4.2.2.4 A new rank-based spatial scan statistic for univariate functional data

A pointwise approach based on ranks and the nonparametric scan statistic for univariate data (Jung and Cho , 2015) can be proposed in the univariate functional framework by adapting the approach of Frévent et al. (2021b). For a time t Jung and Cho (2015) proposed to test $\mathcal{H}_0 : \forall w \in \mathcal{W}, F_{w,t} = F_{w^c,t}$ where $F_{w,t}$ and $F_{w^c,t}$ are the cumulative distribution functions of $X(t)$ in w and outside w, by using the Wilcoxon rank-sum test statistic. For a time t and a potential cluster w, the Wilcoxon rank-sum test statistic is $W(t)^{(w)} = \sum_{i,s_i \in w} R_i(t)$ where $R_i(t)$ is the rank of $X_i(t)$ in $\{X_1(t), \ldots, X_n(t)\}$, using the average rank in the case of tied observations.
Then the standardized version of this statistic is

$$T(t)^{(w)} = \frac{W(t)^{(w)} - \mathbb{E}[W(t)^{(w)}]}{\sqrt{\mathbb{V}[W(t)^{(w)}]}}$$

where $\mathbb{E}[W(t)^{(w)}] = \frac{|w|(n+1)}{2}$ and $\mathbb{V}[W(t)^{(w)}] = \frac{|w||w^c|(n+1)}{12}$ are respectively the expected value and the variance of $W(t)^{(w)}$ under $\mathcal{H}_0$.
Jung and Cho (2015) proposed to minimize the p-value associated with $T(t)^{(w)}$ on the set of potential clusters $\mathcal{W}$. We propose to adapt their approach by simply using $|T(t)^{(w)}|$ as a pointwise statistic.

In the context of cluster detection, the null hypothesis is defined as $\mathcal{H}_0$: $\forall w \in \mathcal{W},\ \forall t \in \mathcal{T},\ F_{w,t} = F_{w^c,t}$. The alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w is $\mathcal{H}_1^{(w)}$: $\exists t \in \mathcal{T},\ F_{w,t}(x) = F_{w^c,t}(x - \Delta_t),\ \Delta_t \neq 0$.

As before, we propose to globalize the information over the time with

$$T^{(w)} = \sup_{t \in \mathcal{T}} |T(t)^{(w)}|$$

and to use this quantity as a concentration index, yielding to the following univariate rank-based functional spatial scan statistic (URBFSS):

$$\Lambda_{\text{URBFSS}} = \max_{w \in \mathcal{W}} T^{(w)}.$$

4.2.3 Spatial scan statistics for multivariate functional data

In a few words. This subsection considers the case where several continuous variables are observed simultaneously in each spatial unit over time: $\{X(t),\ t \in \mathcal{T}\}$ is a p-dimensional vector-valued stochastic process ($p \geq 2$) where $\mathcal{T}$ is an interval of $\mathbb{R}$. The objective is to detect multivariate functional spatial clusters that is aggregations of sites in which the curves are higher or lower than elsewhere.

For example we can observe the concentration of several pollutants over time in different locations. Thus at each location we observe several processes (air pollutant concentrations) and these processes can be correlated. In this context a cluster is an aggregation of sites overexposed or underexposed to multiple pollutants over time. Several methods will be presented: a parametric method based on a functional MANOVA, a distribution-free approach based on a pointwise Hotelling T^2-test and finally a pointwise rank-based method. On normal data, all approaches show high power and high true positive rates for non localized clusters in time. However the performances of the methods based on the MANOVA and the Hotelling T^2-test decrease on non-normal data. For localized clusters in time the pointwise approaches should be favored, especially the pointwise rank-based method on non-Gaussian data. By localized clusters in time we mean aggregations of sites that take higher or lower values for X only in a small interval of time.

4.2.3.1 A parametric spatial scan statistic for multivariate functional data

Here, the process X is supposed to take values in a semi-metric space, in particular the Hilbert space $L^2(\mathcal{T}, \mathbb{R}^p)$ of p-dimensional vector-valued square-integrable functions on $\mathcal{T}$, equipped with the inner product $\langle X, Y \rangle = \int_{\mathcal{T}} X(t)^\top Y(t)\ \mathrm{d}t$.
Frévent et al. (2021b) proposed a parametric scan statistic for multivariate functional data based on a functional MANOVA Lawley–Hotelling trace test (Górecki and Smaga , 2017).

In this context, the null hypothesis $\mathcal{H}_0$ is $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^\mathsf{c}} = \mu_S$, where μ_w, μ_{w^c} and μ_S stand for the mean functions in w, outside w and over S, respectively. And the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w is $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^\mathsf{c}}$. Górecki and Smaga (2017) presented the following adaptation of the Lawley-Hotelling trace test statistic:

$$\mathrm{LH}^{(w)} = \mathrm{Trace}(H_w E_w^{-1})$$

where

$$\begin{aligned} H_w &= |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)][\bar{X}_w(t) - \bar{X}(t)]^\top \mathrm{d}t \\ &+ |w^\mathsf{c}| \int_{\mathcal{T}} [\bar{X}_{w^\mathsf{c}}(t) - \bar{X}(t)][\bar{X}_{w^\mathsf{c}}(t) - \bar{X}(t)]^\top \mathrm{d}t \end{aligned}$$

and

$$\begin{aligned} E_w &= \sum_{j, s_j \in w} \int_{\mathcal{T}} [X_j(t) - \bar{X}_w(t)][X_j(t) - \bar{X}_w(t)]^\top \mathrm{d}t \\ &+ \sum_{j, s_j \in w^\mathsf{c}} \int_{\mathcal{T}} [X_j(t) - \bar{X}_{w^\mathsf{c}}(t)][X_j(t) - \bar{X}_{w^\mathsf{c}}(t)]^\top \mathrm{d}t \end{aligned}$$

with $\bar{X}_g(t) = \dfrac{1}{|g|}\sum_{i,s_i\in g} X_i(t)$ the empirical estimators of $\mu_g(t)$ for $g \in \{w, w^{\mathsf{c}}\}$ and $\bar{X}(t) = \dfrac{1}{n}\sum_{i=1}^{n} X_i(t)$ the empirical estimator of $\mu_S(t)$.

Then, Frévent et al. (2021b) considered $\mathrm{LH}^{(w)}$ as a concentration index and proposed the parametric multivariate functional spatial scan statistic (MPFSS):

$$\Lambda_{\mathrm{MPFSS}} = \max_{w\in\mathcal{W}} \mathrm{LH}^{(w)}.$$

In fact Górecki and Smaga (2017) proposed four test statistics using the matrices H_w and E_w to compare the mean functions in w and w^{c}: (1) the Lawley–Hotelling trace test statistic $\mathrm{LH}^{(w)} = \mathrm{Trace}(H_w E_w^{-1})$, (2) the Pillai trace test statistic $\mathrm{P}^{(w)} = \mathrm{Trace}(H_w(H_w+E_w)^{-1})$, (3) the Roy's largest root test statistic $\mathrm{R}^{(w)} = \lambda_{\max}(H_w E_w^{-1})$ where $\lambda_{\max}(H_w E_w^{-1})$ is the maximum eigenvalue of $H_w E_w^{-1}$ and (4) the Wilks lambda test statistic $\mathrm{W}^{(w)} = \dfrac{\det(E_w)}{\det(H_w+E_w)}$.
Thus each of these quantities (or the opposite for the Wilks lambda test statistic) can be considered as a concentration index and maximized over $\mathcal{W}$ which results in the following parametric multivariate functional spatial scan statistics:

$$\Lambda_{\mathrm{LH}} = \max_{w\in\mathcal{W}} \mathrm{LH}^{(w)}, \quad \Lambda_{\mathrm{P}} = \max_{w\in\mathcal{W}} \mathrm{P}^{(w)}, \quad \Lambda_{\mathrm{R}} = \max_{w\in\mathcal{W}} \mathrm{R}^{(w)}, \quad \Lambda_{\mathrm{W}} = \min_{w\in\mathcal{W}} \mathrm{W}^{(w)}.$$

These four approaches are implemented in the package **HDSpatialScan**.

4.2.3.2 A distribution-free spatial scan statistic for multivariate functional data

Frévent et al. (2021b) proposed a distribution-free spatial scan statistic for multivariate functional data which is the counterpart of the distribution-free spatial scan statistic for univariate functional data developed by Frévent et al. (2021a). They supposed that for each time t, $\mathbb{V}[X_i(t)] = \Sigma(t,t)$ for all $i \in [\![1;n]\!]$, where Σ is a $p \times p$ covariance matrix function.

Thus, as previously, in the context of cluster detection, the null hypothesis $\mathcal{H}_0$ can be defined as follows: $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^{\mathsf{c}}} = \mu_S$, where μ_w, $\mu_{w^{\mathsf{c}}}$ and μ_S stand for the mean functions in w, outside w and over S, respectively. And the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as follows: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^{\mathsf{c}}}$. Next, Qiu et al. (2021) proposed to compare the mean function μ_w in w with the mean function $\mu_{w^{\mathsf{c}}}$ in w^{c} by using the following statistic:

$$T_{n,\max}^{(w)} = \sup_{t\in\mathcal{T}} T_n(t)^{(w)}$$

where $T_n(t)$ is a pointwise statistic defined by the Hotelling T^2-test statistic

$$T_n(t)^{(w)} = \frac{|w||w^{\mathsf{c}}|}{n}(\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t))^{\top}\hat{\Sigma}(t,t)^{-1}(\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)).$$

$\bar{X}_w(t)$ and $\bar{X}_{w^c}(t)$ are the empirical estimators of the mean functions defined previously, and

$$\hat{\Sigma}(s,t) = \frac{1}{n-2}\left[\sum_{i,s_i\in w}(X_i(s)-\bar{X}_w(s))(X_i(t)-\bar{X}_w(t))^\top + \sum_{i,s_i\in w^c}(X_i(s)-\bar{X}_{w^c}(s))(X_i(t)-\bar{X}_{w^c}(t))^\top\right]$$

is the pooled sample covariance matrix function.

Then Frévent et al. (2021b) proposed to use $T^{(w)}_{n,\max}$ as a concentration index and to maximize it over the set of potential clusters $\mathcal{W}$: the multivariate distribution-free functional spatial scan statistic (MDFFSS) is

$$\Lambda_{\text{MDFFSS}} = \max_{w\in\mathcal{W}} T^{(w)}_{n,\max}.$$

4.2.3.3 A rank-based spatial scan statistic for multivariate functional data

Finally Frévent et al. (2021b) also proposed to consider as a pointwise test statistic the multivariate extension of the Wilcoxon rank-sum test statistic developed by Oja and Randles (2004) and detailed in subsection 4.2.1. They defined the pointwise multivariate ranks as $R_i(t) = \frac{1}{n}\sum_{j=1}^{n}\text{sgn}(A_X(t)(X_i(t)-X_j(t)))$ where the pointwise transformation matrix $A_X(t)$ is so that $\frac{p}{n}\sum_{i=1}^{n}R_i(t)R_i(t)^\top = \frac{1}{n}\sum_{i=1}^{n}R_i(t)^\top R_i(t) I_p$, and the sgn function is the same as in subsection 4.2.1.

Then for each time t, the pointwise multivariate extension of the Wilcoxon rank-sum test statistic is defined as

$$W(t)^{(w)} = \frac{pn}{\sum_{i=1}^{n}R_i(t)^\top R_i(t)}\left[\ |w|\ ||\bar{R}_w(t)||_2^2 + |w^c|\ ||\bar{R}_{w^c}(t)||_2^2\ \right]$$

where

$$\bar{R}_g(t) = \frac{1}{|g|}\sum_{i,s_i\in g}R_i(t)\ (g\in\{w,w^c\}).$$

In the context of cluster detection, the null hypothesis is defined as $\mathcal{H}_0$: $\forall w\in\mathcal{W},\ \forall t\in\mathcal{T},\ F_{w,t}=F_{w^c,t}$ where $F_{w,t}$ and $F_{w^c,t}$ correspond respectively to the cumulative distribution functions of $X(t)$ in w and outside w. The alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w is $\mathcal{H}_1^{(w)}$: $\exists t\in\mathcal{T},\ F_{w,t}(x)=F_{w^c,t}(x-\Delta_t)$, $\Delta_t\neq 0$.

Finally Frévent et al. (2021b) proposed to globalize the information over the time with the quantity

$$W^{(w)} = \sup_{t\in\mathcal{T}} W(t)^{(w)}$$

and to use it as a concentration index to be maximized over the set of potential clusters $\mathcal{W}$. The multivariate rank-based functional spatial scan statistic (MRBFSS) is then

$$\Lambda_{\text{MRBFSS}} = \max_{w \in \mathcal{W}} W^{(w)}.$$

4.2.4 Computing the significance of the MLC

Once the most likely cluster (MLC) is detected, its significance must be evaluated. The distribution of the scan statistic $\mathcal{S}$ ($\mathcal{S} = \lambda_{\text{MG}}$, λ_{MNP}, Λ_{PFSS}, Λ_{NPFSS}, Λ_{DFFSS}, Λ_{URBFSS}, Λ_{MPFSS}, Λ_{MDFFSS} or Λ_{MRBFSS}) is untractable under $\mathcal{H}_0$ due to the overlapping nature of $\mathcal{W}$. Then we choose to obtain a large set of simulated datasets by randomly permuting the observations X_i in the spatial locations. This technique was already used in spatial scan statistics (Cucala et al. , 2017; Kulldorff et al. , 2009).

Let M denote the number of random permutations of the original dataset and $\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(M)}$ be the observed scan statistics on the simulated datasets. According to Dwass (1957) the p-value for $\mathcal{S}$ observed in the real data is estimated by

$$\hat{p} = \frac{1 + \sum_{m=1}^{M} \mathbb{1}_{\mathcal{S}^{(m)} \geq \mathcal{S}}}{M + 1}. \tag{4.2}$$

Finally, the MLC is considered to be statistically significant if the associated $\hat{p}$ is less than the type I error.

4.2.5 How to choose the method to apply on the data ?

According to Cucala et al. (2019) the MNP method tends to present a better power and higher true positive rates for non-Gaussian data than the MG one. Although the false positive rates are often higher for this approach than the MG one, it remains moderate. In the functional framework, the approaches that present the best results are the DFFSS and the URBFSS in the univariate context and the MDFFSS and the MRBFSS in the multivariate one (Frévent et al. , 2021b; Frévent et al. , 2021a). The URBFSS and the MRBFSS tend to show higher powers and higher true positive (TP) rates although they detect more false positives (FP) than the DFFSS and the MDFFSS respectively. Table 4.1 summarizes the methods and their performances. The symbols ✓ and ✗ indicate respectively a high and a low performance on the criterion. If there is no symbol it means that for this criterion the approach offers medium performances. The terminology "localized clusters in time" in the functional cases refers to aggregations of sites that take higher or lower values for the process only in a small interval of time.

	Gaussian distribution			**Non-Gaussian distribution**		
Method	Power	TP rate	FP rate	Power	TP rate	FP rate
Multivariate data				Multivariate data		
MG	✓	✓	✓	X	X	✓
MNP	✓	✓		✓	✓	
Functional univariate data				Functional univariate data		
Non localized clusters in time				Non localized clusters in time		
PFSS			✓	X		✓
DFFSS	✓	✓	✓	✓	✓	✓
NPFSS		✓			✓	
URBFSS	✓	✓		✓	✓	
Localized clusters in time				Localized clusters in time		
PFSS	X	X		X	X	
DFFSS	✓	✓	✓	✓	✓	✓
NPFSS	X			X		
URBFSS	✓	✓	✓	✓	✓	✓
Functional multivariate data				Functional multivariate data		
Non localized clusters in time				Non localized clusters in time		
MPFSS			✓	X	X	✓
MDFFSS	✓	✓	✓			✓
NPFSS		✓		✓	✓	
MRBFSS	✓	✓		✓	✓	
Localized clusters in time				Localized clusters in time		
MPFSS	X	X		X	X	✓
MDFFSS	✓	✓	✓			✓
NPFSS	X			X		
MRBFSS	✓	✓	✓	✓	✓	✓

Table 4.1: Performance in terms of power, TP rate and FP rate of spatial scan statistics for multivariate data (MG and MNP), univariate functional data (PFSS, DFFSS, NPFSS and URBFSS) and multivariate functional data (MPFSS, MDFFSS, NPFSS and MRBFSS)

4.3 Software

4.3.1 Computing the spatial scan statistic

The package **HDSpatialScan** provides the functions to compute all the spatial scan statistics described above: `MG()` and `MNP()` functions apply respectively the parametric and nonparametric spatial scan statistics approaches on multivariate data. Their univariate counterparts (when $p = 1$) can be computed with the functions `UG()` and `UNP()` respectively. Then the functions `PFSS()`, `DFFSS()` and `URBFSS()` apply the parametric, the distribution-free and the new rank-based functional approaches on univariate functional data, and `MPFSS()`, `MDFFSS()`, and `MRBFSS()` are their multivariate counterparts. Finally the function `NPFSS()` applies the nonparametric spatial scan statistic for functional data developed by Smida et al. (2021) on both univariate and multivariate functional data.

4.3.1.1 Type of the data

Depending on the type of approach (univariate, multivariate, functional univariate or functional multivariate), the data must be formatted in a specific way. For univariate approaches, the data must be a vector in which each element corresponds to a site. If the data is individual and many individuals share the same site, the data can remain in an individual format with one element of the vector per individual. Then for real-valued multivariate methods or functional univariate methods, the data must be a matrix in which each row corresponds to a site (or an individual) and each column corresponds to a variable or an observation time in the functional framework. For multivariate functional methods the data must be a list in which each element is a matrix corresponding to a site (or an individual). In the matrices, the rows correspond to the variables and the columns to the observation times. Note that the observation times must be the same for each site or individual and they must be equally spaced for the functions `NPFSS()`, `PFSS()` and `MPFSS()`. However if it is not the case of the raw data, they can be easily transformed by smoothing the data (Ramsay and Silverman , 2005), by using for example the R package **fda** (Ramsay et al. , 2020).

4.3.1.2 Parameters of the functions

All the functions listed above share the same arguments: `data`, `sites_coord`, `system`, `mini`, `maxi`, `type_minimaxi`, `mini_post`, `maxi_post`, `type_minimaxi_post`, `sites_areas`, `MC`, `typeI` and `nbCPU` (except for the functions `UG()` and `UNP()`).

The first argument, `data`, is the data vector, matrix or list on which the approach must be applied. `MC` and `typeI` correspond respectively to the number of permutations of the data while computing the significance of the clusters and the type I error i.e. a cluster is declared significant if its estimated p-value is below this threshold.

The arguments `sites_coord` and `system` are respectively a matrix of two columns corresponding to the coordinates of each site or individual, and to the system of coordinates (“Euclidean” or “WGS84”).

The `sites_areas` argument is optional and corresponds to the areas of the sites (or the site of each individual if the data is individual).

Finally the argument `nbCPU` permits to do parallelization and the arguments `mini`, `maxi`, `type_minimaxi`, `mini_post`, `maxi_post`, `type_minimaxi_post` are described further below.

The `MPFSS()` function has an extra argument `method` which is a character vector that can take its values in {“LH”, “W”, “P”, “R”}. They correspond to the MANOVA approach to use: the Lawley-Hotelling trace test, the Wilks’ Lambda

test, the Pillai's trace test or the Roy's largest root test. Although the Lawley-Hotelling trace test is the most used statistic (Oja and Randles , 2004), it should be noted that all these methods usually provide very similar results. By default the four MANOVA approaches are computed.

***A priori* filtering** The clusters are computed automatically as circular clusters, so we need to define a minimum and a maximum size for these clusters. That is what we call "*a priori* filtering" and this allows to control the computation time. Three types of *a priori* filtering are possible through the argument `type_minimaxi`: "sites/indiv" (the filtering is applied on the number of sites or individuals in the potential clusters, it is the default value), "area" (it is applied on the area of the clusters and is available only if `sites_areas` is provided), or "radius" (the radius of the clusters).
The arguments `mini` and `maxi` are then respectively the minimum number of sites/individuals, or the minimal area or radius and the maximum number of sites/individuals, or the maximal area or radius. For the radius it is specified in km if `system` is "WGS84" or in the user units if `system` is "Euclidean".
It should be noted that this filtering can bias the value of the p-values obtained for the clusters. In order to perform a correct statistical inference Kulldorff and Nagarwalla (1995) recommended to consider a maximum size of half the study region. Thus the default setting is to consider potential clusters comprising at least one site and at most 50% of the sites (equation 4.1). If you want to select clusters according to size (number of sites or individuals), area or radius, it is better to select them *a posteriori* among the detected clusters and if you really want to decrease the computation time we recommend to increase the number of CPU (with the argument `nb_CPU`). Changing the default settings can allow the user to investigate whether there appear to be clusters in a relatively quick first step, although the inference is biased, before applying the scan procedure with the default settings for the *a priori* filtering (50% of the studied region).

***A posteriori* filtering** Sometimes after that the p-value of each potential cluster has been computing, the user may want to retrieve only the significant clusters that satisfy a certain size, area, or radius criteria. That is what we call *a posteriori* filtering. The corresponding arguments are `mini_post`, `maxi_post` and `type_minimaxi_post` and their definitions are the same as `mini`, `maxi` and `type_minimaxi`. If the user only wants to obtain clusters meeting size criteria, this *a posteriori* approach must be prioritized over the *a priori* approach which gives biased results and must therefore be used with great care.

4.3.1.3 Output of the functions

The scan statistics functions outputs are composed of the following elements: `sites_clusters`, `pval_clusters`, `centres_clusters`, `radius_clusters`,

`areas_clusters` and `system`.

The element `sites_clusters` is a list in which each element corresponds to a significant cluster and contains the index of the sites (or the individuals) included in this cluster. The clusters are listed in their order of detection. The secondary clusters are defined according to Kulldorff (1997): they correspond to potential clusters that also present large values for the concentration index. Their p-values are calculated as if they were the most likely cluster themselves which is a bit conservative since the secondary clusters are compared with the most likely cluster of the permutations (Kulldorff , 1997). Finally, only clusters that are significant at the `typeI` threshold and that do not overlap with a more likely cluster are returned, and `pval_clusters` corresponds to the associated p-values.

The element `centres_clusters` corresponds to the coordinates of the centres of each detected cluster and `radius_clusters` is the radius of the clusters in km if `system` is "WGS84" or in the user units otherwise. The system of coordinates is recalled in the element `system`.

Finally `areas_clusters` corresponds to the areas of the clusters (in the same units as `sites_areas`). This element is only provided if `sites_areas` is not `NULL`.

These elements are duplicated in the `MPFSS()` output with one element per method used.

4.3.2 Plot or summarize the results

In order to plot the detected clusters the package **HDSpatialScan** provides three functions that give different types of plot.

The first one, `plot_map()` allows the user to plot a map of the sites and draws the circles corresponding to the circular clusters. The arguments `centres` and `radius` are returned by the scan functions.

In the second function `plot_map2()`, the argument `sites_coord` corresponds to the matrix of the sites' coordinates (or the individuals' coordinates) in the same order as the data in the scan procedure and `output_clusters` is the element `sites_clusters` in the output of the scan functions.

These two functions also have the arguments `system` which takes the value "Euclidean" or "WGS84" to correctly display the map and the detected clusters, and `spobject` which is the spatial object corresponding the sites.

If you do not have this object you can use the third function `plot_schema()` which simply draws a schema of the sites and the clusters, with the arguments `output_clusters` corresponding to the element `sites_clusters` in the output of

the scan functions, `sites_coord` which is the matrix of the sites' coordinates (or the individuals' coordinates) in the same order as the data in the scan procedure, `system` ("Euclidean" or "WGS84") and `system_conv`. The latter allows to correctly project the coordinates. It must be entered as in the PROJ documentation (PROJ contributors , 2021).

One may also want to get some features of one's clusters.

The function `plot_summary()` allows to get a summary of the clusters, either the mean and the standard deviation of each of the variables (if many) over the time (for functional data) if the argument `type_summ` is "param", or the 25th percentiles, the medians and the 75th percentiles if the argument `type_summ` is "nparam". The other important arguments are `output_clusters` corresponding to the element `sites_clusters` in the output of the scan functions, `data` which is the data provided for the scan procedure and `type` which allows to differentiate the univariate functional case ("funct") from the multivariate non-functional case ("multi"). Two types of output are available: if the argument `html` is `TRUE` the function displays an html table which allows a better readability but a less good extraction of the information than if `html` is `FALSE`: in the latter case the table is simply returned in the console.

Other interesting functions are `plot_curves_clusters()` that allows to display cluster curves (in the functional case), and `plot_summary_curves()` which displays the average (if `type` is "mean") or the median (if `type` is "median") curves in the clusters, outside and the global mean or median curves (in the functional case). For the multivariate non-functional framework a function `plot_summary_chart()` is available. It displays a spider chart of means or medians for each variable inside the cluster, outside, or in all the area. All these functions take an argument `variable_names` which is the vector of the names of the variables. By default its value is `NULL`: the names of the variables that appear on the plots are "var1", "var2", etc. Note that the value of this argument is only considered for multivariate functional data or multivariate non-functional data.

4.4 Illustrations

To show the simplicity of use of the package, we will apply the different approaches on the environmental data provided in the package.

4.4.1 Air pollution in northern France

We considered data provided by the French national air quality forecasting platform PREV'AIR which is available in the package **HDSpatialScan**. It consists in the daily concentrations (from May 1, 2020 to June 25, 2020) in $\mu g.m^{-3}$ of four pollutants for each of the 169 *cantons* (administrative subdivisions of France) of

the *Nord-Pas-de-Calais* (a region in northern France) located by their center of gravity: nitrogen dioxide (NO_2), ozone (O_3) and fine particles PM_{10} and $PM_{2.5}$ corresponding respectively to particles whose diameter is less than $10\mu m$ and $2.5\mu m$. The package **HDSpatialScan** provides the full data: `fmulti_data` but also some reduced data for the univariate functional case which consists in condering only the NO_2 concentrations (`funi_data`), and for the multivariate non-functional framework (`multi_data`) which corresponds to the temporal mean concentrations of the four pollutants over the studied period.

- The first step is to load the data:

```
R> library(HDSpatialScan)
R> data("map_sites")
R> data("multi_data")
R> data("funi_data")
R> data("fmulti_data")
```

- The second step is to visualize the pollutants daily concentration curves in each *canton* and the spatial distributions of the temporal mean concentrations for each pollutant over the studied time period (Figures 4.1 and 4.2). This step allows us to see if sites seem to aggregate and therefore if launching a cluster detection is relevant, and if a temporal variation of the concentrations is visible, in which case a functional method will be more relevant than a multivariate approach summarizing each curve by its mean.

  The maps in Figure 4.2 show a spatial heterogeneity of the average concentration for each pollutant. Thus spatial scan statistics seem to be suitable to highlight the presence of *cantons*-level spatial clusters of pollutants concentrations. Moreover since the curves in Figure 4.1 show a marked temporal variability during the period from May 1, 2020 to June 25, 2020 a functional approach is more appropriate. However for sake of completeness we will also perform a multivariate spatial scan statistic approach anyway. Since small clusters of pollution are more relevant for interpretation because the sources of the pollutants are very localized, we will consider an *a posteriori* filtering of maximum radius equal to 10 km.

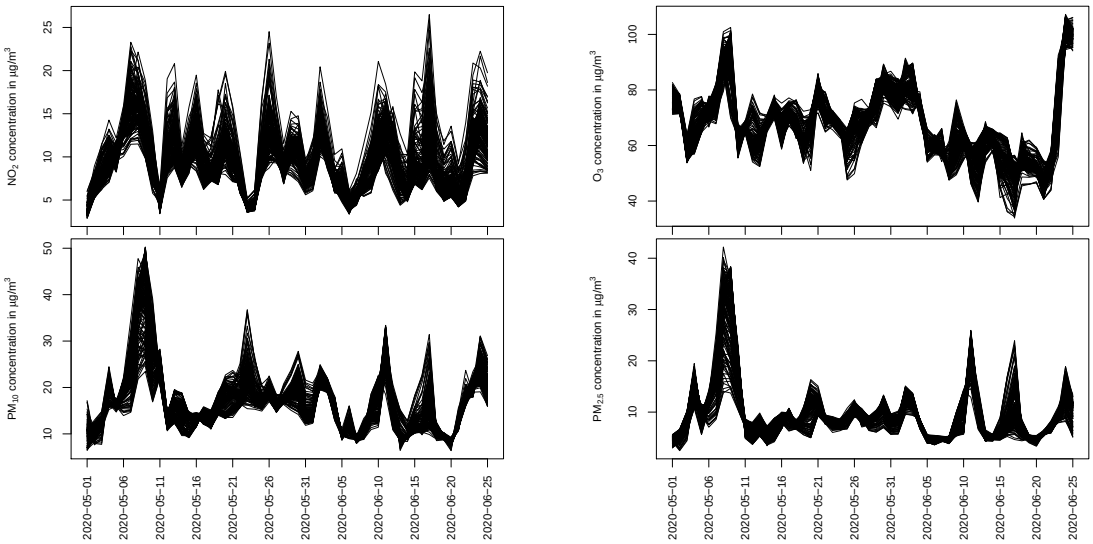

Figure 4.1: Daily concentration curves of NO_2, O_3, PM_{10} and $PM_{2.5}$ (from May 1, 2020 to June 25, 2020) in each of the 169 *cantons* of *Nord-Pas-de-Calais* (a region in northern France).

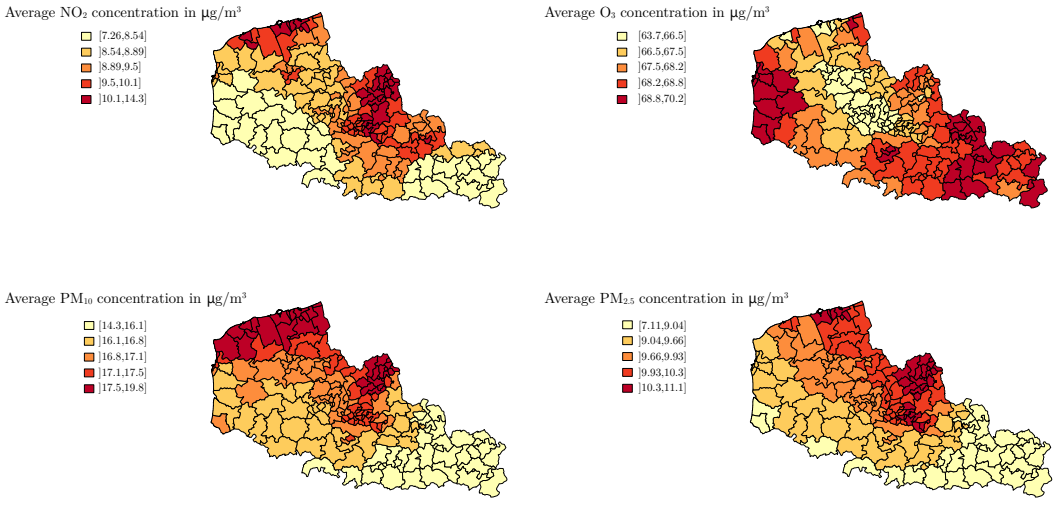

Figure 4.2: Spatial distributions of the average concentrations of NO_2, O_3, PM_{10} and $PM_{2.5}$ over period from from May 1, 2020 to June 25, 2020.

4.4.2 A multivariate spatial scan statistic

First we will investigate a multivariate spatial scan statistic. Since the distribution of the pollutants temporal mean concentrations is non-normal we decide to apply the MNP (function `MNP()`) scan procedure. Here the system of coordinates is "WGS84", it must be filled with the argument `system`. As explained in subsection 4.3.1, Kulldorff and Nagarwalla (1995) recommanded to consider a maximum size of half the study region for the potential clusters so we use this *a priori* filtering with the parameters `mini`, `maxi` and `type_minimaxi`: the potential clusters are circular and they contain between 1 and 50% of the sites. Then as noticed in subsection 4.4.1 we will apply an *a posteriori* filtering of maximum radius equal to 10 km (arguments `mini_post`, `maxi_post` and `type_minimaxi_post`). Here we only want to consider the significant clusters at the 5% threshold. Thus we leave the `typeI` parameter at its default value (0.05). However it should be noted that it is possible to obtain all the clusters (the MLC and the secondary clusters (Kulldorff , 1997)) by setting the `typeI` value at 1.

```
R> library(sp)
R> coords <- coordinates(map_sites)
R> res_mnp <- MNP(data = multi_data, sites_coord = coords, system
= "WGS84", + mini = 1, maxi = nrow(coords)/2, type_minimaxi =
"sites/indiv", + mini_post = 0, maxi_post = 10, type_minimaxi_post
= "radius", nbCPU = 7)
```

Once the scan procedure is completed, the plot functions can be used. For the sake of brevity we choose here and in the following to only focus on the MLC and for the sake of completeness we will show the use of the three possible visualization functions of the clusters. Since we have a spatial object `map_sites` we can use the functions `plot_map()` and `plot_map2()`. However for sake of completeness we also show the use of the function `plot_schema()` which allows to display the clusters otherwise (Figure 4.3). Since the system of the coordinates is "WGS84", this function requires to complete the parameter `system_conv` which allows to correctly project the points. Here we choose the EPSG code 2154 corresponding to the Lambert 93 projection since the data is located in metropolitan France.

```
R> plot_map(spobject=map_sites, centres=res_mnp$centres_clusters[1,],
+ radius = res_mnp$radius_clusters[1], system = "WGS84")
R> plot_map2(spobject = map_sites, sites_coord = coords,
+ output_clusters = res_mnp$sites_clusters[1], system = "WGS84")
R> plot_schema(output_clusters = res_mnp$sites_clusters[1],
sites_coord = coords, + system = "WGS84", system_conv =
"+init=epsg:2154")
```

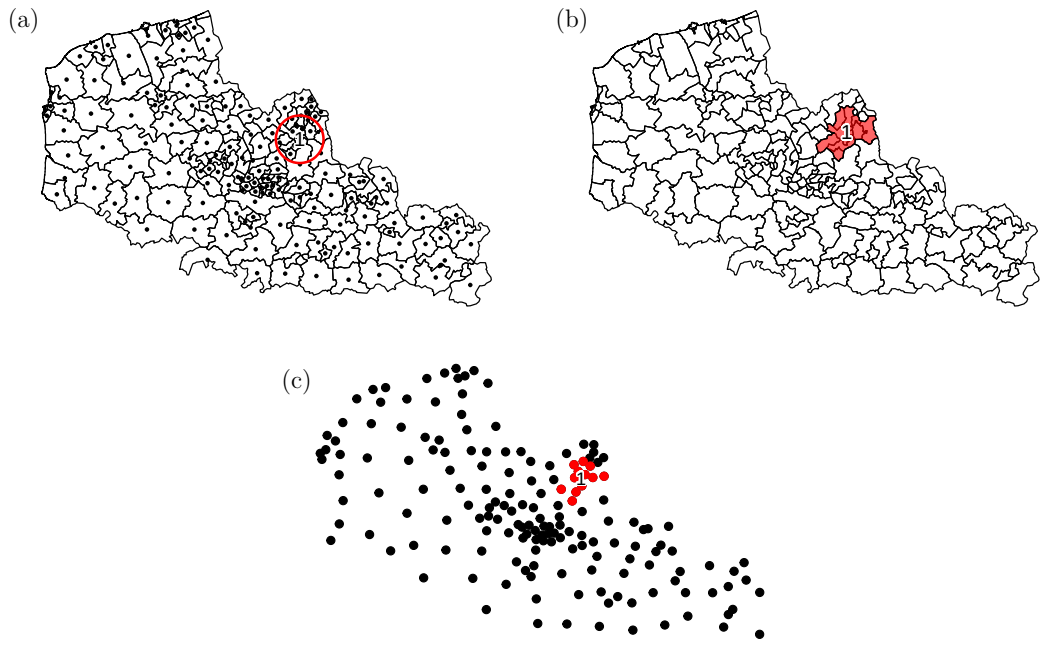

Figure 4.3: Visualization of the most likely cluster with the functions `plot_map()` (panel a), `plot_map2()` (panel b) and `plot_schema()` (panel c) for the MNP scan procedure with the function `MNP()`

Finally users may want to get some summarized characteristics, such as the quantiles of the variables in a html table. This can be achieved by using the function `plot_summary()` with the argument `type_summ` equal to "nparam" (for the quantiles) (Figure 4.4):

```
R> plot_summary(output_clusters = res_mnp$sites_clusters[1],
data = multi_data, + type = "multi", type_summ = "nparam",
nb_digits = 1, html = TRUE, + variable_names = c("NO2","O3","PM10",
"PM2.5"))
```

Or we may prefer to display it in a simpler table in the console to extract more easily the information by setting the `html` parameter to `FALSE`:

```
R> plot_summary(output_clusters = res_mnp$sites_clusters[1],
data = multi_data, + type = "multi", type_summ = "nparam",
nb_digits = 1, html = FALSE, + variable_names = c("NO2","O3","PM10",
"PM2.5"))
```

	Overall	Inside cluster 1	Outside cluster 1
Number of sites	169.0	12.0	157.0
Q25 NO2	8.7	11.3	8.6
Median NO2	9.2	11.7	9.1
Q75 NO2	9.8	12.4	9.7
Q25 O3	66.8	67.5	66.7
Median O3	67.9	67.6	68.0
Q75 O3	68.6	67.9	68.7
Q25 PM10	16.4	17.5	16.2
Median PM10	17.0	17.9	16.9
Q75 PM10	17.4	18.0	17.3
Q25 PM2.5	9.1	10.6	9.1
Median PM2.5	9.8	10.7	9.8
Q75 PM2.5	10.2	10.9	10.1

The user can also use the function `plot_summary_chart()` to display the spider chart corresponding to the detected cluster (Figure 4.5).

```
R> plot_summary_chart(output_clusters = res_mnp$sites_clusters[1],
+ data = multi_data, variable_names = c("NO2","O3","PM10","PM2.5"),
type = "median")
```

The MLC is located in the area of Lille. Figures 4.4 and 4.5 show that it is a cluster of overpollution (except for the pollutant O_3). This result is consistent since it is well-known that the pollutants (except O_3) are more frequent in urban areas.

We have obtained some first results however the curves on Figure 4.1 present a marked temporal variability during the studied period. Thus it could be interesting to apply functional spatial scan statistics.

4.4.3 A univariate functional spatial scan statistic

In this subsection we only consider the pollutant NO_2. We choose to use the URBFSS (function `URBFSS()`) scan procedure since it often presents higher powers and true positive rates than the other univariate functional methods as its multivariate counterpart `MRBFSS()` (Frévent et al. , 2021b). As mentioned in 4.4.2 we decide to use the set of potential clusters *a priori* in the equation 4.1 which corresponds to the recommended approach of Kulldorff and Nagarwalla (1995), and to the default values of the parameters `mini`, `maxi` and `type_minimaxi` in the scan functions. We also set a maximum radius equal to 10 km *a posteriori*.

```
R> res_urbfss <- URBFSS(data = funi_data, sites_coord = coords,
system = "WGS84", + mini = 1, maxi = nrow(coords)/2,
type_minimaxi = "sites/indiv", + mini_post = 0, maxi_post = 10,
type_minimaxi_post = "radius", nbCPU = 7)
```

Show 25 entries Search:

	Overall	Inside cluster 1	Outside cluster 1
Number of sites	169	12	157
Q25 NO2	8.7	11.3	8.6
Median NO2	9.2	11.7	9.1
Q75 NO2	9.8	12.4	9.7
Q25 O3	66.8	67.5	66.7
Median O3	67.9	67.6	68
Q75 O3	68.6	67.9	68.7
Q25 PM10	16.4	17.5	16.2
Median PM10	17	17.9	16.9
Q75 PM10	17.4	18	17.3
Q25 PM2.5	9.1	10.6	9.1
Median PM2.5	9.8	10.7	9.8
Q75 PM2.5	10.2	10.9	10.1

Showing 1 to 13 of 13 entries Previous 1 Next

Figure 4.4: Characterization of the most likely cluster for the MNP scan approach in the context of multivariate data, with the function `plot_summary()` with `html = TRUE`

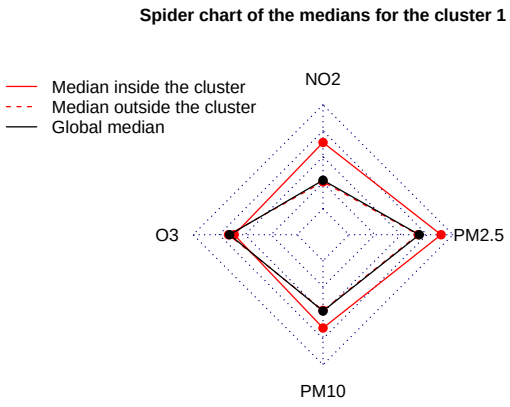

Figure 4.5: Spider chart for most likely cluster detected by the MNP scan procedure, obtained with the function `plot_summary_chart()`

```
R> plot_map2(spobject = map_sites, sites_coord = coords,
+ output_clusters = res_urbfss$sites_clusters[1], system = "WGS84")
```

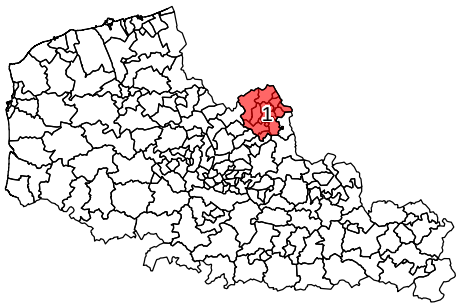

Figure 4.6: Visualization of the most likely cluster for the URBFSS scan procedure with the function `plot_map2()`

Again the MLC is located in the area of Lille (Figure 4.6).

For functional data other functions are provided to give some characteristics of the clusters: we can visualize the curves in the cluster by adding the curve of the global median, as well as the median curves inside and outside the cluster (Figure 4.7): this is a cluster of overexposure to NO_2.

```
R> plot_curves_clusters(output_clusters = res_urbfss$sites_clusters[1],
+ data = funi_data, times = c(1:ncol(funi_data)), add_median = TRUE)
R> plot_summary_curves(output_clusters = res_urbfss$sites_clusters[1],
+ data = funi_data, times = c(1:ncol(funi_data)), type = "median")
```

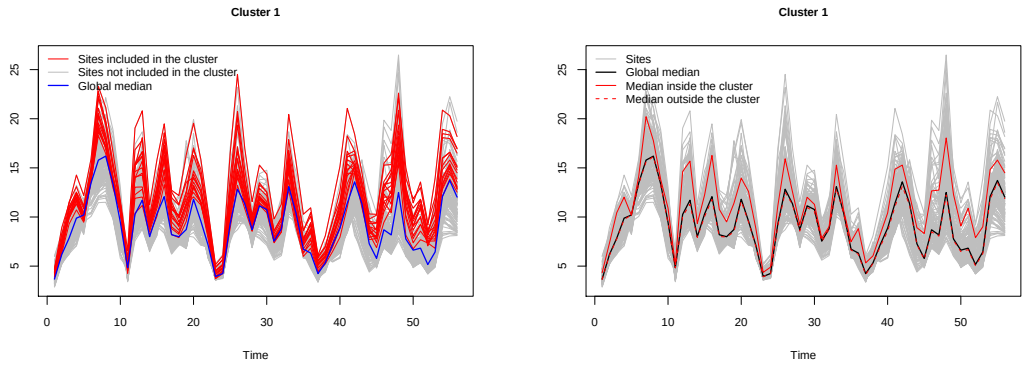

Figure 4.7: Characterization of the most likely cluster for the URBFSS scan approach in the context of univariate functional data with the functions `plot_curves_clusters()` (left panel) and `plot_summary_curves()` (right panel)

4.4.4 A functional multivariate spatial scan statistic

Now we consider the four pollutants together. For the same reason that we have previously chosen to apply the URBFSS scan procedure, we use the MRBFSS (function `MRBFSS()`) in this context, with the same restrictions *a priori* and *a posteriori* as for the MNP and the URBFSS scan approaches.

```
R> res_mrbfss <- MRBFSS(data = fmulti_data, sites_coord = coords,
system = "WGS84", + mini = 1, maxi = nrow(coords)/2, type_minimaxi
= "sites/indiv", + mini_post = 0, maxi_post = 10, type_minimaxi_post
= "radius", nbCPU = 7)
R> plot_map2(spobject = map_sites, sites_coord = coords,
+ output_clusters = res_mrbfss$sites_clusters[1], system = "WGS84")
```

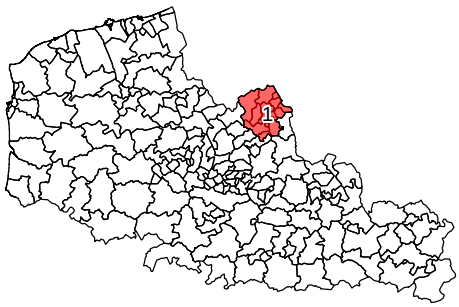

Figure 4.8: Visualization of the most likely cluster for the MRBFSS scan procedure with the function `plot_map2()`

The detected cluster is exactly the same as before and is therefore located in the urban area of Lille (Figure 4.8).

Again we will display the curves in the cluster by adding the curve of the global median (Figure 4.9), as well as the median curves inside and outside the cluster which show that this is a cluster of overexposure to NO_2, PM_{10} and $PM_{2.5}$ (Figure 4.10). As mentioned in subsection 4.4.2, in environmental science it is well-known that those pollutants are more frequent in urban areas so this is consistent with the cluster observed here.

```
R> plot_curves_clusters(output_clusters = res_mrbfss$sites_clusters[1],
+ data = fmulti_data, times = c(1:ncol(fmulti_data[[1]])), add_median
 = TRUE, + variable_names = c("NO2","O3","PM10","PM2.5"))

R> plot_summary_curves(output_clusters = res_mrbfss$sites_clusters[1],
+ data = fmulti_data, times = c(1:ncol(fmulti_data[[1]])), type =
"median", + variable_names = c("NO2","O3","PM10","PM2.5"))
```

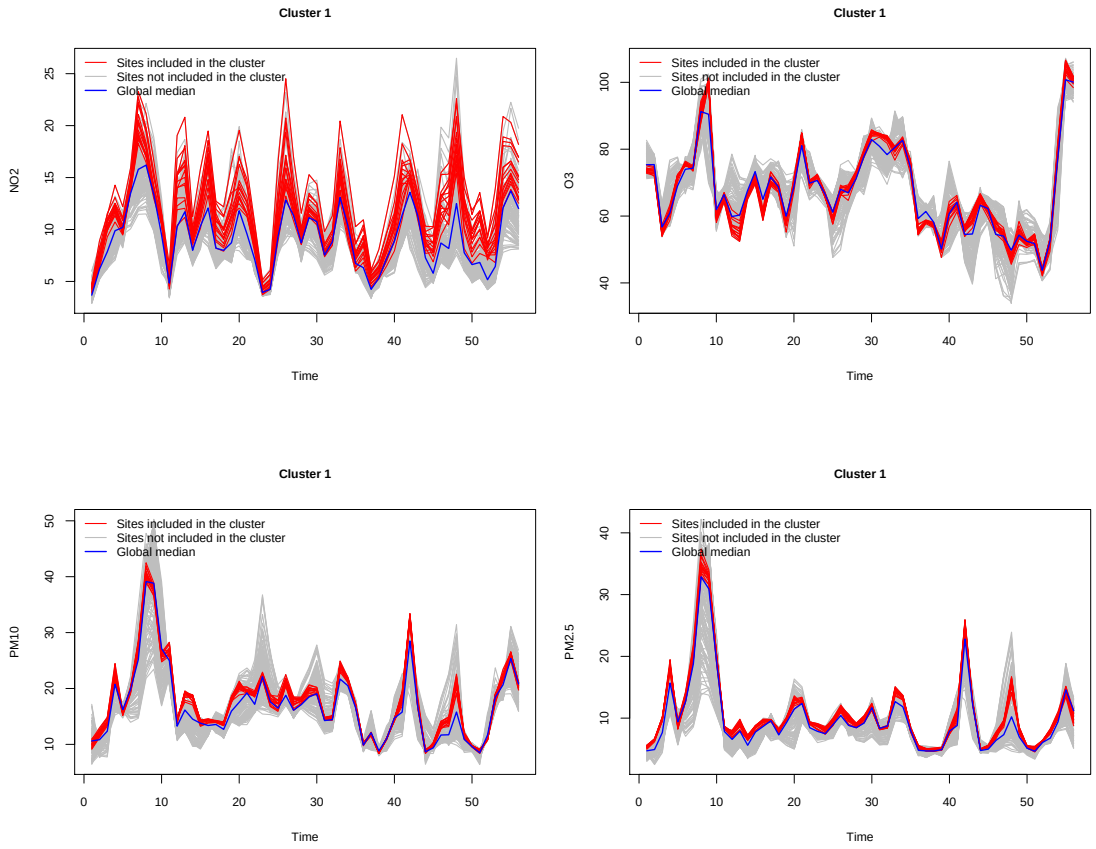

Figure 4.9: Characterization of the most likely cluster for the MRBFSS scan approach in the context of multivariate functional data with the function `plot_curves_clusters()`

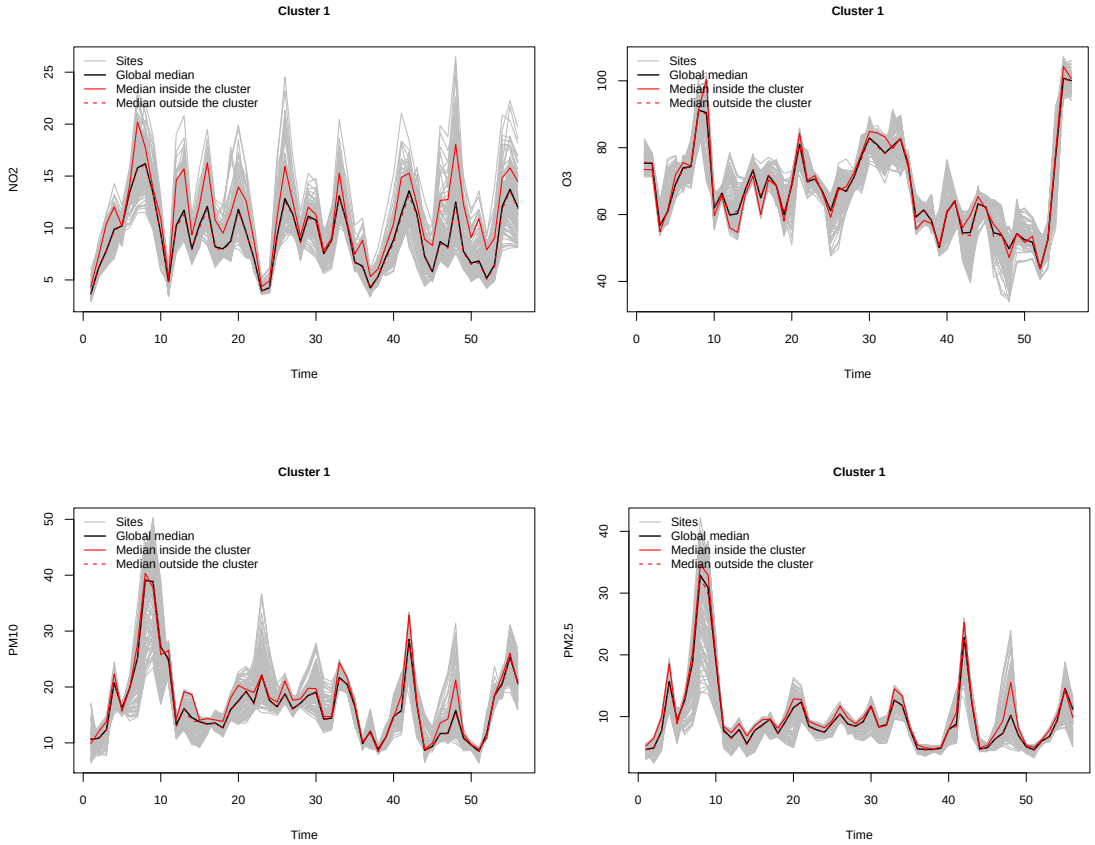

Figure 4.10: Characterization of the most likely cluster for the MRBFSS scan approach in the context of multivariate functional data with the function `plot_summary_curves()`

4.5 Conclusion

In this article we presented the **HDSpatialScan** package. It makes it very easy to apply the existing scan statistics developed for multivariate data or functional data (univariate or multivariate), and the new rank-based scan statistic for univariate functional data presented in the Section 4.2. The potential clusters considered are of variable size and circular. In further updates of the package **HDSpatialScan** other shapes of scanning window such as elliptical or rectangular shapes will be implemented. Our package also contains functions to easily plot and summarize the detected clusters. Then examples of applications of the functions of the package have been shown. **HDSpatialScan** presents the advantage that all scan functions have the same structure which makes it very quick to get started.

Chapter 5

Conclusion and perspectives

"Sometimes the ending of a documented story is really just a new beginning to the unpublished adventure yet to be discovered."

— Jes Fuhrmann

Chapter contents

5.1 General conclusion

In this thesis, we are interested first in the study of the nonparametric tests for the comparison of two independent samples in infinite dimensional space. Then, we are focused in the field of the cluster detection by the method of spatial scan statistics using functional data.

First part of Chapter 1 gives an overview of the existing nonparametric statistical tests for the comparison of two independent samples using real random variables. The second part is devoted to give the methodology of spatial scan statistics using univariate framework too. The last part provides some generalities and reminds some existing works dedicated to functional data in order to understand the constructions made in the next chapters.

Within the Chapter 2, two median test statistics have been proposed in Banach space and its equivalents in Hilbert one using functional data. We have proved the

asymptotic gaussianity of one of them using the central limit theorem in Banach of type 2 separable space (Araujo and Giné , 1980). Then, we have tested its performance compared to the parametric test statistics: two mean-based tests of Horváth et al. (2013) and the ANOVA test for functional data of Cuevas et al. (2004) and to a nonparametric test statistic: Wilcoxon-Mann-Whitney test in infinite dimensional space of Chakraborty and Chaudhuri (2015). Generated datasets from different type of symmetric and heavy-tailed distributions have been used. Simulation study shows that the median test statistics proposed have similar power than the Wilcoxon-Mann-Whitney one using a constant and quadratic location shift and it is more powerful than the parametric test statistics defined above. Also, the power curves of all the tests are close using linear location shift except in the case of the heavy-tailed Cauchy distribution where the median tests and the Wilcoxon-Mann-Whitney one are more poweful than the parametric ones. Application to real data is also made to illustrate the work.

In the Chapter 3, nonparametric spatial scan statistic (Smida et al. , 2021) is introduced using functional data and more precisely it is presented in Hilbert space. It can be extended to a general smooth Banach space by using a sign function adapted to this space as defined in Chapter 1. The construction of this scan statistic is based on the Wilcoxon-Mann-Whitney two-sample test statistic defined by Chakraborty and Chaudhuri (2015) in Hilbert space. To our knowlege, this scan statistic is the first one developed for functional data, so we decided to compare it with the univariate existing ones applied to summaries of the functional datasets. Simulation and real datasets have been used to test its performance. In application to real data, the proposed spatial scan statistic detects a spatial cluster of low demographic evolution rates in Spain. The implementation of a high-speed version of this scan statistic is now available in the `R` package `HDSpatialScan` developed and introduced in the Chapter 4.

Chapter 4 introduces the `R` package `HDSpatialScan` for functional data. In this work, the nonparametric spatial scan statistic introduced in Chapter 3 (Smida et al. , 2021) and all the parametric and nonparametric ones developed very recently (Frévent et al. , 2021b; Frévent et al. , 2021a) are implemented to allow users to apply easily the desired scan method, to plot the detected clusters and to summarize them. Illustrations using real data available in this package are presented and it shows that this package uses a numerous scan functions which have the same structure.

The next section presents some of research perspectives that we would like to develop and some current work progress.

5.2 Perspectives and current studies

Nowadays, functional data analysis becomes a huge branch of statistics due to the expansion of the sensoring and the increasing use of time series data. According to the work done in this manuscript, the median test proposed in Chapter 2 and scan statistics using these type of data are proving to be a field of research offering plenty of perspectives, both theoretically and applied.

In Chapter 2, the asymptotic null distribution of one of the median test statistics proposed, denoted by MED, is obtained in Theorem 2.1. However, we have used, for computing its significance, a test method based on Monte-Carlo simulations whose basic aim is to approximate the null distribution. We have not used the asymptotic null distribution given by Theorem 2.1 because it suffers two limitations which are mentioned in the subsection 2.2.3. A perspective would be to compute the critical values obtained from Theorem 2.1 and based on a suitable estimations of the covariance operators Γ_1 and Γ_2 which are given respectively by equation (2.7) and equation (2.8).

According to Chapter 3, a new nonparametric spatial scan statistic using functional data could be built using the median test statistic introduced in the Chapter 2 since its asymptotic distribution under the null hypothesis is known. More precisely, we introduce this spatial scan statistic in subsection 5.2.1. A first application to the real data analysed in Chapter 3 is presented in the next section. Remark that this scan statistic detects a significant cluster of 19 Spanish provinces. This cluster includes provinces which have the lowest demographic evolution (see Figure 5.2). An application using generated data is planned to compare this new statistic to the nonparametric scan statistic developed in the Chapter 3 and the ones proposed by Frévent et al. (2021a).

Originally, univariate spatial scan statistic was constructed using a likelihood ratio test statistic. As shown in Chapter 1, there is a relationship between the likelihood ratio based on Gaussian model (Kulldorff et al. , 2009) and the Student T^2 test statistic using real random variable. Similary, relationship between the multivariate scan statistic and the Hotelling T^2 is proved in Anderson (2003). Therefore, we propose a new parametric spatial scan statistic for functional data based on one of the test statistics presented in Horváth et al. (2013). Detailed construction of this scan statistic and an application to the Spanish real data is presented now in subsection 5.2.2. Simulation study to compare its performance to the nonparametric and parametric scan statistics is still to be done.

After developing these new spatial scan statistics, we could add them to the `R` package `HDSpatialScan`.

5.2.1 A median spatial scan statistic for functional data

5.2.1.1 Statistic construction

Similarly to the work done in the Chapter 3, we consider a random element taking values in a functional space χ. We suppose that χ is an Hilbert space such as $L^2([0,1],\mathbb{R})$. Let $s_1,\ldots,s_n$ be n different spatial locations included in an observation domain $D \subset \mathbb{R}^2$ and $X_1,\ldots,X_n$ be the observations of X measured in $s_1,\ldots,s_n$.
To construct a new nonparametric spatial scan statistic in the functional case, we consider, the set of potential clusters $\mathcal{S}$ defined in the Chapter 3 (see equation (3.1)). The only nonparametric scan statistics developed in the literature, are constructed using the test of Wilcoxon-Mann-Whitney in the different settings: univariate (Cucala , 2016), multivariate (Cucala et al. , 2019) and recently functional (Smida et al. , 2021). Hence, the idea is to develop a new scan statistic using another nonparametric test for equality of distributions using functional data. More precisely, a possible new concentration index relies on the median test statistic (Chapter 2) since its asymptotic distribution is known and proved in the appendix of Chapter 2.
Assume that $X_1,\ldots,X_n$ are independent and $Z \in \mathcal{S}$ 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. Suppose that the marks in Z and Z^c respectively follow probability measures P_Z and P_{Z^c} on χ. We assume that P_Z and P_{Z^c} differ by a shift $\Delta_Z \in \chi$. Thus, the null hypothesis H_0 (the absence of cluster) can be defined as $H_0 : \Delta_Z = 0$ against the alternative ones $H_{1,Z} : \Delta_Z \neq 0$. Then, the functional median statistic for testing the hypothesis H_0 against $H_{1,Z}$ is defined as

$$\mathrm{MED}(Z) = \frac{1}{n_{Z^c}} \sum_{\{i,s_i \in Z^c\}} \frac{\displaystyle\sum_{\{j,s_j \in Z\}} \frac{X_i - X_j}{\|X_i - X_j\|_\chi}}{\left\| \displaystyle\sum_{\{j,s_j \in Z\}} \frac{X_i - X_j}{\|X_i - X_j\|} \right\|_\chi}, \tag{5.1}$$

where $\|.\|_\chi$ is the norm on the Hilbert space χ. This statistic is the same as (2.3) when m and n are respectively equal to n_Z and n_{Z^c}. We have used this statistic instead of (2.4) because its asymptotic distribution is known. Thus, a standardized concentration index is given by

$$\mathrm{M}(Z) = \sqrt{\frac{n_Z n_{Z^c}}{n}} \mathrm{MED}(Z), \tag{5.2}$$

since the asymptotic mean and covariance of $\mathrm{MED}(Z)$ does not depend of n_Z and n_{Z^c}. This construction is essentially based on the Theorem 2.1. Hence, a new functional nonparametric spatial scan statistic, denoted by $\Lambda_{\mathrm{MEDFSS}}$, is

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

and the most likely cluster (MLC), denoted by $\hat{\mathrm{C}}_{\mathrm{M}}$, is defined as

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

5.2.1.2 Significance of the MLC

Once the scan statistic $\Lambda_{\mathrm{MEDFSS}}$ and the MLC have been computed, its significance must be evaluated. Since computing the null distribution of $\Lambda_{\mathrm{MEDFSS}}$ is untractable, we decided to run the random labelling technique as described in subsection 3.2.3 of Chapter 3 and in subsection 4.2.4 of Chapter 4. First, we generate datasets by randomly associating the functional marks X_i to the spatial locations s_i, where $i \in \{1, \ldots, n\}$. Then, let T be the number of random permutations of the simulated datasets and $\Lambda^{(1)}_{\mathrm{MEDFSS}}, \ldots, \Lambda^{(T)}_{\mathrm{MEDFSS}}$ be the scan statistics associated with the simulated datasets. As mentionned in Dwass (1957), the p-value of the scan statistic $\Lambda_{\mathrm{MEDFSS}}$ is defined as

$$\hat{p}_{\mathrm{M}} = \frac{1 + \sum_{i=1}^{T} \mathbb{1}_{\{\Lambda^{(i)}_{\mathrm{MEDFSS}} > \Lambda_{\mathrm{MEDFSS}}\}}}{T+1}.$$

Finally, the most likely cluster $\hat{C}_M$ is said to be significant if $\hat{p}_{\mathrm{M}}$ is below the type I error.

5.2.1.3 Application to real data

We started by testing the significance of this statistic using the same Spain real data that we have used in Chapter 3 for extracting features in Spanish province population growth. The description and the source of this dataset are given in the Chapter 3, subsection 3.3.2. In order to detect spatial cluster for low or high demographic evolution, we have used the median spatial scan statistic $\Lambda_{\mathrm{MEDFSS}}$. Using $T = 999$ random permutations, the p-value obtained is highly significant ($\hat{p}_{\mathrm{M}} = 0.001$) and the MLC $\hat{\mathrm{C}}_{\mathrm{M}}$ is plotted in Figure 5.1.

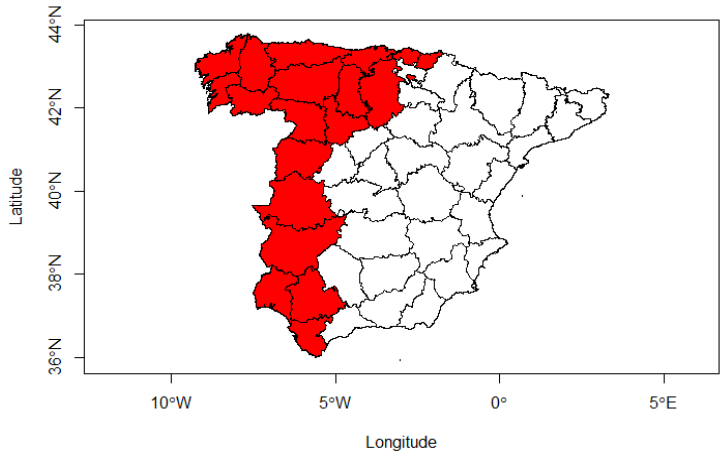

Figure 5.1: The MLC $\hat{C}_M$ detected by the median spatial scan statistic Λ_{MEDFSS}.

The MLC plotted in Figure 5.1 contains 19 provinces of the west of Spain. This cluster includes the same 13 provinces obtained using the Wilcoxon-Mann-Whitney scan statistic Λ_{WMWFSS} (see, Figure 3.4) and 6 more other provinces of the west of Spain. Moreover, similary to the MLC detected by the Wilcoxon-Mann-Whitney, this MLC includes provinces with the lowest demographic evolution compared to the rest of the studied area (see Figure 5.2 and Figure 3.5).

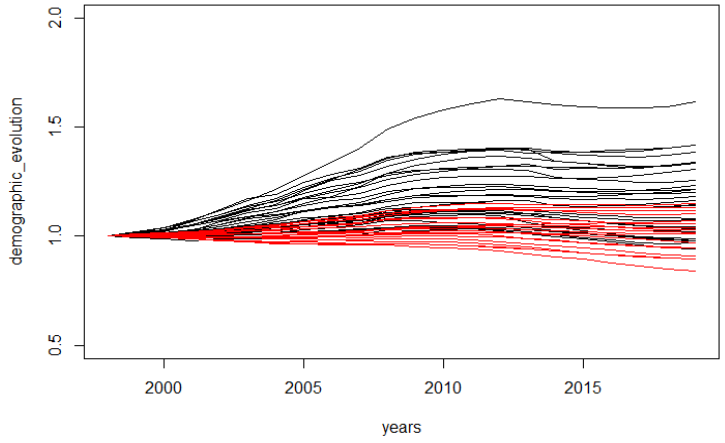

Figure 5.2: The demographic evolution curves (from 1998 to 2019) in each provinces are presented. Curves in red correspond to provinces inside $\hat{C}_M$, curves in black correspond to provinces outside $\hat{C}_M$.

We aim to compare this result also with the parametric spatial scan statistic, denoted by Λ_{CFFSS}, which is developed by Frévent et al. (2021a) based on the

test statistic introduced by Cuevas et al. (2004). A significant p-value is found ($\hat{p}_{\text{PFSS}} = 0.013$) using $T = 999$ permutations. The MLC detected and curves associated are shown in Figure 5.3.

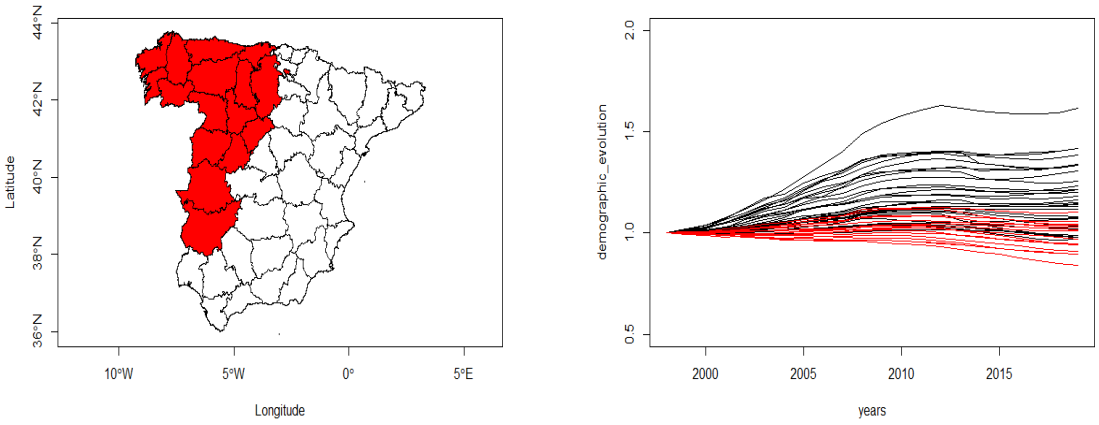

Figure 5.3: The most likely cluster detected by Λ_{CFFSS} and demographic evolution curves associated.

The MLC detected by the parametric scan statistic Λ_{CFFSS} includes 16 provinces. We remark that this cluster contains less provinces than the one detected by Λ_{MEDFSS}. However, the p-value associated in this parametric case is greater than the ones obtained by the nonparametric scan statistics Λ_{MEDFSS} and Λ_{WMWFSS}. Moreover, all nonparametric and parametric scan statistics Λ_{MEDFSS}, Λ_{WMWFSS}, Λ_{CFFSS} detect clusters of low demographic evolution of the Spanish population (see Figure 5.2, Figure 3.5 and Figure 5.3).

5.2.2 A new parametric spatial scan statistic for functional data

In the parametric case, we always use the likelihood ratio between the two models built under the hypotheses H_0 and H_1 to construct a concentration index leading to a scan statistic. Consider X a random element valued in χ. Therefore,

- When the space $\chi = \mathbb{R}$ (X is a real random variable), there is a relationship between the likelihood ratio used for the construction of the univariate spatial Gaussian scan statistic of Kulldorff et al. (2009) and the Student T^2 test statistic. This relation is proved in Chapter 1 and it is defined as equation (1.8).

- When the space $\chi = \mathbb{R}^d$ (X is a real random vector), there is a relationship between the likelihood ratio used for the construction of the multivariate

spatial Gaussian scan statistic of Cucala (2017) and the Hotelling T^2 test statistic (a generalization of the Student T^2 test statistic). This relation is proved and given in Anderson (2003).

- When $\chi = L^2([0,1],\mathbb{R})$ which is an Hilbert space, Horváth et al. (2013) proposed two test statistics for comparaison of two samples. These statistics are based on orthogonal projections of the difference between the sample mean functions. We have used these two statistics, denoted by HKR1 and HKR2, in subsection 2.3.1. As said in Chakraborty and Chaudhuri (2015), HKR1 is the same as the Hotelling T^2 test statistic based on finite number of such projections. Thus, we decided to construct a parametric spatial scan statistic using the test statistic HKR1.

5.2.2.1 Statistic construction

Consider in this case $\chi = L^2([0,1],\mathbb{R})$. Let $s_1,\ldots,s_n$ be n spatial locations in $D \subseteq \mathbb{R}^2$ and $X_1,\ldots,X_n$ are the independent functional marks associated. Similary to previous paragraph and the Chapter 3, we use $\mathcal{S}$ the set of circular potential clusters. Thus, let $Z \in \mathcal{S}$ be any potential cluster with size n_Z and Z^c its complement with size n_{Z^c}. Suppose that the marks in Z and Z^c are respectively from probability measures P_Z and P_{Z^c} on χ and these two measures differ by a shift Δ_Z. Thus, the parametric functional test statistic HKR1 of Horváth et al. (2013), for testing $H_0 : \Delta_Z = 0$ against $H_{1,Z} : \Delta_Z \neq 0$, is defined as

$$\mathrm{HKR}(Z) = \frac{n_Z n_{Z^c}}{n} \sum_{l=1}^{p} \frac{\langle \bar{X}_{n_Z} - \bar{X}_{n_{Z^c}}, \hat{\varphi}_l \rangle}{\hat{\lambda}_l},$$

where, $\langle .,. \rangle$ denotes the inner product in χ, the $\hat{\lambda}_l$'s are the eigenvalues of the usual empirical pooled covariance of the $\{X_i, i \in s_i\}$'s and the $\{X_j, j \in Z^c\}$'s in descending order of magnitude, the $\hat{\varphi}_l$'s are the corresponding empiricial eigenvectors and

$$\bar{X}_{n_Z} = \frac{1}{n_Z} \sum_{\{i, s_i \in Z\}} X_i \quad \text{and} \quad \bar{X}_{n_{Z^c}} = \frac{1}{n_{Z^c}} \sum_{\{j, s_j \in Z^c\}} X_j.$$

The number of projection directions p here is chosen using the cumulative variance method described by Horváth et al. (2013). Under H_0, Horváth et al. (2013) studied the limit of $\mathrm{HKR}(Z)$ and they proved that it converges in law to a random variable W from Chi-squared distribution with p degrees of freedom ($W \sim \chi^2(p)$). Thus, a new concentration index is given by

$$\mathrm{H}(Z) = F(\mathrm{HKR}(Z); p),$$

where $F(.;p)$ is the distribution function of the random variable $W \sim \chi^2(p)$. We remark that, for all $Z \in S$, $\mathrm{H}(Z)$ has a standard uniform distribution $\mathcal{U}(0,1)$.

Consequenty, using the concentration index $\mathrm{H}(Z)$, a new parametric scan statistic, labelled as Λ_{HFSSS}, is given by

$$\Lambda_{\mathrm{HFSSS}} = \max_{Z \in \mathcal{S}} \mathrm{H}(Z)$$

and the MLC associated, denoted by $\hat{\mathrm{C}}_{\mathrm{H}}$, is defined as

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

5.2.2.2 Significance of the MLC

Once the scan statistic Λ_{HFSSS} and the MLC have been computed, its significance must be evaluated. Similary to the paragraph 5.2.1.2, we simulate datasets by randomly associating the functional marks X_i to the spatial locations s_i, where $i \in \{1, \ldots, n\}$. Then, consider T be the number of random permutations of the simulated datasets and $\Lambda^{(1)}_{\mathrm{HFSSS}}, \ldots, \Lambda^{(T)}_{\mathrm{HFSSS}}$ be the scan statistics associated with the generated datasets. Thus, the p-value of the scan statistic Λ_{HFSSS}, is defined as

$$\hat{p}_{\mathrm{H}} = \frac{1 + \sum_{i=1}^{T} \mathbb{1}_{\{\Lambda^{(i)}_{\mathrm{HFSSS}} > \Lambda_{\mathrm{HFSSS}}\}}}{T+1}.$$

Hence, the most likely cluster $\hat{\mathrm{C}}_{\mathrm{H}}$ is said to be significant if $\hat{p}_{\mathrm{H}}$ is below the type I error.

5.2.2.3 Application to real data

Let's take the same real dataset used in the paragraph 5.2.1.3. To detect spatial cluster for low or high demographic evolution, we have used here the parametric functional spatial scan statistic Λ_{HFSSS}. Using $T = 999$ random permutations, the p-value obtained is highly significant ($\hat{p}_{\mathrm{H}} = 0.001$) and the MLC $\hat{\mathrm{C}}_{\mathrm{H}}$ is plotted in Figure 5.4.

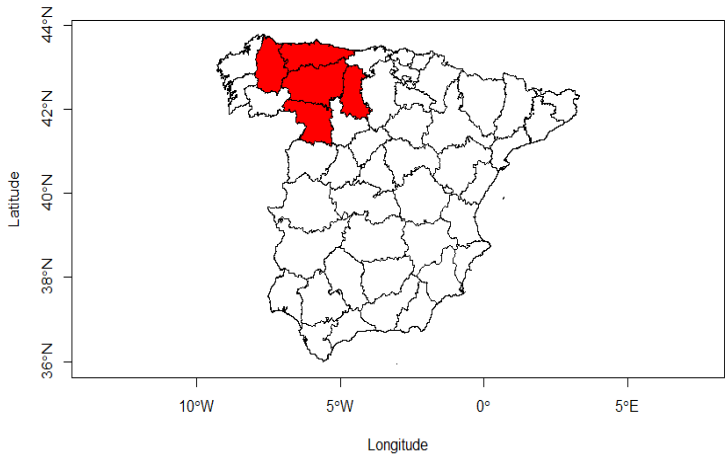

Figure 5.4: The MLC detected by the new parametric spatial scan statistic Λ_{HFSSS}.

The new parametric scan statistic Λ_{HFSSS} identified a significant spatial cluster which is located in the northwest of Spain. This MLC (see Figure 5.4) contains 5 Spanish provinces (*Lugo*, *Asturias*, *León*, *Zamora*, *Palencia*) which are also included in the MLC detected by the parametric scan statistic Λ_{CFFSS} introduced by Frévent et al. (2021a) and the nonparametric ones proposed in this manuscript Λ_{WMWFSS} (Smida et al. , 2021) and Λ_{MEDFSS} (subsection 5.2.1). All the MLC detected by the previous scan statistics are shown in Figure 5.4, Figure 5.3, Figure 3.4 and Figure 5.1. We notice that the characteristic of this new parametric method is that it detects smaller clusters which may be relevant using other applications. Moreover, this new parametric scan statistic Λ_{HFSSS} includes provinces with the lowest demographic evolution of the Spanish population (see, Figure 5.5).

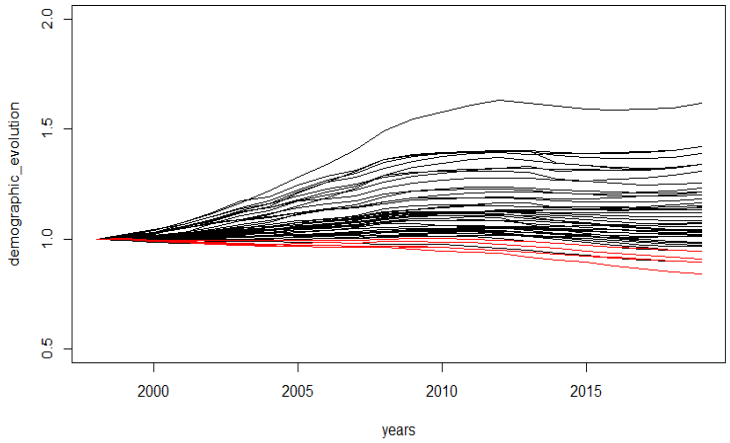

Figure 5.5: The demographic evolution curves (from 1998 to 2019) in each provinces are presented. Curves in red correspond to provinces inside $\hat{\mathrm{C}}_{\mathrm{H}}$, curves in black correspond to provinces outside $\hat{\mathrm{C}}_{\mathrm{H}}$.

Some of this work has already been started. Concerning the median test, some potential improuvements will be implemented:

- The p-values are obtained from random permutations procedure (see, subsection 2.2.3). We propose to determine these values using asymptotic distribution of the test statistic under the null hypothesis presented in Theorem 2.1.
- Asymptotic law of this test statistic under the alternative hypothesis has already been found, demonstrated and is in the process of being written.

Concerning the spatial scan statistic methods, simulation study of the new parametric and nonparametric scan statistics proposed in subsection 5.2.1 and subsection 5.2.2 are in preparation.

Bibliography

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.

Allévius, B. (2018). **scanstatistics**: Space-Time Anomaly Detection using Scan Statistics. *Journal of Open Source Software.* **3**, 515.

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

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

Aneiros, G., Cao, R., Fraiman, R., Genest, C. and Vieu, Ph. (2019). Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis.* **170**, 3–9.

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.

Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. (2007). Local Smoothing Regression with Functional Data. *Computational Statistics.* **22**, 353–369.

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

Berrendero, J.R., Justel, A. and Svarc, M. (2011). Principal Components for Multivariate Functional Data. *Computational Statistics & Data Analysis.* **55**, 2619–2634.

Bhatt, V. and Tiwari, N. (2014). A Spatial Scan Statistic for Survival Data Based on Weibull Distribution. *Statistics in medicine.* **33**, 1867–1876.

Blair, R. C. and Higgins, J. J. (1980). A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to that of Student'st Statistic Under Various Nonnormal Distributions.*Journal of Educational Statistics.* **5**, 309–335.

Boente, G. and Fraiman, R. (2000). Kernel-Based Functional Principal Components. *Statistics & Probability Letters.* **48**, 335–345.

Borovskikh, Y. V. (1996). *U-statistics in Banach spaces.* VSP International Science Publishers.

Borwein, J. M. and Vanderwerff, J. D. (2010). *Convex functions: Constructions, characterizations and counterexamples.* Cambridge University Press, Cambridge.

Capéraà, Ph. and Cutsem, B.V. (1988). *Méthodes et modèles en statistiques non paramétrique. Exposé fondamental.* Presses de l'université Laval.

Chakraborty, A. and Chaudhuri, P. (2014a). A Wilcoxon-Mann-Whitney type test for infinite dimensional data. *arXiv:1403.0201v1.*

Chakraborty, A. and Chaudhuri, P. (2014b). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics.* **42**, 1203–1231.

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

Chakraborty, B. and Chaudhuri, P. (1999). On affine invariant sign and rank tests in one and two-sample multivariate problems. *Multivariate analysis, design of experiments and survey sampling.* **159**, 499–522.

Chen, C. and Kim, AY., Ross, M. and Wakefield, J. (2018). **SpatialEpi**: Methods and Data for Spatial Epidemiology. *https://CRAN.R-project.org/package=SpatialEpi.*

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

Chiou, JM. and Müller, HG. (2007). Diagnostics for Functional Regression via Residual Processes. *Computational Statistics & Data Analysis.* **15**, 4849–4863.

Chong, S., Nelson, M., Byun, R., Harris, L., Eastwood, J. and Jalaludin, B. (2013). Geospatial Analyses to Identify Clusters of Adverse Antenatal Factors for Targeted Interventions. *Int J Health Geogr.* **12**, 46.

PROJ contributors. (2021). PROJ coordinate transformation software library. *Open Source Geospatial Foundation. https://proj.org/.*

Clifford, B. R. and Higgins, J.J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. *Journal of Educational Statistics.* **5**, 309–335.

Cressie, N. (1977). On Some Properties of the Scan Statistic on the Circle and the Line. *Journal of Applied Probability.* **14**, 272–283.

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. Demattei, C., Lopes, P. and Ribeiro, A. (2013). A Spatial Scan Statistic for Case Event Data Based on Connected Components. *Computational Statistics.* **28**, 357–369.

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

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

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference.* **147**, 1–23.

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

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear Functional Regression: The Case of Fixed Design and Functional Response. *The Canadian Journal of Statistics.* **30**, 285–300.

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, JP. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data. *Computational Statistics and Data Analysis,* **51**, 3931–3945.

Dematteï, C., Molinari, N. and Daurès, JP. (2006). **SPATCLUS**: An R Package for Arbitrarily Shaped Multiple Spatial Cluster Detection for Case Event Data. *Computer Methods and Programs in Biomedicine.* **84**, 42–49.

Donnan, P., Parratt, J., Wilson, S., Forbes, R., O'Riordan, J. and Swingler, R. (2005). Multiple Sclerosis in Tayside, Scotland: Detection of Clusters Using a Spatial Scan Statistic. *Multiple Sclerosis (Houndmills, Basingstoke, England)* **11**, 403–8.

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.

Duncan, D., Rienti, M., Kulldorff, M., Aldstadt, J., Castro, M., Frounfelker, R., Williams, J., Sorensen, G., Johnson, R., Hemenway, D. and Williams, D. (2011). Local Spatial Clustering in Youths' Use of Tobacco, Alcohol, and Marijuana in Boston. *The American Journal of Drug and Alcohol Abuse.* **42**, 412–421.

Durbeck, H., Greiling, D., Estberg, L., Long, A., Jacquez, G., Pallicaris, Y. and Hinton, S. (2012). ClusterSeer: Software for the Detection and Analysis of Event Clusters, User Manual Book 2, Version 2.5. *BioMedware.*

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

Estévez-Pérez, G. and Vieu, P. (2021). A new way for ranking functional data with applications in diagnostic test. *Computational Statistics.* **36**, 127–154.

Ferraty, F. and Vieu, P. (2002). Functional Nonparametric Model and Application to Sprectrometric Data. *Computational Statistics.* **17**, 545–564.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis (Theory and practice).* Springer-Verlag, New York.

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

Frévent, C., Ahmed, MS., and Dabo-Niang, S. and Genin, M. (2021b). Investigating Spatial Scan Statistics for Multivariate Functional Data. *arXiv:2103.14401.*

Frévent. C., Ahmed, MS., Marbac. M. and Genin. M. (2021a). Detecting spatial clusters on functional data: new scan statistic approaches. *arXiv:2011.03482v2.*

Frévent, C., Ahmed, MS., Soula, J., Smida, Z., Cucala, L., Dabo-Niang, S. and Genin, M. (2021c). 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.

Gao, J., Zhang, Z., Hu, Y., Bian, J., Jiang, W., Wang, X., Sun, L. and Jiang, Q. (2014). Geographical Distribution Patterns of Iodine in Drinking-Water and Its Associations with Geological Factors in Shandong Province, China. *International Journal of Environmental Research and Public Health.* **11**, 5431–5444.

Geenens, G. (2015). Moments, errors, asymptotic normality and large deviation principle in nonparametric functional regression. *Statistics & Probability Letters.* **107**, 369–377.

Gijbels, I. and Nagy, S. (2017). On a general definition of depth for functional data. *Statistical Science.* **32**, 630–639

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.*

Goia, A. and Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis.* **146**, 1–6

Górecki, T. and Smaga, Ł. (2015). A Comparison of Tests for the One-Way ANOVA Problem for Functional Data. *Computational Statistics.* **30**, 987–1010.

Górecki, T. and Smaga, Ł. (2017). Multivariate Analysis of Variance for Functional Data. *Journal of Applied Statistics.* **44**, 2172–2189.

Greiling, D., Estberg, L., Long, A. and Jacquez, G. (2012). ClusterSeer: Software for the Detection and Analysis of Event Clusters, User Manual Book 1, Version 2.5. *BioMedware.*

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.*

Hoffmann-Jørgensen, J. and Pisier, G. (1976). The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probability.* **4**, 587-599.

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. *The Annals of Mathematical Statistics.* **2**, 360–378.

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.

Jacques, J. and Preda, C. (2016). Model-Based Clustering for Multivariate Functional Data. *Computational Statistics & Data Analysis.* **71**, 92–106.

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

Jung, I., Kulldorff, M. and Klassen, AC. (2007). A Spatial Scan Statistic for Ordinal Data. *Statistics in Medicine.* **26**, 1594–1607.

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

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis.* CRC Press.

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *Ann. Statist.* **25**, 435–477.

Koul, H.L. and Staudte, R. G. (1972). Weak Convergence of Weighted Empirical Cumulatives Based on Ranks. *Ann. Math. Statist.* **43**, 832–841.

Kleinman, K. (2015). **rsatscan**: Tools, Classes, and Methods for Interfacing with SaTScan Stand-Alone Software. *https://CRAN.R-project.org/package=rsatscan.*

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

Kulldorff, M. (1999). Spatial Scan Statistics: Models, Calculations, and Applications. In: Scan Statistics and Applications. *Recent Advances on Scan Statistics and Applications. Birkhäuser, Boston, MA.*

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

Kulldorff, M. (2018). TreeScan User Guide. *https://www.treescan.org.*

Kulldorff, M. (2021). SaTScan User Guide for Version 9.7. *https://www.satscan.org/.*

Kulldorff, M., Athas, WF., Feurer, EJ., Miller, BA. and Key, CR. (1998). Evaluating Cluster Alarms: A Space-Time Scan Statistic and Brain Cancer in Los Alamos, New Mexico. *The American Journal of Public Health.* **88**, 1377–1380.

Kulldorff, M., Fang, Z. and Walsh, SJ. (2003). A Tree-Based Scan Statistic for Database Disease Surveillance. *Biometrics.* **59**, 323–331.

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., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in medicine.* **25**, 3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, WK., 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.

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

Lehmann, E. L. (1986). *Testing Statistical Hypotheses (Second edition).* Springer-Verlag, New York.

Lehmann, E. L and Romano, J.P. (2005). *Testing Statistical Hypotheses (Third edition).* Springer-Verlag, New York.

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

Lin, Z., Lopes, ME. and Müller, HG. (2021). High-dimensional MANOVA via Bootstrapping and its Application to Functional and Sparse Count Data. *arXiv:2007.01058.*

Ling, N. and Vieu, Ph. (2018). Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics.* **52**, 934–949.

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

Loche, R., Giron, B., Abrial, D., Cucala, L., CharrasGarrido, M. and De-Goer, J. (2016). **graphscan**: Cluster Detection with Hypothesis Free Scan Statistic. *https://CRAN.R-project.org/package=graphscan.*

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

Luquero, FJ., Banga, C., Remartínez, D., Palma, PP., Baron, E. and Grai, RF. (2011). Cholera Epidemic in Guinea-Bissau (2008): The Importance of "Place". *PloS One.* **6**.

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.

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.

Mood, A. M. (1950). *Introduction to the Theory of Statistics*. McGraw-Hi1l series in probability and statistics, New York.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *Ann. Math. Statist*. **25**, 514–522.

Moraga, P. (2017). **SpatialEpiApp**: A Shiny Web Application for the Analysis of Spatial and Spatio-Temporal Disease Data. *Spatial and Spatio-temporal Epidemiology.* **23**, 47–57.

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

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

Naus, J. I. (1965). Clustering of random points in two dimensions. *Biometrika.* **52**, 263–267.

Oja, H. (1999). Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: a Review. *Scandinavian Journal of Statistics*. **3**, 319–343.

Oja, H. (2010). *Multivariate Nonparametric Methods with R*. Springer, New York.

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

Otani, T. and Takahashi, K. (2021). **rflexscan**: The Flexible Spatial Scan Statistic. *https://CRAN.R-project.org/package=rflexscan.*

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis.* John Wiley & Sons, Inc, New York-London-Syndney.

Qiu, Z., Chen, J. and Zhang, JT. (2021). Two-Sample Tests for Multivariate Functional Data with Applications. *Computational Statistics & Data Analysis.* **157**.

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

Ramsay, JO., Graves, S. and Hooker, G. (2020). **fda**: Functional Data Analysis. *https://CRAN.R-project.org/package=fda.*

Serfling, R. (2002). A Depth Function and a Scale Curve Based on Spatial Quantiles. In *Statistical Data Analysis Based on the L1-Norm and Related Methods. Stat. Ind. Technol.Birkhäuser Basel*. 25–38.

Smida, Z., Cucala, L. Gannoun, A and Durif, G. (2021). A Wilcoxon-Mann-Whitney spatial scan statistic for functional data. *Computational Statistics and Data Analysis*. **167**, 107378.

Takahashi, K. and Tango, T. (2005). A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters. *International Journal of Health Geographics.* **4**, 11.

Takahashi, K. and Tango, T. (2006). An Extended Power of Cluster Detection Tests. *Statistics in Medicine.* **25**, 841–852.

Takahashi, K., Yokoyama, T. and Tango, T. (2010). FleXScan v 3.1: Software for the Flexible Scan Statistic.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Vardi, Y. and Zhang, C. -H. (2000). The multivariate L1-median and associated data depth. *Proceedings of the National Academy of Sciences.* **97**, 1423–1426.

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

Zapała, A. M. (2000). Jensen's Inequality for Conditional Expectations in Banach Spaces. *Real analysis exchange.* **26**, 541-552.

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.

Résumé

Dans cette thèse, nous nous focalisons d'un côté sur les tests statistiques de comparaison de deux échantillons basés sur les rangs et d'un autre côté sur la méthode de détection d'agrégats basée sur les statistiques de balayage spatial. Dans les deux cas, le travail a été effectué en utilisant des données fonctionnelles. L'objectif est d'étendre les méthodes développées dans le cadre univarié c'est-à-dire à destination des variables aléatoires à valeurs dans $\mathbb{R}$ au cadre fonctionnel c'est-à-dire en utilisant des variables aléatoires à valeurs dans un espace fonctionnel. Dans la première partie, nous étudions le test de la médiane basé sur les rangs dans le cadre univarié. Nous proposons ensuite une extension de ce dernier pour des données fonctionnelles. Puis, nous étudions le comportement asymptotique de sa statistique sous l'hypothèse nulle. Cette extension est comparée à d'autres statistiques paramétriques et non paramétriques existantes en utilisant des données simulées et des données réelles pour étudier sa puissance. Dans la deuxième partie, nous introduisons une statistique de balayage spatial non paramétrique pour des données fonctionnelles. Cette statistique est dérivée de celle de Wilcoxon-Mann-Whitney définie dans un espace de Hilbert. La méthode de balayage proposée est appliquée sur des données simulées pour évaluer sa performance, ensuite sur des données réelles pour extraire des caractéristiques de l'évolution démographique de la population espagnole. Dans la dernière partie, nous développons un package `R` intitulé `HDSpatialScan`. Il permet d'appliquer les statistiques de balayage spatial récemment développées pour des données fonctionnelles, y compris la statistique de balayage introduite dans cette thèse. Ce package facilite l'utilisation des méthodes de balayage et permet de visualiser les agrégats détectés d'une manière simple et rapide.

Mots-clefs : test non paramétrique de comparaison de deux échantillons, test de la médiane, statistique de balayage spatial, test de Wilcoxon-Mann-Whitney, données fonctionnelles, espace de Hilbert, package `R`.

Abstract

In this thesis, we focus on the statistical comparison tests of two samples based on ranks on the one hand and the cluster detection method based on spatial scan statistics on the other hand. In both cases, the work was performed using functional data. The goal is to extend the methods developed in the univariate case, i.e. for $\mathbb{R}$-valued random variables to the functional case, i.e. by using random variables valued in functional space. In the first part of this thesis, we study the median test based on ranks in the univariate case. We propose an extension of this latter for functional data. Then, we study the asymptotic behavior of its statistic under the null hypothesis. This extension is compared to other existing parametric and nonparametric statistics using simulated and real data to study its performance. In the second part, we introduce a nonparametric spatial scan statistic for functional data. It is derived from the Wilcoxon-Mann-Whitney statistic defined in an Hilbert space. The proposed scan method is applied on simulated data to study its performance, then on real data to extract characteristics of the demographic evolution of the Spanish population. In the last part, we develop an `R` package called `HDSpatialScan`. It allows to apply spatial scan statistics developed recently for functional data including the nonparametric scan statistic that we developed in this thesis. This package facilitates the use of the scanning methods and allows to plot the detected clusters with a simple and fast manner.

Keywords : nonparametric test of comparison of two samples, median test, spatial scan statistic, Wilcoxon-Mann-Whitney test, functional data, Hilbert space, `R` package.