Université de Montpellier

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Titre :

Détection d'agrégats temporels, spatiaux et spatio-temporels: contributions aux méthodes de balayage.

Spécialité : **Mathématiques appliquées**

préparé au laboratoire **IMAG**

dans le cadre de l'**École Doctorale I2S**

présenté et soutenu publiquement le **8 Juin 2022** par

Lionel CUCALA

devant le jury suivant:

Liliane BEL	**Rapporteure**
Jean-Noël BACRO	**Examinateur**
Édith GABRIEL	**Examinatrice**
Ali GANNOUN	**Examinateur**
Cécile HARDOUIN	**Examinatrice**
Cristian PREDA	**Rapporteur**
Radu-Stefan STOICA	**Rapporteur**
Christine THOMAS-AGNAN	**Examinatrice**

Table des matières

Introduction

Ce document est la synthèse de mes travaux dans le domaine de la détection d'agrégats sur données ponctuelles menés depuis ma soutenance de thèse, en 2006, jusqu'à maintenant. Les publications personnelles auxquelles je me réfère sont notées entre crochets et sont listées en fin de document, avant la bibliographie.

En Introduction, je détaille le type de données (semis de points), la problématique (détection d'agrégats) et les méthodes (statistiques de balayage) sur lesquelles je me suis concentré. Les Chapitres 1 à 3 décrivent les différentes voies d'innovation suivies. Le Chapitre 1 se concentre sur les différentes façons de construire une famille d'agrégats potentiels sans contrainte de forme, en se basant sur les distances entre événements. Le Chapitre 2 explique comment les indices de concentration classiques, basés sur des rapports de vraisemblance, peuvent être avantageusement remplacés par des indices de concentration non-paramétriques. Le Chapitre 3 s'intéresse à l'extension de la détection d'agrégats à des données multivariées et fonctionnelles. Pour conclure, je donne quelques pistes pour des recherches futures.

1 Processus ponctuels

Dans de nombreuses applications, le jeu de données consiste en une collection d'événements localisés aléatoirement sur un domaine d'observation. Cela peut être l'ensemble des emplacements d'une espèce végétale dans une forêt, les temps d'arrivée de clients dans une boutique, les adresses postales de personnes atteintes par une certaine maladie, les coordonnées de galaxies dans l'espace... Ce genre de données est généralement vu comme la réalisation d'un processus ponctuel observé dans une fenêtre : le nombre N d'événements dans cette fenêtre, ainsi que leurs localisations $S_1, \cdots, S_N$, sont aléatoires (Cressie 1993).

Soit $\{s_i : i \in [\![1, n]\!]\}$ la réalisation d'un processus ponctuel, $s_i \in W$ étant la localisation du $i^{\text{ème}}$ événement et $W \subset \mathbb{R}^d$ la fenêtre d'observation. Les localisations peuvent être temporelles ($d = 1$) ou spatiales ($d = 2$ ou $d = 3$), voire spatio-temporelles ($d = 2+1$ ou $d = 3+1$). Il existe de très nombreux modèles probabilistes adaptés à ces processus, ainsi que des méthodes statistiques pour inférer les paramètres associés (Møller et Waagepetersen 2003). Le modèle de référence, dit de hasard spatial complet ("complete spatial randomness" ou CSR), considère que les localisations sont indépendantes et distribuées uniformément dans W.

2 Détection d'agrégat(s)

La notion originelle d'agrégat ("cluster" en anglais) a été introduite par Naus 1963 durant sa thèse et peut s'expliquer très simplement : un agrégat est un ensemble d'événements géographiquement proches. Il sera d'autant plus significatif que la probabilité de l'observer sous l'hypothèse CSR est faible. Les travaux originaux de Naus consistent donc à calculer la probabilité que, si n événements sont indépendemment et uniformément distribués sur l'intervalle $[0, 1]$ (resp. le carré $[0, 1]^2$), on en trouve k dans un intervalle de longueur l (resp. dans un rectangle de volume v). Si ces questions sont très stimulantes d'un point de vue mathématique (Glaz, Naus et Wallenstein 2001), elles sont néanmoins inadaptées à certaines problématiques que l'on rencontre de manière pratique. Ainsi, cette notion d'agrégat a été légèrement transformée pour mieux coller aux besoins de l'analyse de données. Elle sera différente suivant que le processus ponctuel soit marqué ou non.

2.1 Processus non marqué

L'objectif est de comparer l'échantillon de localisations $\{s_i : i \in [\![1, n]\!]\}$ à une population sous-jacente dont la distribution est décrite par la mesure de probabilité $\mu(\cdot)$, qui sera par défaut la mesure uniforme sur W. Ce que nous appelons agrégat probable est alors un sous-ensemble $Z \subset W$ dans lequel le nombre de localisations s_i est anormal : soit trop élevé, soit trop faible par rapport à $\mu(Z)$, la proportion de la population de W contenue dans Z. Une méthode de détection d'agrégat doit, dans un premier temps, identifier l'agrégat le plus probable puis évaluer sa significativité par rapport à la mesure de référence $\mu(\cdot)$.

2.2 Processus marqué

Notons X une variable aléatoire dont la nature peut être quelconque : binaire, discrète, continue, multivariée, fonctionnelle... On dit que le processus ponctuel générateur de l'échantillon de localisations $\{s_i : i \in [\![1, n]\!]\}$ est marqué par la variable X s'il existe une réalisation x_i de la variable X en chaque localisation s_i. Le jeu de données que nous observons est alors $\{(s_i, x_i) : i \in [\![1, n]\!]\}$ et ce que nous appelons agrégat probable est dans ce cas un sous-ensemble $Z \subset W$ dans lequel se concentrent des observations de X atypiques par rapport à celles dans Z^c, le complémentaire de Z dans W. Une méthode de détection d'agrégat doit, dans un premier temps, identifier l'agrégat le plus probable puis évaluer sa significativité par rapport à une hypothèse nulle d'homogénéité de X dans W.

3 Les statistiques de balayage

Certaines méthodes de détection d'agrégats s'appuient sur une modélisation complète du jeu de données observé et nécessitent l'estimation de nombreux paramètres, liés notamment à la dépendance entre événements, pour parvenir à identifier l'agrégat le plus probable (Lawson et Denison 2002). Les statistiques de balayage ont, elles, l'avantage de ne pas avoir à spécifier ces relations de dépendance entre événements.

3.1 Définition

La première statistique de balayage ("scan statistic" en anglais) (Naus 1963) consistait simplement en le nombre maximum d'événements contenus dans une fenêtre de taille et de volume fixés. Cette fenêtre est balayée de manière continue sur l'ensemble du domaine d'observation S et identifie la zone où la concentration en événements est la plus importante. La distribution de cette statistique a été étudiée abondamment (Glaz, Naus et Wallenstein 2001). Cependant, elle souffre de deux inconvénients majeurs, qui limitent son utilisation pratique : elle nécessite de fixer le volume de la fenêtre de balayage a priori et elle ne s'adapte pas à une distribution de population non-uniforme.

Dans les années 1990, Nagarwalla 1996, pour des processus ponctuels temporels, et Kulldorff 1997, pour des processus ponctuels spatiaux, ont introduit des indices de concentration basés sur des rapports de vraisemblance généralisés afin de pouvoir comparer des zones de volumes différents. Je reviendrai en détail sur ces indices de concentration dans le Chapitre 2. Soit $Z \subset W$, notons $I(Z)$ l'indice de concentration associé à la zone Z, qui va croître lorsque la concentration en événements augmente dans Z (processus non-marqué) ou lorsque la concentration en marques anormales augmente dans Z (processus marqué). Nous allons chercher à maximiser cet indice de concentration sur un ensemble de zones appelées agrégats potentiels. Cet ensemble est noté $\mathcal{C}$ et je reviendrai en détail sur sa composition dans le Chapitre 1. La statistique de balayage est définie de la manière suivante :

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

et l'agrégat potentiel sur lequel l'indice de concentration est maximisé,

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

est noté agrégat le plus probable. Notons que rien n'interdit de s'intéresser aux agrégats secondaires : ceux pour lesquels l'indice de concentration est le deuxième plus grand, troisième plus grand... (Zhang, Kulldorff et Assunção 2010)

3.2 Significativité

La dernière partie de la méthode de balayage consiste maintenant à évaluer la significativité de la valeur de λ observée : quelle est la probabilité, sous l'hypothèse nulle, d'obtenir une concentration maximale aussi élevée ?

Même si la distribution des événements (et de leurs marques) est très simple, les calculs de probabilité concernant λ sont compliqués par la dépendance entre tous les agrégats potentiels $Z \in \mathcal{C}$. Certains résultats théoriques ont été obtenus (Glaz, Naus et Wallenstein 2001) mais uniquement lorsque les agrégats potentiels sont tous de même taille.

Pour évaluer la significativité de λ, il est donc nécessaire de se tourner vers des méthodes de simulation de type Monte-Carlo. Notons $\lambda^{(1)}, \cdots, \lambda^{(T)}$ les valeurs des statistiques de balayage associées à T échantillons simulés sous l'hypothèse nulle. D'après Dwass 1957, la p-valeur de la statistique observée λ est $\dfrac{R}{T+1}$, où R est le rang de λ sur l'ensemble des $T+1$ statistiques de balayage calculées

$(\lambda^{(1)}, \cdots, \lambda^{(T)}, \lambda)$. Notons que cette p-valeur est sans biais dans le sens où, sous l'hypothèse nulle, la probabilité d'observer une p-valeur inférieure ou égale à p vaut exactement p. Suivant la théorie des tests statistiques, l'agrégat $\hat{C}$ sera significatif si la p-valeur associée est inférieure au niveau de test α.

La clé du problème consiste donc à pouvoir simuler des semis de points sous l'hypothèse nulle. Dans le cas de processus non-marqués, il faut choisir le nombre d'événements N ainsi que leurs positions $S_1, \cdots, S_N$. Concernant le nombre d'événements, deux choix sont possibles : soit l'on fixe ce nombre à n, le nombre d'événements du semis de points initial ; soit l'on considère que le semis de points doit être une réalisation d'un processus ponctuel de Poisson, et alors N doit suivre une loi de Poisson dont le paramètre sera égal à l'estimateur du maximum de vraisemblance, n. Ensuite, le nombre d'événements étant fixé, les positions peuvent être simulées suivant la mesure de référence $\mu(\cdot)$ par une méthode d'acceptation-rejet (Kulldorff 1997).

Dans le cas de processus marqués, il est généralement admis que seules les marques, et non les positions des événements, ont à être modifiées dans les semis de points simulés. Le nombre n d'événements et les positions $s_1, \cdots, s_n$ du semis de points initial sont donc conservés. La seule question à résoudre est donc la suivante : comment simuler des marques $X_1, \cdots, X_n$ sous l'hypothèse nulle d'homogénéité dans W ? Deux solutions sont proposées dans la littérature. La première consiste à spécifier totalement la distribution de X et à en simuler de nouvelles réalisations (Jung, Kulldorff et Klassen 2007). La seconde s'affranchit du besoin de modéliser X : les marques observées sur le semis de points initial $x_1, \cdots, x_n$ sont réutilisées mais après avoir été permutées aléatoirement (Kulldorff, Heffernan et al. 2005). Si on note $\{(s_i^{(t)}, x_i^{(t)}) : i \in [\![1, n]\!]\}$ le $t^{\text{ème}}$ semis de points simulé, on a

$$(x_1^{(t)}, \cdots, x_n^{(t)}) = (x_{\sigma(1)}, \cdots, x_{\sigma(n)})$$

où $\sigma(.)$ est une permutation aléatoire sur les entiers de 1 à n. Par cette permutation, la structure spatiale de la distribution des marques est fracturée et le semis de points simulé respecte bien l'hypothèse d'homogénéité.

Chapitre 1

Des agrégats potentiels originaux

Publications : [2], [3], [4].

Dans ce chapitre, je décris en détail les différentes façons de construire une famille d'agrégats potentiels. Néanmoins, dans un premier temps, il me semble nécessaire d'expliquer comment des données groupées peuvent également être concernées par cette problématique. Alors que de nombreuses méthodes fixent a priori une forme et parcourent l'ensemble des zones respectant cette forme contrainte, certaines cherchent à construire des agrégats potentiels uniquement à partir des localisations des événements, sans aucune restriction sur leur géométrie. Parmi ces dernières, on peut distinguer celles qui cherchent à parcourir l'ensemble des événements dans un ordre pertinent, comme dans [2], de celles qui font émerger des agrégats potentiels en se basant uniquement sur les distances entre événements, comme dans [4]. Une difficulté supplémentaire apparaît lorsque le processus ponctuel observé est de type spatio-temporel : dans ce cas-là, les méthodes sans restriction de forme nécessitent la définition d'une distance spatio-temporelle entre événements, comme définie dans [3].

1 Données ponctuelles versus données groupées

Dans l'introduction de ce mémoire, j'ai spécifié que les données auxquelles je m'intéressais étaient de type ponctuel : à chaque événement est associé une localisation (temporelle, spatiale ou spatio-temporelle) précise $s_i \in W$ où $W \subset \mathbb{R}^d$ est la fenêtre d'observation. Malheureusement, dans de nombreux cas de figure, cette localisation précise est inconnue ou inaccessible (pour des raisons de confidentialité par exemple) et l'on doit se contenter, pour chaque événement, de l'intervalle de temps (jour, semaine, mois ...) et/ou de l'unité administrative (commune, canton, département...) à laquelle il est rattaché : on parle alors de données groupées (ou données de comptage) car de nombreux événements se retrouvent associés à la même localisation alors que ce n'est pas le cas en réalité. Soit $(W_1, \cdots, W_K)$ une partition de la fenêtre d'observation W en K unités administratives :

$$W = W_1 \cup \cdots \cup W_K \text{ et } \forall k \neq l,\ W_k \cap W_l = \emptyset.$$

On notera $u_i \in \{1, \cdots, K\}$ l'indice de l'intervalle de temps ou de l'unité administrative à laquelle est rattaché le $i^{\text{ème}}$ événement :

$$s_i \in W_k \Leftrightarrow u_i = k.$$

Lorsque l'on cherche à détecter des agrégats dans des données groupées, deux possibilités se présentent. La première consiste à utiliser une information sur le voisinage entre les éléments de la partition $(W_1, \cdots, W_K)$. Cette information peut être contenue dans une matrice de voisinage $M : M_{k,l}$ vaudra 1 si les sous-ensembles W_k et W_l sont voisins (au sens temporel ou géographique) et 0 sinon. L'objectif est alors de considérer comme agrégats potentiels tous les ensembles d'unités voisines. Pour expliciter cela mathématiquement, utilisons la théorie des graphes. Considérons le graphe $G = (V, E)$ dont l'ensemble des sommets est $V = \{1, \cdots, K\}$ et l'ensemble des arêtes $E = \{(k, l) : M_{k,l} = 1\}$. Les agrégats potentiels peuvent alors être identifiés à tous les sous-graphes connexes de G (Speakman, McFowland et Neill 2015).

La seconde possibilité consiste à transformer ces données groupées en données ponctuelles. Pour cela, à chaque unité d'observation W_k est associée une localisation centrale $c_k \in W_k$. Cela peut être, par exemple, le milieu de l'intervalle si W_k est un intervalle de temps, ou la position géographique de la préfecture si W_k est un département. Ainsi, au jeu de données initial $(u_1, \cdots, u_N)$, recensant uniquement les unités d'observation de n événements, se substitue le semis de points $x = (c_{u_1}, \cdots, c_{u_N})$.

Dans ce chapitre, nous nous concentrerons sur les méthodes de constitution des agrégats potentiels issus de données ponctuelles, tout en ayant à l'esprit qu'elles peuvent s'étendre à des données groupées par la simple transformation décrite ci-dessus.

2 Des agrégats potentiels à géométrie contrainte

A partir de maintenant, nous allons supposer que les localisations des événements sont toutes différentes :

$$\forall i \neq j,\ s_i \neq s_j.$$

On parle de processus ponctuel simple (Cressie 1993) lorsque la probabilité que deux événements aient la même localisation est nulle. S'il y a des répétitions dans l'ensemble des localisations $(s_1, \cdots, s_n)$, il suffira de les éliminer avant de construire l'ensemble des agrégats potentiels $\mathcal{C}$. Bien sûr, ces répétitions seront prises en compte lors du calcul des indices de concentration associés à l'ensemble des agrégats potentiels $\{I(Z) : Z \in \mathcal{C}\}$ que nous détaillerons dans le Chapitre 2.

Comment définir un agrégat potentiel ? Il s'agit bien sûr d'un sous-ensemble $Z \subset W$ de la fenêtre d'observation. Mais n'importe quel sous-ensemble de W ne peut pas être considéré comme un agrégat potentiel. Un critère essentiel pour l'interprétabilité d'un agrégat est sa convexité. En effet, on cherche à identifier un phénomène, dans le temps et/ou dans l'espace, qui modifie les observations. Ce phénomène se produit de manière continue et son effet doit donc se ressentir sur un ensemble convexe. Un autre point essentiel est que, entre deux agrégats potentiels Z et Z' contenant exactement les mêmes événements et inclus l'un dans l'autre, le plus probable est le plus petit. En effet, lorsqu'on traite de processus non marqués, on a

$$\begin{cases} \{i : s_i \in Z\} = \{i : s_i \in Z'\} \\ Z \subset Z' \end{cases} \Rightarrow I(Z) \geq I(Z').$$

Dans le cas où on travaille avec des processus marqués, la valeur de l'indice de concentration $I(Z)$ ne dépend même plus de la géométrie de Z mais uniquement des

marques associées aux événements dans Z donc

$$\{i : s_i \in Z\} = \{i : s_i \in Z'\} \Rightarrow I(Z) = I(Z').$$

Pour ces raisons, un choix logique pour l'ensemble des agrégats potentiels serait l'ensemble des enveloppes convexes de n'importe quel sous-ensemble de $S = (s_1, \cdots, s_n)$:

$$\mathcal{C} = \{\text{Conv}(A) : A \in \mathcal{P}(S)\}$$

où $\mathcal{P}(S)$ désigne l'ensemble des parties de S et $\text{Conv}(A)$ l'enveloppe convexe de A. En effet, on voit aisément que, quel que soit l'agrégat potentiel Z convexe, on a

$$\text{Conv}(\{s_i \in Z\}) \subset Z \Rightarrow I(Z) \leq I\big(\text{Conv}(\{s_i \in Z\})\big).$$

Malheureusement, en pratique, ce choix est généralement impossible car il nécessite de tester un nombre d'agrégats potentiels de l'ordre de 2^n, ce qui est beaucoup trop grand même pour des valeurs de n raisonnables.

Un choix classique pour réduire cet ensemble d'agrégats potentiels est de leur fixer des contraintes de forme. Bien sûr, ces contraintes seront différentes suivant que les localisations $s_1, \cdots, s_n$ soient temporelles, spatiales ou spatio-temporelles.

Il est à noter également que certains auteurs considèrent qu'un sous-ensemble $Z \subset W$ ne peut être considéré comme un agrégat si, dans le cas d'un processus non-marqué, il contient plus de la moitié de la population sous-jacente $\big($i.e. $\mu(Z) > \mu(Z^c)\big)$ ou, dans le cas d'un processus marqué, il contient plus de la moitié des localisations $s_1, \cdots, s_n$. Leur argument est que la détection d'agrégat consiste à exhiber un comportement minoritaire sur l'ensemble des observations. Pour ma part, je préfère généralement ne pas rajouter cette contrainte sur l'ensemble des agrégats potentiels et laisser l'indice de concentration faire le tri entre eux.

2.1 Localisations temporelles

Lorsque les localisations $s_1, \cdots, s_n$ sont temporelles, donc unidimensionnelles, le choix des agrégats potentiels est peu problématique puisqu'il existe un ordonnancement naturel entre les événements. Il suffit donc de considérer tous les intervalles de temps débutant par un événement et se terminant par un autre,

$$\mathcal{C} = \{[s_i, s_j] : s_i < s_j\},$$

puisque seuls ces intervalles pourront maximiser l'indice de concentration. Le cardinal de cet ensemble étant de l'ordre de n^2, cela ne pose généralement pas de problème calculatoire.

2.2 Localisations spatiales

Pour des localisations spatiales, le choix fréquemment rencontré, initié par Kulldorff 1997, est de se concentrer sur des agrégats potentiels circulaires. En effet, le disque semble la forme la plus naturelle pour modéliser un phénomène se produisant en un point et se diffusant tout autour. On peut penser par exemple à une source de pollution qui se répand ensuite dans le voisinage.

Il n'est bien sûr pas question de considérer tous les disques contenus dans W puisque l'objectif est de réduire le nombre d'agrégats potentiels. Le choix classique consiste à considérer tous les disques centrés en un événement et dont la frontière passe par un autre :

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

où $D_{i,j}$ est le disque centré en s_i et de rayon $d(s_i, s_j)$ et $d(\cdot, \cdot)$ désigne la distance euclidienne. Là encore, on retrouve un ensemble d'agrégats potentiels dont le cardinal est de l'ordre de n^2. De plus, il est très facile, pour chacun de ces disques $D_{i,j}$, de déterminer l'ensemble des localisations qui sont à l'intérieur.

Toutefois, il existe quelques cas de figure où la forme circulaire de ces agrégats potentiels n'est pas adaptée. On peut penser notamment à des phénomènes se propageant le long de cours d'eau ou selon des vents dominants. Pour pallier à cet inconvénient, Kulldorff, Huang, Pickle et al. 2006 ont proposé d'utiliser des agrégats potentiels de forme elliptique. Il est nécessaire pour l'utilisateur d'indiquer les différentes valeurs possibles pour l'orientation des axes et l'excentricité de ces ellipses. Bien sûr, le nombre d'agrégats potentiels est ainsi grandement multiplié par rapport au cas circulaire et ce choix ne devra être fait que si l'utilisateur croit fortement à son utilité.

2.3 Localisations spatio-temporelles

Très vite après l'introduction d'une statistique de balayage spatial à fenêtre variable ont été introduites des versions spatio-temporelles : Kulldorff, Athas et al. 1998 définissent des agrégats potentiels de type cylindrique dont la base, circulaire, correspond à la zone géographique, et la hauteur correspond à l'intervalle de temps considéré. Notons $s_i = (y_i, t_i)$ où $y_i \in \mathbb{R}^{d-1}$ représente la localisation géographique et $t_i \in \mathbb{R}$ la localisation temporelle. L'ensemble des agrégats potentiels est alors :

$$\mathcal{C} = \{C_{i,j} : 1 \leq i, j \leq n\}$$

où $C_{i,j}$ désigne le cylindre dont la base géographique est le disque centré en y_i et de rayon $d(y_i, y_j)$ et la hauteur l'intervalle de temps $[\min(t_i, t_j), \max(t_i, t_j)]$. Encore une fois, on retrouve un ensemble d'agrégats potentiels dont le cardinal est de l'ordre de n^2 et dont il est aisé de déterminer quel est l'ensemble des localisations qui sont à l'intérieur.

On peut bien sûr imaginer, si l'on souhaite prendre en compte des zones géographiques non circulaires, des agrégats potentiels dont la base soit une ellipse, telle que définie par Kulldorff, Huang, Pickle et al. 2006, et la hauteur l'intervalle de temps $[\min(t_i, t_j), \max(t_i, t_j)]$. Le prix à payer en termes de temps de calcul est similaire à celui observé dans le cas uniquement spatial.

3 Des agrégats potentiels sans forme contrainte

Dans de nombreux cas de figure, l'utilisation d'agrégats potentiels de forme contrainte ne pose pas de problème et n'empêche pas d'exhiber un ou des agrégat(s) significatif(s). Néanmoins, il peut parfois être pertinent de s'appuyer uniquement sur les données afin d'obtenir des agrégats potentiels de formes irrégulières (Duczmal,

Kulldorff et Huang 2006). De nombreuses méthodes (Duczmal, Cançado et al. 2007) permettent de le faire mais uniquement lorsque l'on dispose de données groupées. En effet, ces méthodes s'appuient sur la relation de voisinage entre unités géographiques.

Concernant les données ponctuelles, il existe très peu de travaux introduisant des agrégats potentiels de forme irrégulière. On peut noter que Duczmal, Moreira et al. 2011 proposent une astuce pour pouvoir utiliser les méthodes pour données agrégées sans avoir à introduire de découpage arbitraire. En effet, il est possible de s'appuyer sur le découpage du domaine d'observation suivant les cellules de Voronoï (Allard et Fraley 1997) $V_1, \cdots, V_n$ induites par les localisations $s_1, \cdots, s_n$. On rappelle que la cellule V_i contient tous les points de la fenêtre W qui sont plus près de la localisation s_i que de n'importe quelle autre :

$$V_i = \{s \in W : \forall j \in [\![1, n]\!],\, d(s, s_i) \leq d(s, s_j)\}.$$

3.1 Parcourir les événements de façon pertinente

Dans les années 2000, Dematteï, Molinari et Daurès 2007 proposent une procédure originale pour introduire un ordonnancement entre des localisations spatiales et obtenir un processus ponctuel unidimensionnel, sur lequel la détection d'agrégats est facilitée. Deux ans plus tard, je reprends dans [2] le même ordonnancement, mais en modifiant le processus ponctuel unidimensionnel obtenu afin qu'il satisfasse certaines propriétés mathématiques.

L'idée de Dematteï, Molinari et Daurès 2007 est de parcourir les localisations $s_1, \cdots, s_n$ de proche en proche. De manière arbitraire, on partira de la localisation la plus proche de la frontière du domaine d'observation, que l'on notera $s_{(1)}$:

$$s_{(1)} = \underset{s \in \{s_1, \cdots, s_n\}}{\arg\min}\ d(s, \partial W)$$

où ∂W désigne la frontière de la fenêtre d'observation W. Ensuite, de manière itérative, on cherchera, parmi les localisations non encore parcourues, laquelle est la plus proche de $s_{(i)}$:

$$\forall i \in [\![2, n]\!], \quad s_{(i)} = \underset{s \in \{s_1, \cdots, s_n\}, s \notin \{s_{(1)}, \cdots s_{(i-1)}\}}{\arg\min} d(s, s_{(i)}).$$

Ainsi, les localisations géographiques initiales sont toutes parcourues dans un ordre tel que des localisations appartenant à un même agrégat significatif ont de fortes chances d'être visitées les unes après les autres. Il est néanmoins possible que ce parcours "s'échappe" d'un agrégat avant d'avoir visité toutes ses localisations et y revienne un peu plus tard. On verra par la suite que ce problème n'est pas irrémédiable. Une solution envisageable serait alors de s'appuyer sur cet ordonnancement $\{s_{(1)}, \cdots, s_{(n)}\}$ pour proposer des agrégats potentiels spatiaux. Un choix dans ce sens serait :

$$\mathcal{C} = \{\text{Conv}(s_{(i)}, \cdots, s_{(j)}) : 1 \leq i \leq j \leq n\},$$

l'ensemble des enveloppes convexes de tous les groupes de localisations consécutives. Par cet ordonnancement, on construit des agrégats potentiels pertinents, de forme non contrainte et dont le cardinal est de l'ordre de n^2.

Néanmoins, en se limitant à cela, on n'exploite pas totalement l'information apportée par l'ordonnancement engendré. En effet, en plus de la succession des événements, nous disposons d'éléments concernant la proximité entre eux.

L'idée originelle de Dematteï, Molinari et Daurès 2007 était de s'appuyer sur les distances euclidiennes entre événements successifs, $d_1, \cdots, d_n$, où

$$d_1 = d(\partial W, s_{(1)}) \text{ et } \forall j \in [\![2, n]\!],\ d_j = d(s_{(j-1)}, s_{(j)}).$$

Cela semble naturel puisque deux événements consécutifs auront plus de chance d'appartenir à un agrégat significatif si la distance entre les deux est faible. Cependant, même sous l'hypothèse CSR, ces distances ne sont pas identiquement distribuées et leur structure de dépendence est très complexe. Pour ces raisons, nous proposons une approche différente basée sur une propriété probabiliste.

Une propriété de distribution

A partir de maintenant, nous noterons $S_{(i)}$ et D_i les variables aléatoires associées respectivement aux localisations ordonnées $s_{(i)}$ et aux distances entre elles d_i. Notons $E_1, \cdots, E_{n+1}$ des espacements uniformes, c'est-à-dire des espacements issus d'un n-échantillon de la loi uniforme sur $[0, 1]$ (Pyke 1965). Les espacements sont juste les intervalles entre valeurs consécutives. La distribution de $\{E_1, \cdots, E_{n+1}\}$ est classique : il s'agit de la loi de Dirichlet. On peut montrer que l'ordonnancement des événements spatiaux que nous venons de décrire précédemment donne naissance à des variables aléatoires ayant, sous l'hypothèse CSR, la même loi que ces espacements uniformes.

Soit $A_1 = \{s \in W : d(s, \partial W) < D_1\}$ le sous-ensemble contenant tous les points de W qui sont plus proches de la frontière que l'événement $s_{(1)}$. On peut alors voir $\mu(A_1)$ comme la proportion de la population qu'il a fallu explorer avant de rencontrer le premier événement, en partant de la frontière de W. Or, sous H_0, la distribution de $\mu(A_1)$ est la même que celle de E_1 et ne dépend pas de la géométrie de la fenêtre W. En effet, pour tout $r > 0$, on a

$$\begin{aligned}
\mathbb{P}\big(\mu(A_1) \leq \mu(A_{1,r})\big) &= 1 - \mathbb{P}\big(\mu(A_1) > \mu(A_{1,r})\big) \\
&= 1 - \mathbb{P}(\forall i \in [\![1, n]\!],\ S_i \in A_{1,r}^c) \\
&= 1 - \big(\mu(A_{1,r}^c)\big)^n = 1 - \big(1 - \mu(A_{1,r})\big)^n,
\end{aligned}$$

où $A_{1,r} = \{s \in W : d(s, \partial W) \leq r\}$ est le sous-ensemble de points de W dont la distance à la frontière est inférieure à r. La seconde égalité provient de la definition de A_1 : $A_{1,r}$ est inclus dans A_1 si et seulement si la distance de tous les événements $S_1, \cdots, S_n$ à la frontière ∂W est supérieure à r. La troisième égalité vient de la loi des S_i sous H_0 : les localisations sont i.i.d. selon la mesure de référence $\mu(\cdot)$. Puisque, pour tout $t \in [0, 1]$, $\mathbb{P}(E_1 < t) = 1 - (1 - t)^n$, alors $\mu(A_1)$ et E_1 ont même loi.

De manière similaire, notons maintenant $A_2 = \{s \in A_1^c : d(s, S_{(1)}) < D_2\}$ et

$A_{2,r} = \{s \in A_1^c : d(s, S_{(1)}) \leq r\}$. On a, pour tout $r > 0$,

$$\begin{aligned}
& \mathbb{P}\big(\mu(A_2) \leq \mu(A_{2,r}) \;\big|\; \mu(A_1)\big) = 1 - \mathbb{P}\big(\mu(A_2) > \mu(A_{2,r}) \;\big|\; \mu(A_1)\big) \\
= \;& 1 - \mathbb{P}\big(\forall i \in \{1 \leq j \leq n : S_j \neq S_{(1)}\},\, S_i \in A_{2,r}^c \,\big|\, \forall i \in \{1 \leq j \leq n : S_j \neq S_{(1)}\},\, S_i \in A_1^c\big) \\
= \;& 1 - \frac{\mathbb{P}\big(\forall i \in \{1 \leq j \leq n : S_j \neq S_{(1)}\},\, S_i \in A_{2,r}^c \cap A_1^c\big)}{\mathbb{P}\big(\forall i \in \{1 \leq j \leq n : S_j \neq S_{(1)}\},\, S_i \in A_1^c\big)} \\
= \;& 1 - \Big(\frac{\mu\big((A_1 \cup A_{2,r})^c\big)}{\mu(A_1^c)}\Big)^{n-1} = 1 - \Big(\frac{1 - \mu(A_1) - \mu(A_{2,r})}{1 - \mu(A_1)}\Big)^{n-1}.
\end{aligned}$$

La seconde égalité découle des définitions de A_1 et A_2, la troisième de la définition de la probabilité conditionnelle et la dernière est une conséquence de $A_1 \cap A_{2,r} = \emptyset$. Puisque, pour tout $s \in [0,1]$ et tout $t \in [0, 1-s]$, $\mathbb{P}(S_2 < t | S_1 = s) = 1 - \Big(\frac{1 - s - t}{1 - s}\Big)^{n-1}$, alors la distribution conditionnelle de $\mu(A_2)$ sachant $\mu(A_1)$ est la même que celle de E_2 sachant E_1. Comme les lois de $\mu(A_1)$ et E_1 sont les mêmes, la loi marginale de $\mu(A_2)$ est identique à celle de E_2. La même méthode s'applique récursivement à tous les $\mu(A_i)$, avec

$$\forall i \in [\![3, n]\!], \quad A_i = \Big\{ s \in \Big(\bigcup_{j \in [\![1, i-1]\!]} A_j \Big)^c : d(s, S_{(i-1)}) < D_i \Big\}$$

et

$$A_{n+1} = \Big(\bigcup_{j \in [\![1, n]\!]} A_j \Big)^c.$$

En conclusion, le vecteur des surfaces d'espacement que nous venons de définir, $\big(\mu(A_1), \cdots, \mu(A_{n+1})\big)$, est distribué identiquement au vecteur des espacements uniformes $\big(E_1, \cdots, E_{n+1}\big)$.

Les aires d'espacement $A_1, \cdots, A_{n+1}$ construites à partir d'un exemple jouet ($n = 8$ localisations sur $W = [0,1]^2$) sont représentées sur la Figure 1.1. Attention ! L'aire d'espacement A_9 est ici composée de deux parties non connexes : l'une en haut à droite de la fenêtre d'observation et l'autre en bas à gauche. La proximité géographique entre les événements $S_{(2)}$, $S_{(3)}$ et $S_{(4)}$ a pour conséquence des aires d'espacement consécutives A_3 et A_4 très réduites, pouvant donner lieu à la détection d'un agrégat.

Du cas spatial au cas temporel

On peut définir le processus ponctuel observé sur $[0,1]$ dont les événements ont pour localisations $\{T_1, \cdots, T_n\}$, où $\forall i \in [\![1, n]\!]$, $T_i = \sum_{j=1}^{i} \mu(A_j)$: il s'agit du processus ponctuel dont les espacements sont les surfaces d'espacement $\mu(A_1), \cdots, \mu(A_{n+1})$. La propriété que nous venons de prouver nous assure que, sous H_0, les événements $\{T_1, \cdots, T_n\}$ sont distribués comme des statistiques d'ordre issues d'un n-échantillon de la loi uniforme sur $[0,1]$. De plus, les distances uni-dimensionnelles entre T_i consecutifs représentent les distances spatiales entre $X_{(i)}$ consecutifs. Ainsi, à un agrégat temporel détecté sur $\{T_1, \cdots, T_n\}$ correspond une séquence inhabituelle de petites surfaces d'espacement, donc un agrégat spatial. D'un problème initial de détection

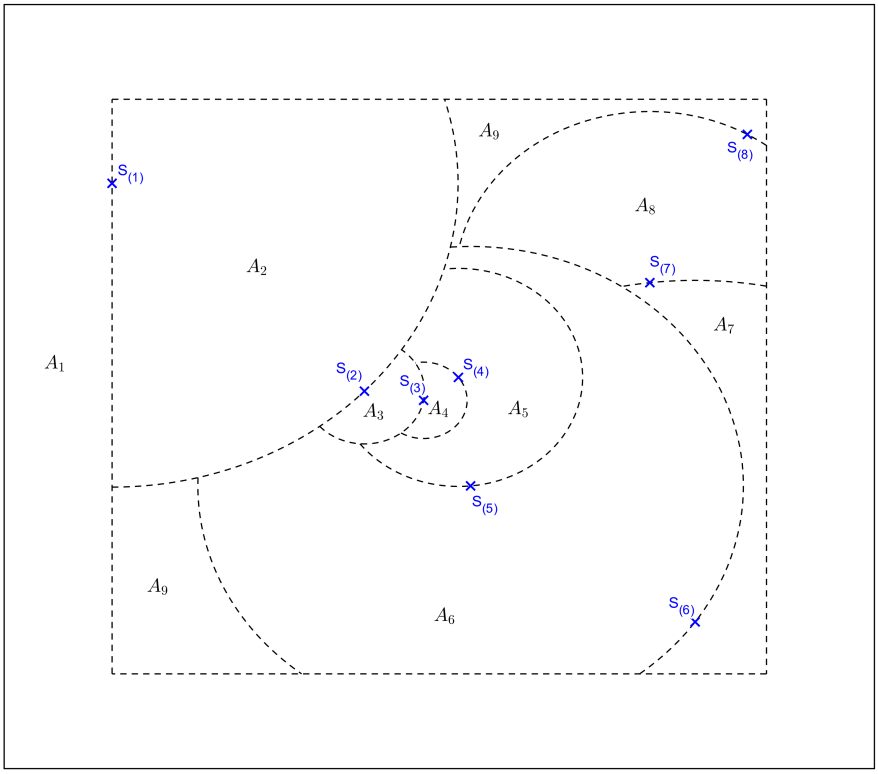

FIGURE 1.1 – Construction des aires d'espacement.

d'agrégats spatiaux, nous sommes passés à un problème de détection d'agrégats temporels. Ce problème peut être résolu par le calcul d'une statistique de balayage temporel et l'estimation de la significativité associée. Je détaillerai ce type de méthodes dans le Chapitre 2.

Une fois la détection d'agrégats temporels effectuée, il reste à les transformer en agrégats spatiaux. Supposons qu'une méthode de balayage temporel appliquée au processus $\{T_1, \cdots, T_n\}$ ait permis de détecter l'agrégat temporel le plus probable, que nous noterons $[t_a, t_b]$, $t_1, \cdots, t_n$ représentant les observations des variables aléatoires $T_1, \cdots, T_n$. L'ensemble des événements du jeu de données initial qui appartiennent à l'agrégat spatial correspondant est donc

$$\{s_{(a)}, \cdots, s_{(b)}\}.$$

Nous savons quels sont les événements inclus dans cet agrégat spatial, mais il nous faut définir sa forme plus précisément. Deux possibilités sont proposées par DEMATTEÏ, MOLINARI et DAURÈS 2007. La première consiste à identifier l'agrégat spatial à la réunion des cellules de Voronoi des événements inclus. La seconde consiste à définir pour chaque événement une zone d'influence : le cercle centré sur cet événement et dont la population est la population de W divisée par le nombre d'événements. L'agrégat est alors identifié à la réunion des zones d'influence des événements inclus. Ces deux propositions sont pertinentes mais j'ai préféré néanmoins adopter une autre solution. En effet, nous disposons des aires d'espacement $A_1, \cdots, A_{n+1}$ et nous savons dans lesquelles la faible population a permis d'identifier

l'agrégat temporel le plus probable. On peut donc considérer que l'agrégat spatial le plus probable est la réunion des aires d'espacement associées à l'agrégat temporel, i.e.

$$\hat{C} = \bigcup_{i \in [\![a+1,b]\!]} A_i.$$

La significativité de cet agrégat spatial est bien sûr donnée par celle de son équivalent temporel. J'ai baptisé cette procédure du nom de méthode flexible de balayage spatial. Il est à noter qu'il peut être utile de s'intéresser aux agrégats temporels secondaires (le deuxième plus probable, troisième plus probable...) et de considérer leurs équivalents spatiaux car la méthode d'ordonnancement des événements spatiaux décrite précédemment peut conduire à une trajectoire qui ressort d'un agrégat spatial avant de l'avoir parcouru totalement, avant d'y revenir par la suite.

3.2 Des agrégats potentiels basés sur les distances

Comme je l'ai dit en Section 1, de nombreuses méthodes ont été proposées pour donner naissance à des agrégats potentiels de forme non contrainte, mais elles sont toutes destinées à traiter des données groupées (Patil et Taillie 2004, Duczmal et Assuncão 2004, Tango et Takahashi 2005). Ces méthodes s'appuient toutes sur la notion de voisinage entre unités administratives. Puisque nous disposions de données ponctuelles, Christophe Dematteï et moi avons introduit une procédure de construction d'agrégats potentiels basée sur les distances entre événements.

Nous nous sommes appuyés sur un algorithme proposé par Bar-Hen, Koskas et Picard 2007 qui consiste à associer une famille de graphes au processus ponctuel original. Pour tout $\delta \in \mathbb{R}^+$, un graphe non-orienté, noté $\mathcal{G}(\delta)$, est défini de la manière suivante : l'ensemble des sommets est $\{1, \cdots, n\}$ et l'ensemble de ses arêtes est $\{(i,j) : d(s_i, s_j) \leq \delta \text{ et } 1 \leq i < j \leq n\}$. Chaque sommet i étant associé à l'événement de localisation s_i, il s'agit juste de relier les couples d'événements dont la distance est inférieure à δ. La composante connexe du sommet i dans ce graphe est notée $\mathcal{N}_i(\delta)$ et nous noterons $V_i(\delta) = \{s \in W : \exists j \in \mathcal{N}_i(\delta),\, d(s, s_j) \leq \delta\}$ le δ-voisinage associé. Puisque ces δ-voisinages regroupent des événements proches les uns des autres, nous pensons que ce sont des agrégats potentiels adéquats et nous définissons donc l'ensemble des agrégats potentiels de la manière suivante :

$$\mathcal{C} = \{V_i(\delta) : 1 \leq i \leq n \text{ et } \delta \in \mathbb{R}^+\}.$$

À première vue, le cardinal de cet ensemble peut paraître élevé. Néanmoins, il est possible de le réduire de manière drastique. Premièrement, l'ensemble des distances δ à analyser se limite à l'ensemble des distances entre couples d'événements $d_{i,j} = d(s_i, s_j)$, puisque le graphe $\mathcal{G}(\delta)$ reste inchangé quand δ croît entre deux $d_{i,j}$ consécutives. De plus, une nouvelle arête se rajoute au graphe $\mathcal{G}(\delta)$ lorsque δ atteint $d_{i,j}$ mais les composantes connexes du graphe, et par conséquent les δ-voisinages, peuvent rester identiques. Enfin, seul le δ-voisinage $V_i(d_{i,j}) = V_j(d_{i,j})$ doit être analysé lorsque δ atteint $d_{i,j}$ puisque les autres composantes connexes restent inchangées.

Soit $\mathcal{G}^-(\delta)$ le graphe dont l'ensemble des sommets est $\{1, \cdots, n\}$ et l'ensemble des arêtes est $\{(i,j) : d(s_i, s_j) < \delta \text{ et } 1 \leq i < j \leq n\}$. La composante connexe du sommet i dans ce graphe est notée $\mathcal{N}_i^-(\delta)$. Finalement, l'ensemble des agrégats

potentiels peut se réécrire

$$\mathcal{C} = \{V_i(d_{i,j}) : 1 \leq i < j \leq n \text{ et } \mathcal{N}_i^-(d_{i,j}) \neq \mathcal{N}_i(d_{i,j})\}.$$

Le processus que nous venons de décrire est similaire à la création d'un arbre couvrant euclidien par la liaison des sommets les plus proches, tant qu'aucune boucle n'est formée. On peut remarquer que cet arbre couvrant n'a aucune raison d'être de poids minimal (West 2000). Par conséquent, le nombre d'agrégats potentiels dans $\mathcal{C}$ est égal au nombre d'arêtes nécessaire pour relier les n sommets d'un graphe sans création de boucle, i.e. le nombre d'arêtes d'un arbre couvrant, soit $n-1$. Bien qu'initialement le nombre de distances entre couples d'événements $d_{i,j}$ à analyser soit de l'ordre de n^2, le nombre d'agrégats potentiels à considérer ici est très fortement réduit par rapport aux familles classiques d'agrégats potentiels de forme contrainte.

Sur le même exemple jouet que précédemment, la figure 1.2 représente la construction des six premiers δ-voisinages. Là encore, la proximité géographique entre les événements $S_{(2)}$, $S_{(3)}$ et $S_{(4)}$ a pour conséquence la création d'un δ-voisinage d'aire réduite contenant ces trois localisations (en haut, à droite), pouvant donner lieu à la détection d'un agrégat.

3.3 Une distance spatio-temporelle

Les méthodes que nous venons de décrire dans les deux sous-sections précédentes sont basées sur l'utilisation de la distance euclidienne dans un sous-espace de $\mathbb{R}^2$ (ou $\mathbb{R}^3$ si les localisations spatiales sont tri-dimensionnelles). Malheureusement, lorsque les localisations des événements sont spatio-temporelles, la notion de distance euclidienne n'est plus clairement définie puisque les dimensions spatiales et temporelles ne jouent pas le même rôle. Pour surmonter cette difficulté, Christophe Dematteï et moi avons proposé la création d'une distance spatio-temporelle, fonction des distances euclidiennes dans l'espace et dans le temps.

Pour cela, nous avons besoin d'un paramètre de correspondance entre l'espace et le temps, dont nous fixons la valeur de la manière suivante. Nous supposons que la fenêtre d'observation du processus se décompose ainsi :

$$W = A \times T,$$

où A, sous-ensemble de $\mathbb{R}^2$ (ou $\mathbb{R}^3$), est la surface d'observation et T l'intervalle de temps d'observation. Notons $|A|$ l'aire de A et $|T|$ la longueur de T. Soit $D = 2\sqrt{\frac{|A|}{\pi}}$ le diamètre du disque dont l'aire est $|A|$, ce qui représente également la distance spatiale maximale entre deux points de ce disque. Nous considérons que la distance temporelle maximale entre deux événements, égale à $|T|$, doit avoir le même poids que cette distance spatiale maximale D. Ainsi, la distance spatio-temporelle que nous proposons, notée $d_{ST}(\cdot)$, est définie par

$$d_{ST}\big((y,t),(y',t')\big) = \sqrt{d_S(y,y')^2 + \frac{D^2}{|T|^2} d_T(t,t')^2},$$

où $d_S(\cdot)$ et $d_T(\cdot)$ représentent respectivement les distances euclidiennes spatiale et temporelle. Intuitivement, cette distance spatio-temporelle peut se voir comme une

distance euclidienne dans $\mathbb{R}^3$ (ou $\mathbb{R}^4$ si les localisations spatiales initiales sont tridimensionnelles), après changement d'échelle sur l'axe temporel. De manière plus rigoureuse, on a

$$d_{ST}\big((y,t),(y',t')\big) = d\big((y,\frac{D}{T}t),(y',\frac{D}{T}t')\big),$$

ce qui montre bien que $d_{ST}(\cdot)$ est une distance correctement définie. En utilisant cette distance spatio-temporelle, il est donc possible d'utiliser l'ensemble des techniques de création d'ensemble d'agrégats potentiels basés sur les distances décrites précédemment, même lorsque les localisations sont spatio-temporelles.

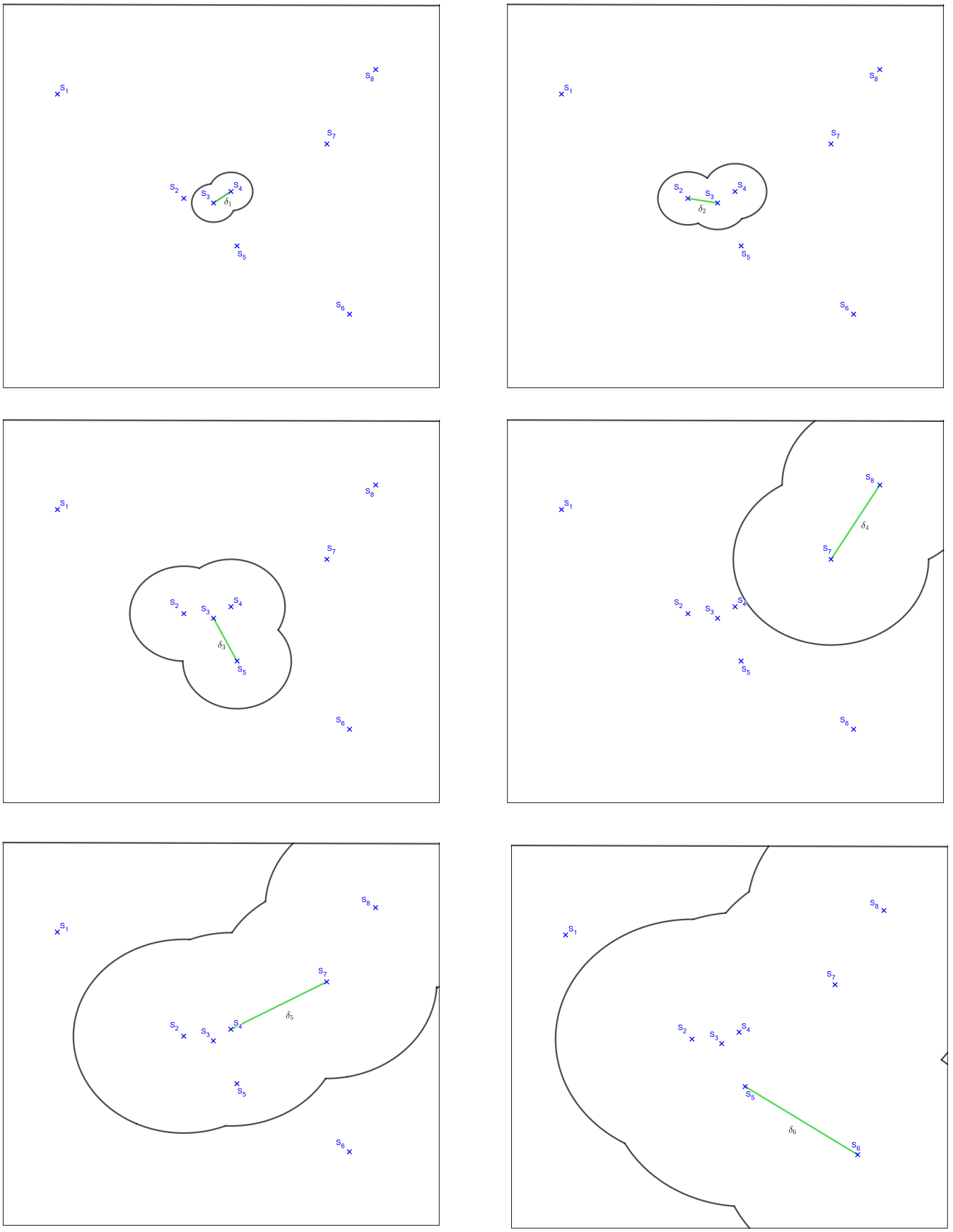

FIGURE 1.2 – Construction des δ-voisinages.

Chapitre 2

Des indices de concentration alternatifs

Publications : [1], [5], [6], [7], [8], [11].

Dans ce chapitre, je m'intéresse précisément aux différentes façons de construire un indice de concentration $I(Z)$ permettant de départager les agrégats potentiels $\{Z \in \mathcal{C}\}$, l'ensemble des agrégats potentiels $\mathcal{C}$ ayant été créé par n'importe laquelle des méthodes décrites au Chapitre 1. Historiquement, les premiers indices de concentration sont basés sur le calcul de rapports de vraisemblance. Dans un premier temps, je me consacrerai au cas de processus non marqués. Après avoir détaillé les méthodes s'appuyant sur la vraisemblance, je montrerai comment je m'en suis affranchi en utilisant une propriété de distribution des espacements directement exploitée dans le cadre temporel dans [1], puis adaptée aux cadres spatial et spatio-temporel dans [11]. Ensuite, nous basculerons dans le cadre de processus marqués par une variable réelle. La plupart du temps, on cherche à identifier des zones où la moyenne des marques est significativement différente. Là encore, les indices de concentration issus de rapports de vraisemblance peuvent être remplacés par des indices de concentration non-paramétriques, basés sur les moments, comme dans [5] et [8], ou sur les rangs, comme dans [6]. Néanmoins, on peut aussi avoir besoin de chercher des zones où c'est la variance des marques qui diffère significativement, comme dans [7].

1 Processus non marqué

Rappellons que l'objectif est ici de comparer les localisations d'événements observés $\{s_i : i \in [\![1, n]\!]\}$ à une population sous-jacente dont la distribution est décrite par la mesure de probabilité $\mu(\cdot)$ sur la fenêtre d'observation W, qui sera par défaut la mesure uniforme. L'indice de concentration $I(Z)$ doit nous permettre de déterminer dans quel agrégat potentiel $Z \in \mathcal{C}$ le nombre de localisations s_i est le plus élevé (ou le plus faible) par rapport à $\mu(Z)$, la proportion de la population de W contenue dans Z.

Supposons tout d'abord que la mesure de probabilité $\mu(\cdot)$ soit absolument continue par rapport à la mesure de Lebesgue et qu'elle soit associée à la densité de

probabilité $\phi(\cdot)$, de telle sorte que

$$\forall Z \subset W, \quad \mu(Z) = \int_Z \phi(s)ds.$$

On peut signaler que, si $\mu(\cdot)$ n'est pas absolument continue par rapport à la mesure de Lebesgue, le processus non marqué peut être transformé en processus marqué : nous y reviendrons en Section 2.

Notons $N(\cdot)$ le processus de comptage associé à ce processus ponctuel : pour tout $Z \subset W$, la variable aléatoire $N(Z)$ représente le nombre d'événements dans Z et on a donc $N(W) = N$. A la naissance des statistiques de balayage, Naus 1963 se contentait de comparer des sous-ensembles Z et Z' contenant la même proportion de population $p \in [0, 1]$, i.e. tels que $\mu(Z) = \mu(Z') = p$. On parle de méthode de balayage à fenêtre fixe. L'énorme avantage de ce choix est qu'alors, pour comparer la concentration d'événements dans Z et dans Z', il suffit de comparer $N(Z)$ et $N(Z')$. Néanmoins, deux inconvénients majeurs ne plaident pas en faveur de l'application de cette méthode. D'abord, la valeur du paramètre p doit être fixée arbitrairement. Ensuite, lorsque $\mu(\cdot)$ n'est pas la mesure uniforme, il n'est pas facile de construire des sous-ensembles Z tels que $\mu(Z) = p$.

Ce n'est que trente ans plus tard que Nagarwalla 1996 et Kulldorff 1997 ont proposé des méthodes de balayage à fenêtre variable en s'appuyant sur des rapports de vraisemblance généralisés.

1.1 Un indice de concentration basé sur la vraisemblance

L'idée générale des méthodes de balayage basées sur la vraisemblance est la suivante :

1. On considère un modèle paramétrique $\mathcal{M}_0$ synonyme d'absence d'agrégat significatif dans W.
2. On introduit, pour chaque agrégat potentiel $Z \in \mathcal{C}$, un modèle paramétrique $\mathcal{M}_{1,Z}$ traduisant 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 maximale des observations sous le modèle $\mathcal{M}_{1,Z}$ et L^*_0 la vraisemblance maximale 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

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

Construire un indice de concentration basé sur ce rapport de vraisemblance s'avère totalement pertinent puisque ce dernier sera d'autant plus élevé que le modèle $\mathcal{M}_{1,Z}$ sera vraisemblable. Nous allons maintenant préciser les modèles paramétriques proposées par Kulldorff 1997 et Nagarwalla 1996, calculer les rapports de vraisemblance correspondants et construire un indice de concentration associé.

Le modèle Poissonnien

KULLDORFF 1997 propose de considérer que les observations $\{s_1, \cdots, s_n\}$ sont issues d'un processus ponctuel de Poisson inhomogène, d'intensité $\psi(.)$. Par conséquent, le nombre d'observations N dans W suit une loi de Poisson de paramètre $\int_W \psi(s)ds$ et les localisations des événements sont indépendantes et identiquement distribuées selon la fonction densité $\dfrac{\psi(.)}{\int_W \psi(s)ds}$.

Sous le modèle $\mathcal{M}_0$, l'intensité est supposée totalement proportionnelle à la densité de probabilité de la population sous-jacente :

$$\forall s \in W, \quad \psi(s) = k\phi(s),$$

où $k \in \mathbb{R}^+$ est un paramètre du modèle qui sera estimé par maximum de vraisemblance. Sous le modèle $\mathcal{M}_{1,Z}$, le rapport entre l' intensité et la densité de probabilité de la population sous-jacente diffère suivant que l'on se trouve dans Z ou pas :

$$\forall s \in W, \quad \psi(s) = \phi(s)\big(k_Z \mathbb{1}_Z(s) + k_{Z^c}\mathbb{1}_{Z^c}(s)\big),$$

où $\mathbb{1}_Z(.)$ est la fonction indicatrice sur Z et $k_Z \in \mathbb{R}^+$ et $k_{Z^c} \in \mathbb{R}^+$ sont deux paramètres du modèle qui seront estimés par maximum de vraisemblance.

Sous le modèle $\mathcal{M}_0$, l'espérance du nombre d'événements dans W vaut donc

$$\int_W \psi(s)ds = \int_W k\phi(s)ds = k$$

et la vraisemblance maximale des observations est donc donnée par

$$\begin{aligned} L_0^* &= \max_{k\in\mathbb{R}^+} \exp(-k)\frac{k^n}{n!}\prod_{i=1}^n \frac{k\phi(s_i)}{\int_W k\phi(s)ds} \\ &= \max_{k\in\mathbb{R}^+} \exp(-k)\frac{k^n}{n!}\prod_{i=1}^n \phi(s_i) \\ &= \exp(-n)\frac{n^n}{n!}\prod_{i=1}^n \phi(s_i). \end{aligned}$$

Sous le modèle $\mathcal{M}_{1,Z}$, l'espérance du nombre d'événements dans W vaut

$$\int_W \psi(s)ds = \int_W \phi(s)\big(k_Z\mathbb{1}_Z(s) + k_{Z^c}\mathbb{1}_{Z^c}(s)\big) = k_Z\mu(Z) + k_{Z^c}\mu(Z^c)$$

et la vraisemblance maximale des observations est donnée par

$$\begin{aligned} L_{1,Z}^* &= \max_{(k_Z,k_{Z^c})\in\mathbb{R}^{+2}} \exp(-k_Z\mu(Z) - k_{Z^c}\mu(Z^c))\frac{\big(k_Z\mu(Z) + k_{Z^c}\mu(Z^c)\big)^n}{n!} \\ &\quad \prod_{i=1}^n \frac{\phi(s_i)\big(k_Z\mathbb{1}_Z(s_i) + k_{Z^c}\mathbb{1}_{Z^c}(s_i)\big)}{k_Z\mu(Z) + k_{Z^c}\mu(Z^c)} \\ &= \max_{(k_Z,k_{Z^c})\in\mathbb{R}^{+2}} \exp(-k_Z\mu(Z) - k_{Z^c}\mu(Z^c))\frac{k_Z^{n(Z)}k_{Z^c}^{n(Z^c)}}{n!}\prod_{i=1}^n \phi(s_i) \\ &= \exp\big(-n(Z) - n(Z^c)\big)\frac{\left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)}\left(\frac{n(Z^c)}{\mu(Z^c)}\right)^{n(Z^c)}}{n!}\prod_{i=1}^n \phi(s_i). \end{aligned}$$

Finalement, le rapport de vraisemblance entre les deux modèles est :

$$RV(Z) = \frac{L^*_{1,Z}}{L^*_0} = \frac{\left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)} \left(\frac{n(Z^c)}{\mu(Z^c)}\right)^{n(Z^c)}}{n^n}.$$

L'approche conditionnelle

L'approche de Nagarwalla 1996 est légérement différente puisqu'il considère un nombre d'événements dans W déterministe et suppose des localisations $S_1, \cdots, S_n$ indépendantes et identiquement distribuées selon la fonction densité $\gamma(\cdot)$. Nous l'appelons approche conditionnelle car, conditionnellement à $N = n$, elle est totalement identique au modèle Poissonnien décrit précédemment.

Sous le modèle $\mathcal{M}_0$, la densité des localisations est égale à la densité de la population sous-jacente :

$$\forall s \in W, \quad \gamma(s) = \phi(s).$$

Sous le modèle $\mathcal{M}_{1,Z}$, le rapport entre la densité des localisations et la densité de la population sous-jacente diffère suivant que l'on se trouve dans Z ou pas :

$$\forall s \in W, \quad \gamma(s) = \phi(s)\Big(r_Z \mathbb{1}_Z(s) + \frac{1 - r_Z \mu(Z)}{\mu(Z^c)} \mathbb{1}_{Z^c}(s)\Big),$$

où $r_Z \in \mathbb{R}^+$ est un paramètre du modèle qui sera estimé par maximum de vraisemblance. Sous le modèle $\mathcal{M}_0$, la vraisemblance maximale des observations est donc donnée par

$$L^*_0 = \prod_{i=1}^{n} \phi(s_i)$$

et sous le modèle $\mathcal{M}_{1,Z}$, elle est donnée par

$$\begin{aligned} L^*_{1,Z} &= \max_{r_Z \in \mathbb{R}^+} \prod_{i=1}^{n} \phi(s_i)\Big(r_Z \mathbb{1}_Z(s_i) + \frac{1 - r_Z \mu(Z)}{\mu(Z^c)} \mathbb{1}_{Z^c}(s_i)\Big) \\ &= \max_{r_Z \in \mathbb{R}^+} r_Z^{n(Z)} \Big(\frac{1 - r_Z \mu(Z)}{\mu(Z^c)}\Big)^{n(Z^c)} \prod_{i=1}^{n} \phi(s_i) \\ &= \Big(\frac{n(Z)}{n\mu(Z)}\Big)^{n(Z)} \Big(\frac{n(Z^c)}{n\mu(Z^c)}\Big)^{n(Z^c)} \prod_{i=1}^{n} \phi(s_i) \end{aligned}$$

Finalement, le rapport de vraisemblance entre les deux modèles est :

$$RV(Z) = \frac{L^*_{1,Z}}{L^*_0} = \Big(\frac{n(Z)}{n\mu(Z)}\Big)^{n(Z)} \Big(\frac{n(Z^c)}{n\mu(Z^c)}\Big)^{n(Z^c)} = \frac{\left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)} \left(\frac{n(Z^c)}{\mu(Z^c)}\right)^{n(Z^c)}}{n^n}.$$

On retrouve exactement le même rapport de vraisemblance que sous le modèle Poissonnien, ce qui montre que la prise en compte du caractère aléatoire du nombre total d'observations n'apporte ici aucune modification.

Construction de l'indice de concentration

Comme je l'ai écrit précédemment, ce rapport de vraisemblance traduit la pertinence d'un modèle différenciant les observations dans Z et dans Z^c. Peut-on pour autant l'utiliser directement comme indice de concentration ? Pas tout à fait car il ne fait pas la différence entre un excès et un défaut d'événements dans Z : $\forall Z \subset W, RV(Z) = RV(Z^c)$. Pour différencier ces deux cas, on introduit généralement dans l'indice de concentration un terme indiquant si le nombre d'événements dans Z est supérieur ou inférieur à celui attendu sous le modèle $\mathcal{M}_0$. Afin de manipuler des nombres moins grands, on préfère également utiliser le logarithme du rapport de vraisemblance. Finalement, l'indice de concentration basé sur la vraisemblance est

$$\begin{aligned} I_{RV}(Z) &= \log\big(RV(Z)\big)\Big(\mathbb{1}\big(n(Z) > n\mu(Z)\big) - \mathbb{1}\big(n(Z) < n\mu(Z)\big)\Big) \\ &= \Big[n(Z)\log\Big(\frac{n(Z)}{\mu(Z)}\Big) + \big(n - n(Z)\big)\log\Big(\frac{n - n(Z)}{1-\mu(Z)}\Big) - n\log(n)\Big] \\ &\quad \Big(\mathbb{1}\big(n(Z) > n\mu(Z)\big) - \mathbb{1}\big(n(Z) < n\mu(Z)\big)\Big). \end{aligned}$$

Cet indice est positif lorsque le nombre d'événements est en excès, négatif lorsqu'il est en défaut et nul lorsqu'il est égal au nombre attendu en l'absence d'agrégat. Il est bien sûr croissant selon $n(Z)$ et décroissant selon $\mu(Z)$.

1.2 Un indice de concentration basé sur les espacements

Les indices de concentration décrits ci-dessus sont efficaces mais dépendent fortement de la construction des hypothèses alternatives (constantes par morceaux). Dans cette sous-section, je montre comment on peut s'affranchir de ce choix. Je me focalise d'abord sur les processus ponctuels temporels afin de construire un indice de concentration exploitant les propriétés théoriques des espacements, puis je l'étends aux cas spatial et spatio-temporel.

Construction de l'indice de concentration

Lorsque le processus ponctuel qui nous intéresse est un processus temporel observé sur $W = [0, T]$, les localisations $S_1, \cdots, S_n$ sont unidimensionnelles donc facilement ordonnables. Afin de prendre en compte la densité de population sous-jacente, nous allons d'abord introduire les variables aléatoires suivantes :

$$\forall i \in [\![1, n]\!], \quad T_i = \int_0^{S_i} \phi(s)ds.$$

Plaçons-nous désormais sous l'hypothèse nulle d'absence d'agrégats : les variables $T_1, \cdots, T_n$ sont indépendantes et uniformément distribuées sur $[0, 1]$. Soit

$$0 = T_{(0)} \leq T_{(1)} \leq \cdots \leq T_{(n)} \leq T_{(n+1)} = 1$$

les statistiques d'ordre associées et, pour tout $k \in [\![1, n+1]\!]$,

$$D_k = T_{(k)} - T_{(k-1)}$$

les espacements associés. On rappelle que, dans ce cadre temporel, les agrégats potentiels sont les intervalles $[T_{(i)}, T_{(j)}]$ avec $1 \leq i < j \leq n$. Les longueurs de chacun de ces intervalles,

$$D_{i,j} = T_{(j)} - T_{(i)} = \sum_{k=i+1}^{j} D_k,$$

sont connues sous le nom d'espacements du $(j-i)^{\text{ème}}$ ordre associés à $(T_1, \cdots, T_n)$. Leur distribution est donnée par David 1981 : $D_{i,j} \sim \beta(j-i, n+1-j+i)$. Afin de pouvoir comparer ces longueurs, nous introduisons

$$U_{i,j} = 1 - B_{inc}(D_{i,j}, j-i, n+1-j+i)$$

où $B_{inc}(\cdot, j-i, n+1-j+i)$ est la fonction Beta incomplete, i.e. la fonction de répartition de $D_{i,j}$. Toutes les variables aléatoires $U_{i,j}$ sont distribuées uniformément sur $[0,1]$ donc peuvent être comparées. De plus, la concentration en événements sur l'intervalle $[T_{(i)}, T_{(j)}]$ est d'autant plus grande que $U_{i,j}$ est grande. Nous pouvons désormais introduire l'indice de concentration basé sur les espacements

$$I_{ES}([T_{(i)}, T_{(j)}]) = U_{i,j} = 1 - B_{inc}(T_{(j)} - T_{(i)}, j-i, n+1-j+i). \tag{2.1}$$

Remarques sur le calcul de la significativité

Comme je l'ai écrit dans l'Introduction de ce mémoire, calculer la distribution d'une statistique de balayage à fenêtre variable est tout sauf une sinécure et l'on doit généralement se contenter, pour calculer la p-valeur associée à une telle statistique, de se baser sur des simulations de type Monte-Carlo. Néanmoins, en plus de la possibilité de pouvoir utiliser l'indice de concentration $I_{ES}(\cdot)$, le cadre temporel pourrait permettre également, au prix de quelques efforts, de calculer la distribution sous l'hypothèse nulle de la statistique de balayage correspondante

$$\lambda_{ES} = \max_{1 \leq i < j \leq n} I_{ES}([T_{(i)}, T_{(j)}]).$$

Nous souhaitons exprimer

$$P(t,n) = \mathbb{P}_0(\lambda_{ES} < t),$$

où $\mathbb{P}_0$ est la probabilité sous l'hypothèse nulle. Soit $b_{\alpha,n,m}$ le quantile de la loi Beta tel que $B_{inc}(b_{\alpha,n,m}, m, n+1-m) = \alpha$. On peut écrire

$$\begin{aligned}
P(t,n) &= \mathbb{P}_0\Big(\max_{1 \leq i < j \leq n} I_{ES}(i,j) < t\Big) \\
&= \mathbb{P}_0\big(\forall\, 1 \leq i < j \leq n,\, U_{i,j} < t\big) \\
&= \mathbb{P}_0\big(\forall\, 1 \leq i < j \leq n,\, D_{i,j} > b_{1-t,n,j-i}\big) \\
&= \mathbb{P}_0\Big(\forall\, 1 \leq i < j \leq n,\, \sum_{k=i+1}^{j} D_k > b_{1-t,n,j-i}\Big).
\end{aligned}$$

Or, Huffer et Lin 2001 ont introduit un algorithme permettant de calculer la distribution jointe de combinaisons linéaires d'espacements uniformes (i.e. des espacements issus d'un n-échantillon de la loi uniforme sur $[0,1]$). Cet algorithme,

qui consiste en un usage répété et systématique de deux récursions de base, nous permet d'aboutir à une somme d'éléments simples qui s'écrit sous forme explicite. A ma connaissance, il a principalement été utilisé pour calculer la distribution nulle de statistiques de balayage à fenêtre fixe (HUFFER et LIN 1999) mais l'extension à notre statistique de balayage à fenêtre variable est très simple, la seule différence étant que le nombre de combinaisons linéaires d'espacements à considérer n'est plus de l'ordre de n mais de n^2. En appliquant cet algorithme, on obtient par exemple que

$$P(t,3) = \begin{cases} 0 & \text{si } t \in [0, 1/8], \\ 4b_{1-t,3,1}^3(1-2b_{1-t,3,1}) + (1-2b_{1-t,3,1})^4 & \text{si } t \in (1/8, t^*], \\ (1-b_{1-t,2,2})[(1-b_{1-t,2,2})^2(3b_{1-t,2,2} - 8b_{1-t,3,1} + 1) + 4b_{1-t,3,1}^3] & \text{si } t \in [t^*, 1], \end{cases}$$

où $t^* \approx 1/3$ est le réel tel que $b_{1-t^*,3,1} - 2b_{1-t^*,2,2} = 0$. Pour de grandes valeurs de n, cet algorithme nécessite l'usage d'un ordinateur puisqu'il manipule des matrices de très grande dimension. Cependant, ces matrices peuvent être mises sous la forme de matrices triangulaires inférieures par blocs grâce à la permutation et la suppression de certaines de ses lignes ou de ses colonnes. Malheureusement, l'écriture d'un programme informatique traduisant cet algorithme est loin d'être évidente et le programme en C fourni par les auteurs a été conçu pour fonctionner uniquement avec des termes rationnels, ce que ne sont pas les quantiles $b_{\alpha,n,m}$.

Extension aux cas spatial et spatio-temporel

Lorsque le processus qui nous intéresse est un processus spatial ou spatio-temporel, l'indice de concentration $I_{ES}(\cdot)$ ne peut être appliqué tel quel. Néanmoins, il est facile de l'adapter à ces cas-là. Dans l'équation (2.1), l'indice défini mesure la concentration associée à la présence de $j-i+1$ événements sur un intervalle qui contient une part de la population sous-jacente égale à $T_{(j)} - T_{(i)}$. Si l'on veut généraliser cette expression pour mesurer la concentration dans n'importe quel agrégat potentiel $Z \subset W$, contenant $n(Z)$ événements pour une part de la population sous-jacente égale à $\mu(Z)$, il suffit alors de substituer $n(Z)$ à $j-i+1$ et $\mu(Z)$ à $T_{(j)} - T_{(i)}$ dans l'équation (2.1) pour obtenir

$$I_{ES}(Z) = 1 - B_{inc}\big(\mu(Z), n(Z)-1, n+2-n(Z)\big).$$

Il s'agit bien d'une extension de l'indice introduit dans le cadre temporel puisqu'on le retrouve en prenant $Z = [T_{(i)}, T_{(j)}]$. Bien sûr, dans les cadres spatial et spatio-temporel, il est bien plus difficile d'identifier des propriétés probabilistes communes aux agrégats potentiels $Z \in \mathcal{C}$, et ce quelle que soit la méthode choisie pour générer la famille $\mathcal{C}$ (voir chapitre 1). Dans ces conditions, il nous est impossible de donner la loi de probabilité de $I_{ES}(Z)$ même sous l'hypothèse d'absence d'agrégat.

1.3 Comparaisons

Qualités des différents indices de concentration

Pour comparer les indices de concentration proposés, on doit s'interroger sur les qualités qu'un indice de concentration peut posséder. Idéalement, un indice de

concentration $I(Z)$ doit traiter tous les agrégats potentiels sur un même pied d'égalité, quelle que soit leur taille. Autrement dit, sous l'hypothèse nulle traduisant l'absence d'agrégat, le comportement de $I(Z)$ ne devrait pas dépendre de la valeur de $\mu(Z)$. Mathématiquement, on peut donc considérer la propriété P1 : "En l'absence d'agrégat, la loi de probabilité de $I(Z)$ reste identique pour tout $Z \in \mathcal{C}$".

Dans les cadres spatial et spatio-temporel, il apparaît très difficile de vérifier la véracité de cette propriété car la loi de probabilité de $\mu(Z)$ va dépendre totalement de la façon dont les agrégats potentiels $Z \in \mathcal{C}$ ont été construits. Par contre, dans le cadre temporel, on peut exploiter le fait que les agrégats potentiels sont les intervalles $[T_{(i)}, T_{(j)}]$ avec $1 \leq i < j \leq n$, de longueurs $D_{i,j} = T_{(j)} - T_{(i)}$ et contenant $k = j - i + 1$ événements. On rappelle que, sous l'hypothèse nulle d'absence d'agrégats, ces longueurs suivent une loi Beta : $D_{i,j} \sim \beta(k-1, n+2-k)$. Par conséquent, l'indice de concentration basé sur les espacements

$$I_{ES}([T_{(i)}, T_{(j)}]) = 1 - B_{inc}(D_{i,j}, k-1, n+2-k)$$

suit une loi uniforme sur $[0,1]$, quelle que soit la valeur de k : la propriété P1 est bien respectée par cet indice de concentration. Concernant l'indice de concentration basé sur le rapport de vraisemblance,

$$I_{RV}([T_{(i)}, T_{(j)}]) = \Big[k \log\Big(\frac{k}{D_{i,j}}\Big) + (n-k)\log\Big(\frac{n-k}{1-D_{i,j}}\Big) - n\log(n)\Big]\Big(\mathbb{1}\big(k > nD_{i,j}\big) - \mathbb{1}\big(k < nD_{i,j}\big)\Big),$$

le calcul de sa distribution est bien plus complexe. Comme le mentionne Wilks 1938, le rapport de vraisemblance $RV(Z)$ tend vers une loi du chi-deux dont le degré de liberté ne dépend que du nombre de paramètres associé aux modèles $\mathcal{M}_0$ et $\mathcal{M}_{1,Z}$, et pas du tout de la valeur de $\mu(Z)$. Mais ce n'est qu'un résultat asymptotique, et l'expression de la loi de probabilité pour tout n et tout k semble inatteignable. Néanmoins, on peut sans grande difficulté fixer les valeurs de n et de k et donner une valeur approchée de la fonction de répartition de $I_{RV}([T_{(i)}, T_{(j)}])$ en tout point. Un exemple vous est donné par la Figure 2.1 en prenant $n = 100$ et différentes valeurs de k. On constate que les lois de probabilité de $I_{RV}([T_{(i)}, T_{(j)}])$ sont fortement dépendantes de k. Les indices de concentration associés à des intervalles contenant peu d'événements ont une probabilité plus élevée d'atteindre de grandes valeurs, ce qui risque d'avoir une influence sur les résultats de la méthode de balayage associée.

Etude de simulation

Nous souhaitons maintenant comparer les performances des méthodes de balayage basées sur les différents indices de concentration décrits précédemment. Pour cela, nous simulons sur une fenêtre d'observation W des jeux de données exhibant un agrégat, noté $C \subset W$. L'efficacité d'une méthode de détection d'agrégat se mesurant à la fois par sa capacité à générer une alarme en cas d'agrégat et à identifier les contours de cet agrégat, nous calculons les critères suivants :

— la puissance, i.e. la proportion de jeux de données simulés pour lesquels la p-valeur est inférieure au niveau du test α.

— des taux de "vrai positifs" et de "faux positifs" mesurant la correspondance entre le "vrai agrégat" C et l'agrégat le plus probable $\hat{C}$, donnés par

$$VP = \frac{\mu(C \cap \hat{C})}{\mu(C)} \text{ et } FP = \frac{\mu(C^c \cap \hat{C})}{\mu(C^c)}.$$

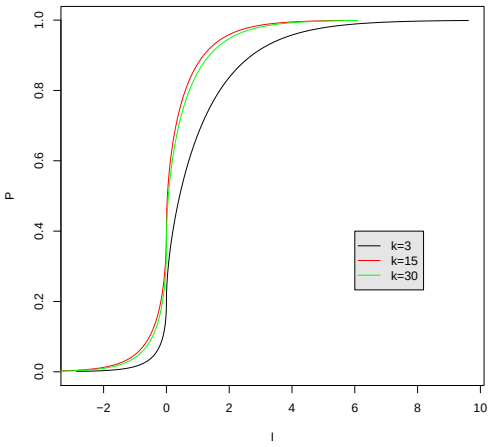

FIGURE 2.1 – FdR de $I_{RV}([T_{(i)}, T_{(j)}])$.

Localisations temporelles Nous simulons $n = 100$ événements sur l'intervalle $[0, 1]$ selon différents scénarios d'agrégation, mais toujours avec l'agrégat $C = [0.4, 0.6]$. Un agrégat de type "palier" est obtenu en utilisant la fonction densité

$$f_1(s) = \frac{1}{0.8 + 0.2r}\mathbb{1}_{[0,0.4]\cup[0.6,1]}(s) + \frac{r}{0.8 + 0.2r}\mathbb{1}_{[0.4,0.6]}(s).$$

Un agrégat de type "cloche" est obtenu en utilisant la fonction densité

$$\begin{aligned} f_2(s) = & \frac{15}{2r + 13}\mathbb{1}_{[0,0.4]\cup[0.6,1]}(s) \\ & + \frac{15}{2r + 13}\{1 + (r - 1) * [1 - 100(s - 0.5)^2]\}\mathbb{1}_{[0.4,0.6]}(s). \end{aligned}$$

A chaque fois, le paramètre r est le ratio entre le maximum et le minimum de la fonction densité : il mesure donc l'intensité de l'agrégat. Ces deux fonctions densité sont représentées sur la Figure 2.2 pour différentes valeurs de r.

Nous avons simulé 1000 jeux de données selon chaque fonction densité. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 9999$ simulations selon la loi uniforme sur $[0, 1]$. La table 2.1 donne les résultats obtenus pour les deux types d'agrégat. Les valeurs en gras indiquent les meilleurs résultats obtenus parmi toutes les méthodes.

Les résultats sont assez éloquents : la méthode qui utilise l'indice de concentration basé sur la loi des espacements est bien plus puissante que celle basée sur le rapport de vraisemblance, et ce quel que soit le scénario d'agrégation et l'intensité d'agrégat. Concernant l'identification du "vrai" agrégat, les taux de vrais positifs sont aussi bien plus élevés pour cette méthode Λ_{ES}. Quant aux taux de faux positifs, ils sont légèrement plus faibles pour la méthode Λ_{RV} et restent de toute façon dans des valeurs très faibles. Il est à noter que le scénario d'agrégat palier correspond à une densité constante par morceaux, du même type que celles qui sous-tendent l'indice de concentration basé sur le rapport de vraisemblance. Cela peut expliquer que la méthode Λ_{RV} donne des résultats un peu moins décevants contre ce type d'alternative.

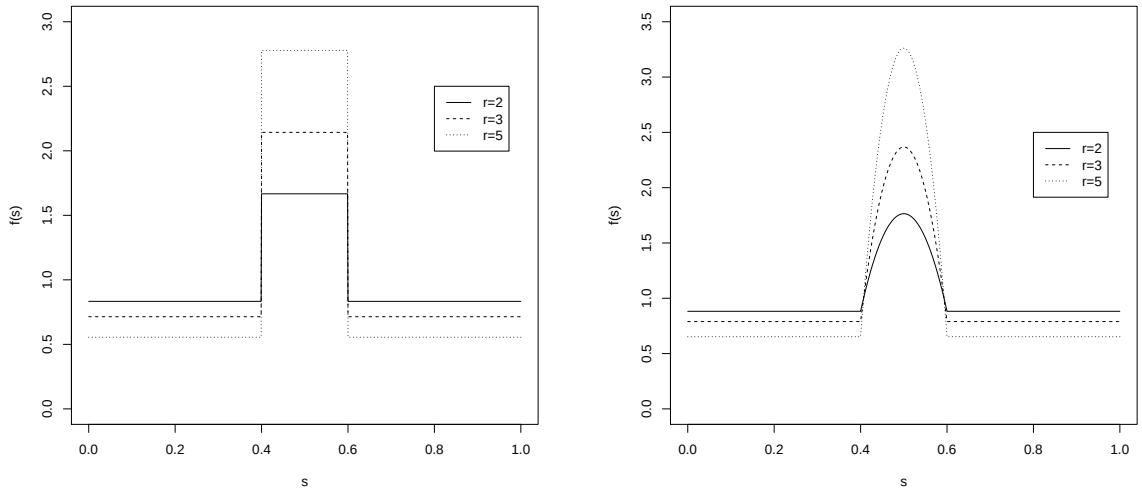

FIGURE 2.2 – Fonctions densité d'agrégat.

TABLE 2.1 – Tests appliqués à deux types d'agrégats sur $[0, 1]$

r		Palier		Cloche	
		Λ_{ES}	Λ_{RV}	Λ_{ES}	Λ_{RV}
2	Power	**0.489**	0.133	**0.315**	0.073
	TP	**0.735**	0.477	**0.556**	0.165
	FP	0.035	**0.015**	0.031	**0.004**
2.5	Power	**0.812**	0.368	**0.576**	0.174
	TP	**0.804**	0.712	**0.618**	0.463
	FP	0.018	**0.013**	0.022	**0.008**
3	Power	**0.954**	0.627	**0.835**	0.362
	TP	**0.863**	0.847	**0.652**	0.556
	FP	0.015	**0.011**	0.013	**0.006**

Localisations spatiales Nous simulons $n = 100$ événements sur l'intervalle $[0, 1]^2$ selon différents scénarios d'agrégation, mais toujours avec l'agrégat $C = [0.3, 0.7]^2$. Un agrégat de type "palier" est obtenu en utilisant la fonction densité

$$f_1(s) \propto 1 + (r - 1)\mathbb{1}_C(s).$$

Un agrégat de type "cloche" est obtenu en utilisant la fonction densité

$$f_2(s) \propto 1 + (r - 1)\big(1 - 12.5d(s, s_C)\big)\mathbb{1}_C(s)$$

où $s_C = [0.5, 0.5]$ est le centre de l'agrégat C. Encore une fois, le paramètre r est le ratio entre le maximum et le minimum de la fonction densité et mesure donc l'intensité de l'agrégat.

Nous avons simulé 100 jeux de données selon chaque fonction densité. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 99$ simulations

selon la loi uniforme sur $[0,1]^2$. La table 2.2 donne les résultats obtenus pour les deux types d'agrégat.

TABLE 2.2 – Tests appliqués à deux types d'agrégats sur $[0,1]^2$

r		Palier		Cloche	
		Λ_{ES}	Λ_{RV}	Λ_{ES}	Λ_{RV}
2	Power	**0.33**	0.11	**0.18**	0.06
	TP	**0.651**	0.309	**0.400**	0.071
	FP	0.046	**0.012**	0.042	**0.002**
2.5	Power	**0.70**	0.19	**0.44**	0.11
	TP	**0.662**	0.574	**0.562**	0.443
	FP	0.044	**0.039**	0.032	**0.018**
3	Power	**0.92**	0.44	**0.69**	0.29
	TP	0.745	**0.755**	**0.652**	0.587
	FP	0.049	**0.038**	0.034	**0.023**

On retrouve le même type de résultats que dans le cadre temporel, ce qui confirme l'efficacité de l'indice de concentration basé sur les espacements même en l'absence de justification théorique.

Conclusion

Au vu des résultats obtenus sur l'étude de simulation, on peut légitimement penser que le fait que l'indice de concentration basé sur les espacements respecte la propriété P1 dans le cadre temporel, contrairement à celui basé sur le rapport de vraisemblance, est tout sauf anodin. L'indice de concentration I_{RV} accorde aux agrégats potentiels de petite taille une influence plus grande, ce qui porte à conséquence sur les résultats de la méthode de balayage associée. Il est à noter que NAGARWALLA 1996, lorsqu'il introduit la méthode de balayage basée sur le rapport de vraisemblance dans le cadre temporel, préconise de ne prendre en compte comme agrégats potentiels que les intervalles contenant un nombre minimum d'événements n_0 (il propose de choisir $n_0 = 5$). Cette recommandation est une manière d'atténuer le problème évoqué ci-dessus, mais se heurte au choix arbitraire de la valeur de n_0.

2 Processus marqué par une variable réelle

Comme nous l'avons dit dans l'Introduction, on s'intéresse ici au cas où une variable aléatoire X est associée à chaque événement du processus ponctuel. A partir du jeu de données $\{(s_i, x_i) : i \in [\![1, n]\!]\}$, l'indice de concentration $I(Z)$ doit nous permettre de déterminer dans quel agrégat potentiel $Z \in \mathcal{C}$ les observations de X sont les plus atypiques par rapport à l'ensemble des observations de X dans W. Dans ce chapitre, nous nous restreindrons au cas où la variable aléatoire X est réelle. Nous verrons dans le chapitre suivant des indices de concentration adaptés aux cas multivarié et fonctionnel.

Rappelons que, comme nous l'avons dit dans le Chapitre 1, les observations $x_1, \cdots, x_n$ peuvent être à l'origine associées à des unités administratives $W_1, \cdots, W_K$: les données sont dites groupées. Le jeu de données $\{(s_i, x_i) : i \in [\![1, n]\!]\}$ est alors obtenu en remplaçant chaque unité administrative W_k par une localisation centrale $c_k \in W$.

2.1 Des indices de concentration basés sur la vraisemblance

L'idée générale est sensiblement la même que dans le cas non marqué :

1. On considére un modèle paramétrique $\mathcal{M}_0$ considérant des marques indépendantes et identiquement distribuées dans W.
2. On introduit, pour chaque agrégat potentiel $Z \in \mathcal{C}$, un modèle paramétrique $\mathcal{M}_{1,Z}$ considérant des marques indépendantes mais différemment distribuées dans Z et dans Z^c, traduisant 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

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

Construire un indice de concentration basé sur ce rapport de vraisemblance s'avére totalement pertinent puisque ce dernier sera d'autant plus élevé que le modèle $\mathcal{M}_{1,Z}$ sera vraisemblable. Le choix des modèles $\mathcal{M}_0$ et $\mathcal{M}_{1,Z}$ va bien sûr dépendre de la nature de la variable X. Nous allons maintenant préciser les modèles paramétriques proposées par Kulldorff 1997 et Kulldorff, Huang et Konty 2009 pour deux cas de figure, calculer les rapports de vraisemblance correspondants et construire les indices de concentration associés.

Le modèle de Bernoulli

De nombreux processus ponctuels sont associés à des marques de type binaire. C'est notamment le cas en épidémiologie lorsque l'on observe des individus sur un intervalle de temps et/ou dans une zone géographique qui sont soit porteurs d'une certaine affection, soit sains. Dans ce cas-là, il semble naturel, comme l'a proposé Kulldorff 1997, d'utiliser la loi de Bernoulli. On a donc :

$$\mathcal{M}_0 : X_1, \cdots, X_n \text{ i.i.d. } \sim \mathcal{B}(p)$$

et la vraisemblance sous ce modèle est donnée par

$$L_0\big((s_1, x_1), \cdots, (s_n, x_n); p\big) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Notons, pour tout $Z \subset W$,

$$X(Z) = \sum_{i=1}^{n} X_i \mathbb{1}_Z(s_i) \quad \text{et} \quad \bar{X}(Z) = \frac{X(Z)}{n(Z)}$$

respectivement la somme et la moyenne des marques dans Z. Dans l'exemple épidémiologique donné précédemment, cela représente respectivement le nombre de contaminations et le taux moyen de contamination dans la zone Z. La vraisemblance maximale est obtenue lorsque $p = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}(W)$ et son logarithme vaut

$$\log(L_0^*) = x(W)\log\big(\bar{x}(W)\big) + \big(n - x(W)\big)\log\big(1 - \bar{x}(W)\big).$$

Le modèle considérant un agrégat dans Z est donné par :

$$\mathcal{M}_{1,Z} : X_1, \cdots, X_n \text{ indépendantes et } \begin{cases} X_i \sim \mathcal{B}(p_Z) & \text{si } s_i \in Z, \\ X_i \sim \mathcal{B}(p_{Z^c}) & \text{si } s_i \in Z^c, \end{cases}$$

et la vraisemblance sous ce modèle est donnée par

$$L_{1,Z}\big((s_1,x_1),\cdots,(s_n,x_n); p_Z, p_{Z^c}\big) = p_Z^{x(Z)}(1-p_Z)^{n(Z)-x(Z)} p_{Z^c}^{x(Z^c)}(1-p_{Z^c})^{n(Z^c)-x(Z^c)}.$$

La vraisemblance maximale est obtenue lorsque $p_Z = \bar{x}(Z)$ et $p_{Z^c} = \bar{x}(Z^c)$ et son logarithme vaut

$$\begin{aligned} \log(L_{1,Z}^*) &= x(Z)\log\big(\bar{x}(Z)\big) + \big(n(Z) - x(Z)\big)\log\big(1-\bar{x}(Z)\big) \\ &+ x(Z^c)\log\big(\bar{x}(Z^c)\big) + \big(n(Z^c) - x(Z^c)\big)\log\big(1-\bar{x}(Z^c)\big). \end{aligned}$$

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

$$\begin{aligned} \log\big(RV_B(Z)\big) &= \log(L_{1,Z}^*) - \log(L_0^*) \\ &= x(Z)\log\big(\bar{x}(Z)\big) + \big(n(Z) - x(Z)\big)\log\big(1-\bar{x}(Z)\big) \\ &+ x(Z^c)\log\big(\bar{x}(Z^c)\big) + \big(n(Z^c) - x(Z^c)\big)\log\big(1-\bar{x}(Z^c)\big) \\ &- x(W)\log\big(\bar{x}(W)\big) - \big(n - x(W)\big)\log\big(1-\bar{x}(W)\big). \end{aligned}$$

Le rapport de vraisemblance étant toujours supérieur à 1, l'indice de concentration

$$I_B(Z) = \log\big(RV_B(Z)\big)$$

sera toujours positif. Il est à noter que la maximisation de cet indice de concentration permettra d'identifier la zone dans laquelle la moyenne de X est la plus significativement différente : il peut s'agir d'un "agrégat positif" (moyenne de X plus élevée qu'ailleurs) ou d'un "agrégat négatif" (moyenne de X moins élevée qu'ailleurs). Si l'on n'est intéressé que par la détection d'agrégats positifs, il sera plus pertinent d'utiliser l'indice de concentration

$$I_B^+(Z) = I_B(Z)\mathbb{1}\big(\bar{x}(Z) > \bar{x}(Z^c)\big).$$

Le modèle Gaussien

Il arrive que les marques de type binaire associées aux localisations soient en fait issues d'une variable continue X. Par exemple, si l'on cherche à identifier des zones géographiques où le taux de nouveau-nés en déficit pondéral est plus élevé qu'ailleurs, il doit être plus pertinent de s'intéresser directement au poids à la naissance. De nombreuses lois de probabilité peuvent être utilisées pour modéliser la variable X et chacun de ces choix donnera naissance à une nouvelle statistique de balayage. Ici, nous allons nous concentrer sur le modèle Gaussien introduit par KULLDORFF, HUANG et KONTY 2009. On a donc :

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

et la log-vraisemblance sous ce modèle est donnée par

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

Notons, pour tout $Z \subset W$,

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

la moyenne du carré des marques dans Z. La log-vraisemblance maximale est obtenue lorsque $m = \bar{x}(W)$ et $\sigma^2 = \overline{x^2}(W) - \big(\bar{x}(W)\big)^2 := \sigma^{2*}$ et elle vaut

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

Le modèle considérant un agrégat dans Z suppose que les marques ont même variance mais que l'espérance diffère suivant la localisation associée. Il est donné par :

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

et la log-vraisemblance sous ce modèle est

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

La log-vraisemblance maximale est obtenue lorsque $m_Z = \bar{x}(Z)$, $m_{Z^c} = \bar{x}(Z^c)$ et

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

Elle 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 donc :

$$\log\big(RV_G(Z)\big) = \log(L_{1,Z}^*) - \log(L_0^*) = -\frac{n}{2}\big(\log(\sigma^{2*}_{Z,Z^c}) - \log(\sigma^{2*})\big).$$

On utilisera donc l'indice de concentration

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

qui est toujours positif. Ici encore, cet indice de concentration détecte à la fois des agrégats positifs et des agrégats négatifs. Si l'on n'est intéressé que par la détection d'agrégats positifs, il sera plus pertinent d'utiliser l'indice de concentration

$$I_G^+(Z) = I_G(Z)\mathbb{1}\big(\bar{x}(Z) > \bar{x}(Z^c)\big).$$

2.2 Des indices de concentration non-paramétriques

Bien que ces indices de concentration basés sur la vraisemblance soient parfaitement qualifiés pour la recherche d'agrégats, on peut se demander si, comme dans le cas des processus non marqués, il n'existerait pas des choix plus pertinents. C'est dans cette optique que j'ai introduit deux indices de concentration qui ne s'appuient sur aucun modèle paramétrique, mais uniquement construits à partir de l'hypothèse nulle d'absence d'agrégats $H_0 : X_1, \cdots, X_n$ i.i.d.

Un indice basé sur les moments

Plaçons-nous sous l'hypothèse H_0 et notons

$$\forall i \in [\![1, n]\!], \quad \mathbb{E}(X_i) = m \quad \text{et} \quad \mathbb{V}(X_i) = \sigma^2.$$

De nombreux tests de comparaison de deux populations, comme le test de Student par exemple (Saporta 2011), consistent à tester l'égalité des espérances en s'appuyant sur la différence entre les moyennes empiriques. Il paraît donc logique, pour comparer les marques dans Z et dans Z^c, de considérer l'écart entre les moyennes

$$D(Z) = \bar{X}(Z) - \bar{X}(Z^c).$$

L'espérance de $D(Z)$ est nulle mais sa variance dépend de $n(Z)$ et $n(Z^c)$ et il faut donc standardiser cet écart. Par indépendance des X_i, on a

$$\mathbb{V}\big(D(Z)\big) = \mathbb{V}\big(\bar{X}(Z)\big) + \mathbb{V}\big(\bar{X}(Z^c)\big) = \sigma^2\Big(\frac{1}{n(Z)} + \frac{1}{n(Z^c)}\Big) = \frac{n(Z) + n(Z^c)}{n(Z)n(Z^c)}\sigma^2 = \frac{n}{n(Z)n(Z^c)}\sigma^2.$$

Pour détecter des agrégats positifs, on peut donc utiliser l'indice de concentration basé sur les moments

$$I_M^+(Z) = \frac{\sqrt{n(Z)n(Z^c)}}{\sqrt{n}}\big(\bar{X}(Z) - \bar{X}(Z^c)\big)$$

qui est d'espérance nulle et de variance égale à σ^2 sous H_0. Puisque cela ne dépend pas de $n(Z)$, cet indice de concentration est pertinent pour comparer des agrégats potentiels comportant des nombres de marques quelconques. Si l'on cherche à détecter à la fois des agrégats positifs et négatifs, on pourra utiliser l'indice

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

Remarquons également que, si l'on souhaite tester l'égalité de l'espérance de X dans les zones Z et Z^c à l'aide d'un test de Student, dans le cas où la variance commune σ^2 est connue, on calcule la statistique

$$T_{Z|Z^c} = \frac{\bar{X}(Z) - \bar{X}(Z^c)}{\sqrt{\hat{\sigma}^2_{Z,Z^c}\left(\frac{1}{n(Z)} + \frac{1}{n(Z^c)}\right)}} = \frac{I_M^+(Z)}{\sqrt{\sigma^2}}.$$

La maximisation de l'indice de concentration $I_M(Z)$ est donc équivalente à la recherche de la statistique $T_{Z|Z^c}$ la plus significative (i.e. la plus éloignée de 0).

Un indice basé sur les rangs

Une autre manière de comparer deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$ est de tester non pas l'égalité des espérances mais que, parmi deux observations indépendantes issues chacune d'une population, la plus grande n'a pas plus de chances de provenir de $\mathcal{P}_1$ que de $\mathcal{P}_2$. Le test le plus utilisé pour cela est celui dit de Wilcoxon-Mann-Whitney car il a été introduit sous deux formes différentes par Wilcoxon 1945 et Mann et Whitney 1947. Suivant la méthode de Wilcoxon, nous écrirons la statistique correspondante en utilisant les rangs associés à chaque marque. Commençons par ordonner les marques par ordre croissant :

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

On notera $R_1, \cdots, R_n$ les rangs associés respectivement aux marques $X_1, \cdots, X_n$, c'est-à-dire tels que

$$\forall i \in [\![1, n]\!], \quad X_{\left(R(i)\right)} = X_i.$$

Notons qu'il existe une manière de gérer les ex-aequo que nous ne décrirons pas ici. La statistique à laquelle s'intéresse Wilcoxon pour comparer les marques dans Z et dans Z^c est la somme des rangs dans Z,

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

et la distribution de cette statistique sous H_0 est tabulée pour des petites valeurs de $n(Z)$ et $n(Z^c)$. Néanmoins, pour des raisons de simplicité de calcul, nous préférons suivre la même voie que ce qui est fait pour de grandes valeurs de $n(Z)$ et $n(Z^c)$, à savoir standardiser cette somme des rangs. Sous l'hypothèse H_0, ces rangs sont tirés indépendamment selon un tirage uniforme sans remise dans $\{1, \cdots, n\}$, ce qui permet d'obtenir

$$\mathbb{E}_0\big(SR(Z)\big) = \frac{n(Z)(n+1)}{2} \text{ et } \mathbb{V}_0\big(SR(Z)\big) = \frac{n(Z)n(Z^c)(n+1)}{12}.$$

Pour détecter des agrégats positifs, on peut donc utiliser l'indice de concentration basé sur les rangs

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

qui est d'espérance nulle et de variance unitaire sous H_0. Il est donc lui aussi pertinent pour comparer des agrégats potentiels comportant des nombres de marques quelconques. Si l'on cherche à détecter à la fois des agrégats positifs et négatifs, on pourra utiliser l'indice

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

La maximisation de cet indice correspond à la recherche de la statistique de Wilcoxon $SR(Z)$ la plus significative.

2.3 Comparaisons

Qualités des différents indices de concentration

Nous commençons par montrer que l'indice de concentration basé sur les moments que nous avons introduit est en fait équivalent à l'indice de concentration basé sur le rapport de vraisemblance dans le modèle Gaussien, c'est-à-dire que

$$\forall Z \subset W, \forall Z' \subset W, \quad I_M(Z) > I_M(Z') \Leftrightarrow I_G(Z) > I_G(Z').$$

En effet, on a

$$\begin{aligned}
\big(I_M(Z)\big)^2 &= \frac{n(Z)n(Z^c)}{n}\big(\bar{X}(Z) - \bar{X}(Z^c)\big)^2 \\
&= \frac{n(Z)n(Z^c)}{n}\Big(\frac{X(Z)}{n(Z)} - \frac{X(W) - X(Z)}{n(Z^c)}\Big)^2 \\
&= \frac{n(Z)n(Z^c)}{n}\Big(\frac{n(Z^c)X(Z) - n(Z)\big(X(W) - X(Z)\big)}{n(Z)n(Z^c)}\Big)^2 \\
&= \frac{n(Z)n(Z^c)}{n}\Big(\frac{nX(Z) - n(Z)X(W)}{n(Z)n(Z^c)}\Big)^2 \\
&= \frac{n\Big(X(Z) - n(Z)\frac{X(W)}{n}\Big)^2}{n(Z)n(Z^c)}.
\end{aligned}$$

De son côté,

$$\begin{aligned}
-n\sigma^{2*}\exp\Big(-\frac{2}{n}I_G(Z)\Big) &= -n\sigma^{2*}_{Z,Z^c} \\
&= -n(Z)\Big(\overline{X^2}(Z)-\big(\bar{X}(Z)\big)^2\Big)-n(Z^c)\Big(\overline{X^2}(Z^c)-\big(\bar{X}(Z^c)\big)^2\Big) \\
&= \frac{\big(X(Z)\big)^2}{n(Z)}+\frac{\big(X(Z^c)\big)^2}{n(Z^c)}-n\overline{X^2}(W) \\
&= \frac{n(Z^c)\big(X(Z)\big)^2+n(Z)\big(X(W)-X(Z)\big)^2}{n(Z)n(Z^c)}-n\overline{X^2}(W) \\
&= \frac{n\big(X(Z)\big)^2+n(Z)\big(X(W)\big)^2-2n(Z)X(W)X(Z)}{n(Z)n(Z^c)}-n\overline{X^2}(W) \\
&= \frac{n\Big(X(Z)-n(Z)\frac{X(W)}{n}\Big)^2+n(Z)\big(X(W)\big)^2-\frac{n(Z)^2}{n}\big(X(W)\big)^2}{n(Z)n(Z^c)}-n\overline{X^2}(W) \\
&= \frac{n\Big(X(Z)-n(Z)\frac{X(W)}{n}\Big)^2+\frac{n(Z)n(Z^c)}{n}\big(X(W)\big)^2}{n(Z)n(Z^c)}-n\overline{X^2}(W) \\
&= \big(I_M(Z)\big)^2+\frac{\big(X(W)\big)^2}{n}-n\overline{X^2}(W) \\
&= \big(I_M(Z)\big)^2-n\Big(\overline{X^2}(W)-\big(\bar{X}(W)\big)^2\Big).
\end{aligned}$$

L'indice de concentration basé sur les moments n'est donc pas en soi une nouveauté mais son écriture, différente de l'indice de concentration basé sur le modèle Gaussien, nous pousse à l'utiliser également pour des marques dont la distribution est très éloignée des lois Gaussiennes, comme les marques binaires.

Justement, pour des marques binaires, lequel, de l'indice basé sur la vraisemblance I_B ou de l'indice basé sur les moments I_M, est le plus adéquat ? Avant de comparer leurs performances un peu plus tard sur des jeux de données simulées, interrogeons-nous sur leurs qualités respectives. La première constatation que l'on peut faire est qu'aucun des deux ne satisfait la propriété P1 énoncée précédemment. En effet, même sous l'hypothèse d'absence d'agrégats, la loi de $I_B(Z)$ ne peut pas être invariante selon Z car les ensembles de valeurs discrètes de $I_B(Z)$ dépendent de $n(Z)$. Il en est de même pour $I_M(Z)$.

Notre objectif, on le rappelle, à travers l'étude de cette propriété P1, est de vérifier qu'un indice de concentration donne des chances égales à chaque agrégat potentiel, quelle que soit sa taille. On peut alors introduire une autre propriété, moins contraignante que P1, notée P2 : "En l'absence d'agrégat, pour tout $Z \in \mathcal{C}$ et tout $Z' \in \mathcal{C}$, $\mathbb{P}\big(I(Z) > I(Z')\big) = \mathbb{P}\big(I(Z') > I(Z)\big)$." Si un indice satisfait la propriété P1, il satisfait aussi P2. Pour vérifier si les indices I_B et I_M satisfont P2, ou au moins s'en rapprochent, j'ai étudié un exemple dans lequel $n = 100$ marques binaires sont i.i.d. selon la loi de Bernoulli de paramètre $p = 0.2$. Les lois de probabilité de $I_B(Z)$ et $I_M(Z)$ sont alors calculables quelle que soit la valeur de $n(Z)$. La Figure 2.3 représente la valeur de $\mathbb{P}\big(I(Z) > I(Z')\big) - \mathbb{P}\big(I(Z') > I(Z)\big)$ lorsque $n(Z') = n(Z) + 1$, en fonction de $n(Z)$. On s'aperçoit que, lorsqu'on compare des agrégats potentiels de tailles "moyennes", les écarts de probabilité sont très proches

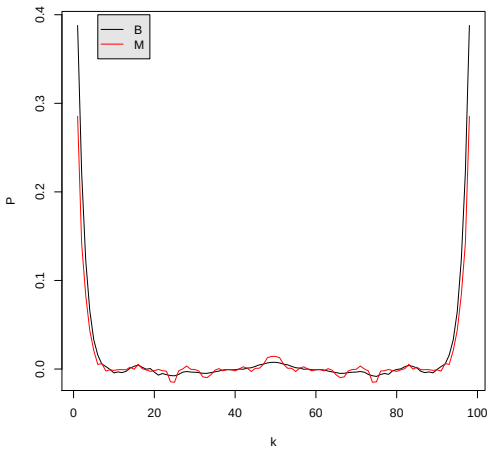

FIGURE 2.3 – Ecarts de probabilités.

de 0, donc ces agrégats potentiels sont considérés de manière équitable par les deux indices de concentration. Par contre, le problème se pose en présence d'agrégats de petite taille : les valeurs très largement positives prises par les écarts de probabilité lorsque $n(Z)$ est petit montrent que les petits agrégats potentiels ont tendance à être "avantagés" par rapport aux autres (comme c'était le cas pour l'indice I_{RV} dans le cadre non-marqué) et ce déséquilibre est encore plus marqué pour l'indice I_B que pour l'indice I_M.

Pour des marques continues, deux choix possibles sont l'indice basé sur les moments (équivalent à celui basé sur le rapport de vraisemblance dans le modèle Gaussien) et celui basé sur les rangs. Le premier est relié à l'utilisation du test de Student et le second à l'utilisation du test de Wilcoxon-Mann-Whitney. Comme indiqué par LEHAMNN 1999, pour tester si des observations issues de deux populations suivent une même loi, l'utilisation du premier est préconisée lorsque les données suivent une loi Gaussienne, et l'utilisation du second sinon. Pour la détection d'agrégats, on peut supposer qu'il en sera de même. En effet, sous l'hypothèse d'absence d'agrégats, l'indice de concentration basé sur les moments

$$I_M^+(Z) = \frac{\sqrt{n(Z)n(Z^c)}}{\sqrt{n}}\big(\bar{X}(Z) - \bar{X}(Z^c)\big)$$

suit la loi normale $\mathcal{N}(0, \sigma^2)$ à condition que les marques X_i soient elles-mêmes gaussiennes : dans ce cas de figure, cet indice de concentration satisfait la propriété P1 et semble donc parfaitemant adapté. A contrario, si les marques ne sont pas gaussiennes, alors que la loi de $I_M^+(Z)$ sera proche d'une loi normale pour des valeurs élevées de $n(Z)$ et $n(Z^c)$ (par application du théorème central limite), ce sera beaucoup moins le cas lorsque l'on s'intéresse à des agrégats de très petite taille ($n(Z)$ proche de 1) ou de très grande taille ($n(Z^c)$ proche de 1). Cet écart de loi, s'il est très marqué, risque d'être préjudiciable aux performances de la méthode de balayage basée sur l'indice I_M.

De son côté, sous l'hypothèse d'absence d'agrégats, la loi de l'indice de concen-

tration basé sur les rangs

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

dépend uniquement de $n(Z)$ et $n(Z^c)$ et nullement de la distribution des marques X_i. Ici encore, cette loi se rapproche de la Gaussienne centrée-réduite pour des valeurs élevées de $n(Z)$ et $n(Z^c)$, par application du théorème central limite, et c'est beaucoup moins le cas lorsque l'on s'intéresse à des agrégats de très petite taille ($n(Z)$ proche de 1) ou de très grande taille ($n(Z^c)$ proche de 1). Néanmoins, cet écart entre la loi exacte de la statistique de Wilcoxon-Mann-Whitney et la loi Gaussienne centrée-réduite, étudié par Lehamnn 1999, semble assez ténu même pour des petites valeurs de $n(Z)$ ou $n(Z^c)$.

Etude de simulation

On rappelle que, dans le cas d'un processus marqué, la détection d'agrégats ne prend pas en compte le caractère aléatoire des localisations géographiques des événements, mais uniquement celui des marques associées. Ainsi, pour les études de simulation suivantes, nous utiliserons comme localisations géographiques les coordonnées des chefs-lieux des 94 départements français métropolitains (hors Corse). Seules les marques associées seront générées aléatoirement. Le "vrai" agrégat que nous utiliserons, noté C, sera l'ensemble des 8 départements d'Île-de-France : Paris, Seine-et-Marne, Yvelines, Essonne, Hauts-de-Seine, Seine-Saint-Denis, Val-de-Marne et Val-d'Oise.

Comme dans le cadre non-marqué, l'efficacité d'une méthode de détection d'agrégat se mesurant à la fois par sa capacité à générer une alarme en cas d'agrégat et à identifier les contours de cet agrégat, nous calculons les critères suivants :

— la puissance, i.e. la proportion de jeux de données simulés pour lesquels la p-valeur est inférieure au niveau du test α.

— des taux de "vrai positifs" et de "faux positifs" mesurant la correspondance entre le "vrai agrégat" C et l'agrégat le plus probable $\hat{C}$, donnés par

$$VP = \frac{n(C \cap \hat{C})}{n(C)} \text{ et } FP = \frac{n(C^c \cap \hat{C})}{n(C^c)}.$$

Marques binaires Nous simulons 100 marques binaires pour chacun des 94 départements français, soit un total de $n = 9400$ marques indépendantes, mais toujours avec l'agrégat C précisé ci-dessus. La marque X_i associée à la localisation géographique s_i suit la loi suivante :

$$\begin{cases} X_i \sim \mathcal{B}(rp) & \text{si } s_i \in C, \\ X_i \sim \mathcal{B}(p) & \text{si } s_i \in C^c. \end{cases}$$

Dans ce cas de figure, le paramètre r est le ratio entre l'espérance des marques dans l'agrégat C et l'espérance des marques hors de l'agrégat : il mesure donc l'intensité de l'agrégat.

Nous avons simulé 100 jeux de données selon chaque fonction densité. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 99$ permutations aléatoires des marques, comme détaillé dans l'introduction de ce mémoire. La table 2.3 donne les résultats obtenus pour les méthodes de balayage Λ_B et Λ_M, utilisant respectivement les indices de concentration I_B et I_M.

TABLE 2.3 – Tests appliqués à des marques binaires.

r		Résultats :	
		Λ_B	Λ_M
1.1	Power	0.07	**0.08**
	TP	0.875	**0.891**
	FP	0.272	**0.241**
1.2	Power	0.26	**0.30**
	TP	**0.913**	0.908
	FP	0.198	**0.152**
1.3	Power	**0.59**	**0.59**
	TP	**0.858**	0.809
	FP	0.077	**0.048**
1.4	Power	**0.95**	0.94
	TP	**0.857**	0.835
	FP	0.026	**0.025**
1.5	Power	0.98	**0.99**
	TP	**0.852**	0.836
	FP	0.017	**0.015**

Comme attendu, les résultats s'améliorent lorsque l'intensité de l'agrégat augmente. On se rend compte que les deux méthodes obtiennent ici des résultats très similaires, notamment les puissances qui sont quasiment identiques. Néanmoins, la méthode basée sur le rapport de vraisemblance dans le modèle de Bernoulli tend à sélectionner des agrégats $\hat{C}$ de taille plus élevée, ce qui permet d'obtenir des indices de vrais positifs parfois légèrement meilleurs mais aussi des indices de faux positifs toujours plus mauvais, surtout lorsque l'intensité de l'agrégat est modérée.

Marques continues Nous simulons une marque continue indépendamment pour chacun des 94 départements français, toujours avec l'agrégat C précisé ci-dessus, selon trois modèles différents. Dans le modèle Gaussien, la marque X_i associée à la localisation géographique s_i suit la loi suivante :

$$\begin{cases} X_i \sim \mathcal{N}(r, 1) & \text{si } s_i \in C, \\ X_i \sim \mathcal{N}(0, 1) & \text{si } s_i \in C^c. \end{cases}$$

Dans le modèle logistique, la marque X_i associée à la localisation géographique s_i suit la loi suivante :

$$\begin{cases} X_i \sim \text{logistique}(r, 1) & \text{si } s_i \in C, \\ X_i \sim \text{logistique}(0, 1) & \text{si } s_i \in C^c. \end{cases}$$

Dans le modèle de Cauchy, la marque X_i associée à la localisation géographique s_i suit la loi suivante :

$$\begin{cases} X_i \sim \text{Cauchy}(r, 1) & \text{si } s_i \in C, \\ X_i \sim \text{Cauchy}(0, 1) & \text{si } s_i \in C^c. \end{cases}$$

Quel que soit le modèle, le paramètre r est l'écart entre l'espérance (ou la médiane dans le modèle de Cauchy, puisque l'espérance n'est pas définie) des marques dans l'agrégat C et celle des marques hors de l'agrégat : il mesure donc toujours l'intensité de l'agrégat.

Nous avons simulé 100 jeux de données selon chaque fonction densité. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 99$ permutations aléatoires des marques. Pour des marques gaussiennes, la table 2.4 donne les résultats obtenus pour les méthodes de balayage Λ_M et Λ_R, utilisant respectivement les indices de concentration I_M et I_R. Pour des marques logistiques et de Cauchy, les tables 2.5 et 2.6 donnent les résultats obtenus.

TABLE 2.4 – Tests appliqués à des marques gaussiennes.

r		Résultats :	
		Λ_M	Λ_R
1.0	Power	**0.29**	0.25
	TP	0.853	**0.980**
	FP	**0.151**	0.334
1.5	Power	**0.65**	0.53
	TP	0.827	**0.962**
	FP	**0.055**	0.113
2.0	Power	**0.94**	0.83
	TP	0.848	**0.943**
	FP	**0.022**	0.053

TABLE 2.5 – Tests appliqués à des marques logistiques.

r		Résultats :	
		Λ_M	Λ_R
1.0	Power	**0.09**	**0.09**
	TP	0.444	**1.0**
	FP	**0.101**	0.376
2.0	Power	**0.41**	0.40
	TP	0.771	**0.975**
	FP	**0.076**	0.150
3.0	Power	**0.74**	**0.74**
	TP	0.811	**0.959**
	FP	**0.056**	0.130

TABLE 2.6 – Tests appliqués à des marques de Cauchy.

r		Résultats :	
		Λ_M	Λ_R
3.0	Power	0.32	**0.35**
	TP	0.180	**0.954**
	FP	**0.061**	0.083
4.0	Power	0.49	**0.61**
	TP	0.168	**0.957**
	FP	**0.041**	0.075
5.0	Power	0.49	**0.73**
	TP	0.166	**0.923**
	FP	**0.040**	0.083

On observe ici une différence notable entre les deux méthodes : quel que soit le modèle des marques et l'intensité de l'agrégat, le taux de vrais positifs est toujours supérieur pour la méthode basée sur les rangs et le taux de faux positifs toujours inférieur pour la méthode basée sur les moments. Cela traduit le fait que la méthode basée sur les rangs tend à considérer comme agrégats les plus significatifs des agrégats potentiels de plus grande taille que la méthode basée sur les moments. Autrement dit, l'indice de concentration basé sur les rangs donne moins de poids aux valeurs extrêmes. Lorsque les marques suivent une loi Gaussienne, cela peut être un handicap et c'est pour cela que la puissance associée à Λ_R est plus faible que celle associée à Λ_M. Par contre, lorsque la distribution des marques admet plus souvent des valeurs extrêmes, comme dans le modèle logistique ou encore plus celui de Cauchy, c'est la situation inverse qui se produit : la méthode basée sur Λ_R devient alors plus puissante que celle basée sur Λ_M.

2.4 Recherche d'agrégats de variance atypique

Jusqu'à présent, en présence d'un processus ponctuel marqué par une variable réelle, nous avons recherché des agrégats présentant des moyennes de marques différentes. Cependant, pour des marques continues, on peut être amené à rechercher également des agrégats présentant des variances significativement différentes. C'est notamment le cas lorsque l'on cherche à identifier les territoires où les inégalités de revenus sont les plus marquées, comme le font SHELNUTT et YAO 2005 parmi les comtés des États-Unis.

A partir du jeu de données $\{(s_i, x_i) : i \in [\![1, n]\!]\}$, nous allons construire deux indices de concentration $I(Z)$, l'un basé sur un rapport de vraisemblance et l'autre sur un test d'égalité des variances, nous permettant de déterminer dans quel agrégat potentiel $Z \in \mathcal{C}$ les observations de X sont de variance atypique par rapport à l'ensemble des observations de X dans W.

Un indice de concentration basé sur la vraisemblance

Dans un premier temps, nous construisons un indice de concentration basé sur un rapport de vraisemblance, de manière similaire à ce que nous avons décrit précédemment. Pour celà, nous choisissons d'utiliser un modèle Gaussien, mais en considérant, pour l'hypothèse alternative, des variances différentes à l'intérieur et à l'extérieur de l'agrégat. On a donc :

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

et la log-vraisemblance sous ce modèle est donnée par

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

Comme nous l'avons dit précédemment, la log-vraisemblance maximale est obtenue lorsque $m = \bar{x}(W)$ et $\sigma^2 = \sigma^{2*} = \overline{x^2}(W) - \big(\bar{x}(W)\big)^2$ et elle vaut

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

Le modèle considérant un agrégat dans Z suppose cette fois-ci que les marques ont même espérance mais que la variance diffère suivant la localisation associée. Il est donné par :

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

et la log-vraisemblance sous ce modèle est

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

Afin de maximiser cette log-vraisemblance, nous fixons tout d'abord la valeur de m_{Z,Z^c}. La valeur de σ_Z^2 maximisant

$$-\frac{n(Z)}{2}\log(\sigma_Z^2) - \frac{1}{2\sigma_Z^2}\sum_{i=1}^n \big((x_i - m_{Z,Z^c})^2 \mathbb{1}_Z(s_i)\big)$$

est alors donnée par

$$\frac{1}{n(Z)}\sum_{i=1}^n \big((x_i - m_{Z,Z^c})^2 \mathbb{1}_Z(s_i)\big) = \overline{x^2}(Z) - 2\bar{x}(Z)m_{Z,Z^c} + (m_{Z,Z^c})^2.$$

On obtient un résultat similaire pour $\sigma_{Z^c}^2$. En injectant ces expressions dans la log-vraisemblance, on obtient que la valeur de m_{Z,Z^c} qui maximise cette log-vraisemblance, que nous noterons m_{Z,Z^c}^*, est celle qui maximise

$$\begin{aligned} g(m_{Z,Z^c}) \;=\; & -\frac{n}{2}\log(2\pi) - \frac{n(Z)}{2}\log\big(\overline{x^2}(Z) - 2\bar{x}(Z)m_{Z,Z^c} + (m_{Z,Z^c})^2\big) \\ & -\frac{n(Z^c)}{2}\log\big(\overline{x^2}(Z^c) - 2\bar{x}(Z^c)m_{Z,Z^c} + (m_{Z,Z^c})^2\big) - \frac{n(Z)}{2} - \frac{n(Z^c)}{2}.\end{aligned}$$

La dérivée de cette fonction est donnée par

$$\begin{aligned} g'(m_{Z,Z^c}) &= -\frac{n(Z)}{2}\frac{-2\bar{x}(Z)+2m_{Z,Z^c}}{\overline{x^2}(Z)-2\bar{x}(Z)m_{Z,Z^c}+(m_{Z,Z^c})^2} \\ &\quad -\frac{n(Z^c)}{2}\frac{-2\bar{x}(Z^c)+2m_{Z,Z^c}}{\overline{x^2}(Z^c)-2\bar{x}(Z^c)m_{Z,Z^c}+(m_{Z,Z^c})^2} \\ &= \frac{n}{2}\frac{h(m_{Z,Z^c})}{\left(\overline{x^2}(Z)-2\bar{x}(Z)m_{Z,Z^c}+(m_{Z,Z^c})^2\right)\left(\overline{x^2}(Z^c)-2\bar{x}(Z^c)m_{Z,Z^c}+(m_{Z,Z^c})^2\right)} \end{aligned}$$

où

$$\begin{aligned} h(m_{Z,Z^c}) &= (m_{Z,Z^c})^3+\frac{-n(Z)\bar{x}(Z)-n(Z^c)\bar{x}(Z^c)-2n(Z)\bar{x}(Z^c)-2n(Z^c)\bar{x}(Z)}{n}(m_{Z,Z^c})^2 \\ &\quad +\frac{n(Z)\overline{x^2}(Z^c)+2n\bar{x}(Z)\bar{x}(Z^c)+n(Z^c)\overline{x^2}(Z)}{n}m_{Z,Z^c} \\ &\quad +\frac{-n(Z)\bar{x}(Z)\overline{x^2}(Z^c)-n(Z^c)\bar{x}(Z^c)\overline{x^2}(Z)}{n}. \end{aligned}$$

En suivant la méthode de Cardan (Jacobson 2009), on calcule les racines de cette fonction cubique. Ainsi, m^*_{Z,Z^c} est soit la seule racine (si les deux autres sont complexes) de la fonction $h(\cdot)$, soit celle des trois racines (si elles sont toutes réelles) qui maximise la fonction $g(\cdot)$. On obtient ensuite les estimateurs du maximum de vraisemblance pour les variances dans Z et dans Z^c :

$$\sigma^{2*}_Z := \overline{x^2}(Z)-2\bar{x}(Z)m^*_{Z,Z^c}+(m^*_{Z,Z^c})^2 \text{ et } \sigma^{2*}_{Z^c} := \overline{x^2}(Z^c)-2\bar{x}(Z^c)m^*_{Z,Z^c}+(m^*_{Z,Z^c})^2.$$

La log-vraisemblance maximale vaut alors

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

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

$$\log\left(RV_{Gvar}(Z)\right) = \log(L^*_{1,Z})-\log(L^*_0) = -\frac{n(Z)}{2}\log(\sigma^{2*}_Z)-\frac{n(Z^c)}{2}\log(\sigma^{2*}_{Z^c})+\frac{n}{2}\log(\sigma^{2*}).$$

On utilisera donc l'indice de concentration

$$I_{Gvar}(Z) = \log\left(RV_{Gvar}(Z)\right)$$

qui est toujours positif. Cet indice de concentration détecte à la fois des agrégats plus variables et moins variables que la normale. Si l'on n'est intéressé que par la détection d'agrégats plus variables, il sera plus pertinent d'utiliser l'indice de concentration

$$I^+_{Gvar}(Z) = I_{Gvar}(Z)\mathbb{1}\left(\overline{x^2}(Z)-\left(\bar{x}(Z)\right)^2 > \overline{x^2}(Z^c)-\left(\bar{x}(Z^c)\right)^2\right).$$

Un indice de concentration basé sur un test

Nous proposons maintenant de construire un autre indicateur de concentration qui s'appuie non pas sur la vraisemblance dans un modèle paramétrique donné, mais sur un test classique d'égalité des variances, appelé F-test (Saporta 2011). Pour

tester l'égalité des variances des marques dans Z et Z^c, ce test s'appuie sur le rapport des variances empiriques sans biais

$$R_{Z|Z^c} = \frac{S^2(Z)}{S^2(Z^c)}$$

où

$$S^2(Z) = \frac{1}{n(Z)-1} \sum_{i=1}^{n} \Big(\big(X_i - \bar{X}(Z)\big)^2 \mathbb{1}_Z(s_i) \Big).$$

Sous l'hypothèse que les marques X_i sont indépendantes et identiquement distribuées selon une loi Gaussienne, ce rapport $R_{Z|Z^c}$ suit une loi de Fisher à $n(Z)-1$ et $n(Z^c)-1$ degrés de liberté. Afin d'évaluer ce rapport quelle que soit la taille de l'agrégat considéré $n(Z)$, nous introduisons donc l'indice de concentration

$$I^+_{Tvar}(Z) = F_{n(Z)-1,n(Z^c)-1}(R_{Z|Z^c})$$

où $F_{d_1,d_2}(\cdot)$ représente la fonction de répartition de la loi de Fisher à d_1 et d_2 degrés de liberté. Sous l'hypothèse que les marques sont indépendantes et identiquement distribuées selon une loi Gaussienne, cet indicateur est distribué selon la loi uniforme sur $[0,1]$ et sera d'autant plus proche de 1 que la variance dans Z excèdera celle dans Z^c : il est donc parfaitement qualifié pour identifier des agrégats de variance significativement plus élevée. Si l'on cherche à détecter des agrégats de variance atypique (plus élevée ou plus faible que la normale), il sera préférable d'utiliser l'indice de concentration

$$I_{Tvar}(Z) = \max\big(I^+_{Tvar}(Z), I^+_{Tvar}(Z^c)\big).$$

Comparaisons

Qualités des différents indices de concentration Avant de comparer les performances de ces deux indices de concentration sur des jeux de données simulées, interrogeons-nous sur leurs qualités respectives. La loi de probalilité de l'indice $I_{Gvar}(Z)$ est particulièrement difficile à obtenir, même sous l'hypothèse que le modèle $\mathcal{M}_0$ est valide, car elle dépend de celle de m^*_{Z,Z^c}, racine d'une fonction cubique dont les coefficients sont aléatoires. Il nous est donc impossible de savoir si cette loi de probabilité dépend des valeurs de $n(Z)$ et $n(Z^c)$. Par contre, cet indice étant basé sur un rapport de vraisemblance, il tend vers une loi du chi-deux dont le degré de liberté ne dépend que du nombre de paramètres associé aux modèles $\mathcal{M}_0$ et $\mathcal{M}_{1,Z}$, et pas du tout de la valeur de $n(Z)$. A contrario, toujours sous l'hypothèse que le modèle $\mathcal{M}_0$ est valide, la loi de $I^+_{Tvar}(Z)$ est une loi uniforme sur $[0,1]$ donc celle de $I_{Tvar}(Z)$ ne dépend pas de la taille de l'agrégat potentiel $n(Z)$. On peut donc penser que cet indice basé sur un test statistique, contrairement à celui basé sur le rapport de vraisemblance, donnera autant de poids aux petits agrégats potentiels qu'aux grands.

Une étude de simulation Nous simulons une marque continue indépendamment pour chacun des 94 départements français, toujours avec le même agrégat C, selon

trois modèles différents. Dans le modèle Gaussien homogène, la marque X_i associée à la localisation géographique s_i suit la loi suivante :

$$\begin{cases} X_i \sim \mathcal{N}(0, r) & \text{si } s_i \in C, \\ X_i \sim \mathcal{N}(0, 1) & \text{si } s_i \in C^c. \end{cases}$$

Dans le modèle exponentiel, la marque $X_i = Y_i - \mathbb{E}(Y_i)$ associée à la localisation géographique s_i est dérivée de la variable aléatoire Y_i qui suit la loi suivante :

$$\begin{cases} Y_i \sim \mathcal{E}(r^{-1/2}) & \text{si } s_i \in C, \\ Y_i \sim \mathcal{E}(1) & \text{si } s_i \in C^c. \end{cases}$$

Dans le modèle Gaussien hétérogène, la marque X_i associée à la localisation géographique s_i suit la loi suivante :

$$\begin{cases} X_i \sim \mathcal{N}(l, r) & \text{si } s_i \in C, \\ X_i \sim \mathcal{N}(0, 1) & \text{si } s_i \in C^c. \end{cases}$$

Quel que soit le modèle, le paramètre r est l'écart entre la variance des marques dans l'agrégat C et celle des marques hors de l'agrégat : il mesure donc toujours l'intensité de l'agrégat. Pour le troisième modèle, le paramètre l mesure la différence d'espérance entre l'intérieur et l'extérieur de l'agrégat : nous l'appellerons le décalage.

Nous avons simulé 1000 jeux de données selon chaque fonction densité. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 999$ permutations aléatoires des marques. Les tables 2.7 et 2.8 donnent les résultats obtenus pour les méthodes de balayage Λ_{Gvar} et Λ_{Tvar}, utilisant respectivement les indices de concentration I_{Gvar} et I_{Tvar}.

TABLE 2.7 – Tests appliqués à des marques gaussiennes homogènes et exponentielles.

r		Gaussiennes		Exponentielles	
		Λ_{Gvar}	Λ_{Tvar}	Λ_{Gvar}	Λ_{Tvar}
1.5	Power	0.080	**0.118**	0.056	**0.057**
	TP	0.567	**0.741**	0.191	**0.352**
	FP	3.702	**3.590**	**1.441**	2.384
2.0	Power	0.265	**0.322**	0.070	**0.071**
	TP	1.723	**2.009**	0.349	**0.472**
	FP	3.998	**3.400**	**1.806**	2.195
2.5	Power	0.528	**0.553**	0.074	**0.093**
	TP	3.492	**3.500**	0.332	**0.659**
	FP	2.844	**2.118**	**1.538**	2.987

On observe ici des performances très différentes suivant le modèle de simulation. Sous le modèle Gaussien homogène, les puissances des deux méthodes augmentent rapidement avec l'intensité de l'agrégat r et la méthode basée sur le test donne des résultats légèrement meilleurs. On retrouve le fait que la méthode basée sur la vraisemblance tend à sélectionner des agrégats plus larges, ce qui augmente le taux

TABLE 2.8 – Tests appliqués à des marques Gaussiennes hétérogènes.

r	l		Résultats :	
			Λ_{Gvar}	Λ_{Tvar}
1.0	1.0	Power	**0.082**	0.075
		TP	**0.612**	0.504
		FP	5.499	**3.480**
1.5	1.0	Power	**0.168**	0.135
		TP	**1.196**	0.960
		FP	4.548	**4.236**
2.0	1.0	Power	**0.390**	0.345
		TP	**2.597**	2.317
		FP	4.822	**4.712**
2.5	1.0	Power	**0.603**	0.571
		TP	**3.970**	3.800
		FP	**2.618**	3.354

de faux positifs. Sous le modèle exponentiel, les deux méthodes ont beaucoup de mal à détecter la présence d'un agrégat, même lorsque la variance y est $2,5$ fois plus élevée. On peut penser que la construction d'une méthode de balayage basée sur un test moins sensible à la normalité, comme le test de Conover (CONOVER 1980) basé sur les rangs, pourrait être profitable dans ce cas-là. Enfin, le dernier tableau nous montre que, même si elles sont construites pour identifier des différences de variance, ces deux statistiques de balayage sont également affectées par les décalages d'espérance, mais Λ_{Gvar} l'est bien plus que Λ_{Tvar}.

Un exemple illustratif

Nous appliquons les deux méthodes de balayage à un jeu de données économiques provenant de l'Institut National de la Statistique et des Etudes Economiques. Dans chacun des 94 départements français (hors Corse et outre-mer), le revenu médian par habitant est illustré par la Figure 2.4. Nous observons clairement un agrégat de départements plus riches autour de Paris. Cependant, on peut aussi se demander s'il n'est pas possible d'exhiber une région où les inégalités entre départements seraient significatives, donc un agrégat de variance anormalement élevée.

Pour répondre à cette question, nous avons appliqué les méthodes basées sur Λ_{Gvar} et Λ_{Tvar} sur ce jeu de données, en utilisant l'ensemble d'agrégats potentiels circulaires décrit au Chapitre 1. Les p-valeurs sont obtenues à partir de 9999 permutations aléatoires des marques. Les agrégats les plus probables sont donnés par la Figure 2.5. L'agrégat le plus probable associé à Λ_{Gvar} correspond exactement à la région Île-de-France. Cela n'a rien de surprenant puisqu'elle contient 7 départements riches plus la Seine-Saint-Denis, l'un des départements les plus pauvres de France. Cet agrégat est très significatif puisque la p-valeur associée est 0.0001 : aucune des 9999 statistiques calculées sur les données permutées n'est plus grande que celle calculée sur les données observées ! L'agrégat le plus probable associé à Λ_{Tvar} est légèrement moins significatif puisque sa p-valeur associée est 0.0075. Il contient 7 départements d'Île-de-France mais aussi 6 départements voisins du Nord de la France, où les revenus médians sont assez faibles.

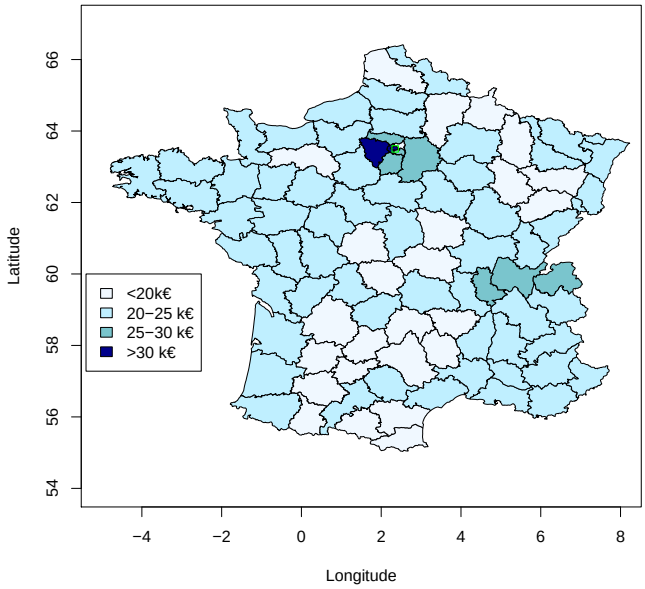

FIGURE 2.4 – Revenu médian

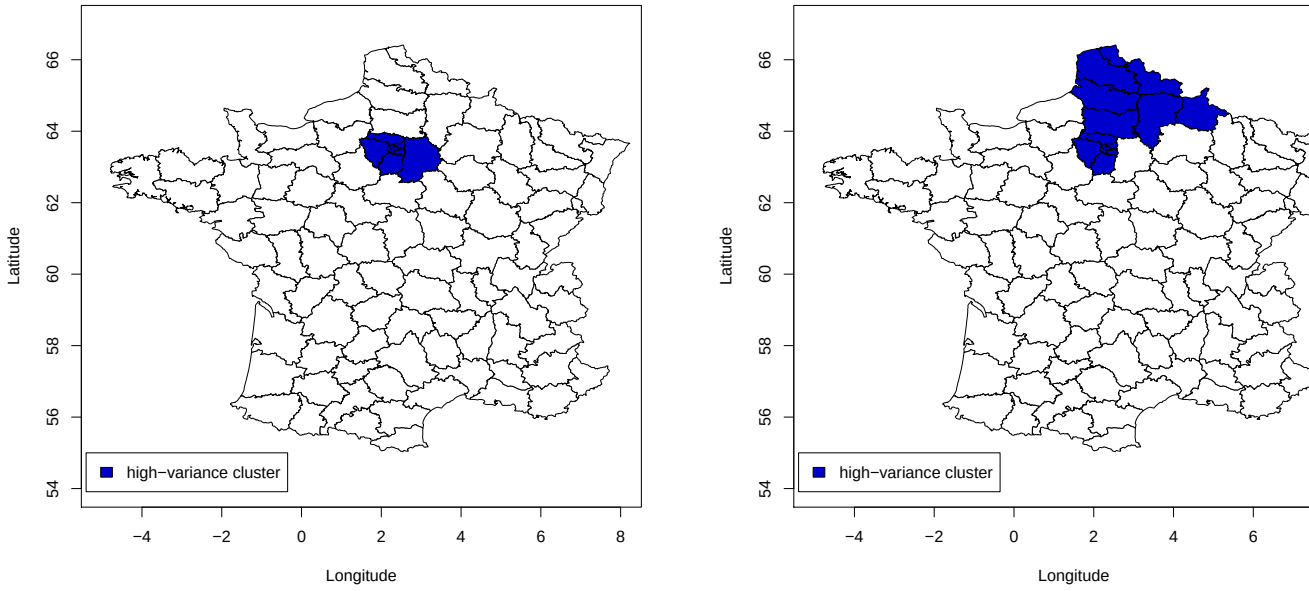

FIGURE 2.5 – Gauche : Agrégat le plus probable associé à Λ_{Gvar}. Droite : Agrégat le plus probable associé à Λ_{Tvar}.

Chapitre 3

Marques multivariées et fonctionnelles

Publications : [8], [9], [11].

Dans le Chapitre 2, j'ai décrit différentes manières de détecter des agrégats lorsque le processus ponctuel observé est marqué par une variable aléatoire réelle X. Dans ce chapitre 3, nous allons voir comment étendre ces méthodes aux cas où la variable X est multivariée, puis fonctionnelle. Le principe général restera le même, à savoir l'utilisation d'une statistique de balayage. L'adaptation principale se fera via l'indice de concentration qui devra correspondre au type de marques rencontrées. Dans une première partie, nous nous concentrerons sur le cas où X est un vecteur aléatoire de dimension finie et nous verrons comment construire un indice de concentration basé sur le rapport de vraisemblance dans un cadre Gaussien multivarié, comme dans [8], puis un indice de concentration non-paramétrique basé sur l'extension du test de Wilcoxon-Mann-Whitney au cas multivarié, comme dans [9]. Ensuite, comme nous l'avons fait dans [11], nous traiterons le cas où X est une fonction aléatoire en introduisant un indice de concentration non-paramétrique basé sur l'extension du test de Wilcoxon-Mann-Whitney au cas fonctionnel.

1 Marques multivariées

La genèse de cette partie du mémoire découle de ma rencontre avec Michaël Genin et Florent Occelli, de l'Université de Lille, dont l'une des problématiques était d'analyser un jeu de données environnementales recueillies dans l'agglomération lilloise, décrit en détail par Occelli, Bavdek et al. 2016 et Occelli, Cuny et al. 2014 et sur lequel nous reviendrons un peu plus tard. Pour résumer, le taux de divers polluants a été mesuré en plusieurs localisations spatiales. On constate une forte corrélation entre les taux des polluants, ainsi qu'une corrélation spatiale marquée. On pourrait envisager de modéliser ces données sous forme d'un champ spatial multivarié mais la gestion à la fois de la corrélation entre les variables et de la corrélation spatiale semble assez difficile. Notre objectif principal étant d'identifier la zone spatiale où la pollution globale peut être considérée la plus élevée, nous avons décidé de mettre en place une statistique de balayage spatial adapté à ces marques multivariées.

Dans cette section, on considèrera un vecteur aléatoire X, de dimension p, associé à chaque événement du processus ponctuel. A partir du jeu de données $\{(s_i, x_i) \ : \ i \ \in \ [\![1, n]\!]\}$, où $x_i = (x_i^1, \cdots, x_i^p) \in \mathbb{R}^p$, l'indice de concentration $I(Z)$ doit nous permettre de déterminer dans quel agrégat potentiel $Z \in \mathcal{C}$ les observations de X sont les plus atypiques par rapport à l'ensemble des observations de X dans W.

1.1 Un indice de concentration multivarié basé sur l'indépendance

Le premier indice que nous proposons utilise la méthode proposée par Kulldorff, Mostashari et al. 2007 pour combiner des indices de concentration associés à différentes variables. Notons $X^1, \cdots, X^p$ chacune des p variables aléatoires constituant le vecteur aléatoire X. Par abus de notation, nous noterons également $X^j = (x_1^j, \cdots, x_n^j)^T$ le vecteur des observations associées à cette variable, et $X = (X^1 | \cdots | X^p)$ la matrice qui rassemble ces p vecteurs.

Pour tout agrégat potentiel $Z \in \mathcal{C}$, nous noterons, pour tout $j \in [\![1, p]\!]$,

$$I_G^j(Z) = \log\left(RV_G^j(Z)\right)$$

l'indice de concentration basé sur le modèle Gaussien appliqué uniquement à la variable X^j. Nous ne rappellerons pas les détails du calcul, qui sont donnés dans le Chapitre 2. Si l'on considère l'indépendance entre les variables $X^1, \cdots, X^p$, la log-vraisemblance de l'ensemble de ces p variables devient la somme des log-vraisemblances individuelles et l'on peut donc définir un indice de concentration multivarié indépendant basé sur le modèle Gaussien :

$$I_{MIG}(Z) = \sum_{j=1}^{p} \log\left(RV_G^j(Z)\right) = \sum_{j=1}^{p} I_G^j(Z).$$

1.2 Un indice de concentration basé sur le modèle Gaussien multivarié

Bien que cette idée soit intéressante, il nous a paru dommage de ne pas utiliser du tout les informations sur les covariances entre les variables dont nous disposons. Pour remédier à cela, nous avons donc élaboré un indice de concentration généralisant l'indice basé sur le rapport de vraisemblance dans le modèle Gaussien de Kulldorff, Huang et Konty 2009 au cadre multivarié. Nous rappellons l'idée générale des indices de concentration basés sur un rapport de vraisemblance énoncée dans le Chapitre 2 :

1. On considère un modèle paramétrique $\mathcal{M}_0$ considérant des marques indépendantes et identiquement distribuées dans W.
2. On introduit, pour chaque agrégat potentiel $Z \in \mathcal{C}$, un modèle paramétrique $\mathcal{M}_{1,Z}$ considérant des marques indépendantes mais différemment distribuées dans Z et dans Z^c, traduisant 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

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

Voici le détail des calculs lorsque l'on se place dans le cadre Gaussien multivarié. Le modèle traduisant l'absence d'agrégat est :

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

où $\mathcal{N}_p(m, \Sigma)$ représente la loi Gaussienne p-variée de vecteur espérance $m \in \mathbb{R}^p$ et de matrice de variance-covariance $\Sigma \in \mathcal{M}_p(\mathbb{R})$. La log-vraisemblance sous ce modèle est donnée par

$$\begin{aligned} \log\Big(L_0\big((s_1,x_1),\cdots,(s_n,x_n);m,\Sigma\big)\Big) &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) \\ &- \frac{1}{2}\sum_{i=1}^{n}(x_i-m)^T\Sigma^{-1}(x_i-m), \end{aligned}$$

où $|\Sigma|$ représente le déterminant de la matrice Σ. Notons, pour tout $Z \subset W$,

$$\left\{ \begin{array}{rcl} n(Z) &=& \displaystyle\sum_{i=1}^{n} \mathbb{1}_Z(s_i) \\ \bar{X}(Z) &=& \displaystyle\frac{1}{n(Z)}\sum_{i=1}^{n} X_i \mathbb{1}_Z(s_i) \\ S(Z) &=& \displaystyle\frac{1}{n(Z)}\sum_{i=1}^{n}\big(X_i-\bar{X}(Z)\big)\big(X_i-\bar{X}(Z)\big)^T\mathbb{1}_Z(s_i) \end{array} \right.$$

respectivement le nombre d'événements, le vecteur moyen et la matrice de variance-covariance empirique dans Z. Comme indiqué par Mardia, Kent et Bibby 1979, la log-vraisemblance maximale est obtenue lorsque $m = \bar{x}(W) := m^*$ et $\Sigma = S(W) := \Sigma^*$

et elle vaut

$$\begin{aligned}
\log(L_0^*) &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - m^*)^T(\Sigma^*)^{-1}(x_i - m^*) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left((x_i - m^*)^T(\Sigma^*)^{-1}(x_i - m^*)\right) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left((x_i - m^*)(x_i - m^*)^T(\Sigma^*)^{-1}\right) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{1}{2}\mathrm{Tr}\left(\sum_{i=1}^{n}(x_i - m^*)(x_i - m^*)^T(\Sigma^*)^{-1}\right) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{1}{2}\mathrm{Tr}\left(n\Sigma^*(\Sigma^*)^{-1}\right) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{n}{2}\mathrm{Tr}(I_p) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma^*|) - \frac{np}{2},
\end{aligned}$$

où $\mathrm{Tr}(A)$ représente la trace de la matrice A. Le modèle considérant un agrégat dans Z suppose que les marques ont même matrice de variance-covariance mais que le vecteur espérance diffère suivant la localisation associée. Il est donné par :

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

et la log-vraisemblance sous ce modèle est

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

La log-vraisemblance maximale est obtenue lorsque $m_Z = \bar{x}(Z) := m_Z^*, m_{Z^c} = \bar{x}(Z^c) := m_{Z^c}^*$ et

$$\Sigma_{Z,Z^c} = \frac{n(Z)S(Z) + n(Z^c)S(Z^c)}{n} := \Sigma_{Z,Z^c}^*.$$

Elle vaut

$$\log(L_{1,Z}^*) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma_{Z,Z^c}^*|) - \frac{np}{2}.$$

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

$$\log\big(RV_{MG}(Z)\big) = \log(L^*_{1,Z}) - \log(L^*_0) = -\frac{n}{2}\big(\log(|\Sigma^*_{Z,Z^c}|) - \log(|\Sigma^*|)\big).$$

On utilisera donc l'indice de concentration

$$I_{MG}(Z) = \log\big(RV_{MG}(Z)\big)$$

qui est toujours positif. Notons que, comme l'a montré Anderson 2003, le rapport de vraisemblance utilisé est équivalent à la statistique de Hotelling, introduite par Hotelling 1931 pour généraliser le test de Student au cas multivarié. La maximisation de cet indice de concentration est donc équivalente à la recherche de la statistique de Hotelling la plus significative.

Ici, dans le cadre multivarié, la notion d'agrégats positifs et négatifs n'est pas aussi claire que dans le cadre univarié. L'indice de concentration que nous venons de définir permettra de détecter des agrégats dans lesquels certaines des p variables observées seront en excès, et les autres en défaut. Si on veut s'assurer de ne détecter que des agrégats dans lesquels toutes les p variables sont en excès, on pourra utiliser l'indice de concentration

$$I^+_{MG}(Z) = I_{MG}(Z)\prod_{j=1}^{p}\mathbb{1}\big(\bar{x}^j(Z) > \bar{x}^j(Z^c)\big).$$

Il est à noter toutefois que l'existence d'un tel agrégat n'est nullement assurée, surtout si certaines variables sont fortement corrélées négativement entre elles.

1.3 Un indice de concentration non-paramétrique

Bien que cet indice de concentration basé sur la vraisemblance soit parfaitement qualifié pour la recherche d'agrégats, on peut se demander si, comme dans le cas des marques réelles, il n'existerait pas de choix plus pertinent ne nécessitant pas de spécifier la distribution des marques. Nous avons choisi de généraliser l'indice de concentration issu du test de Wilcoxon-Mann-Whitney au cadre multivarié. Pour comprendre cette démarche, nous exposons d'abord comment la notion de rangs et les tests associés ont été généralisés au cadre multivarié.

Rangs multivariés

Alors que la notion de rang est naturelle pour les variables aléatoires réelles, il n'en est pas de même lorsque l'on s'intéresse à des vecteurs aléatoires puisque l'ordonnancement de ces vecteurs dans $\mathbb{R}^p$ n'est généralement pas possible. Néanmoins, l'extension de la notion de rangs à des données multivariées a été investiguée par de nombreux auteurs, en utilisant des notions d'interdirections (Randles 1989) ou encore de profondeur (Zuo et Serfling 2000). Nous nous focalisons sur l'approche proposée par Oja et Randles 2004. Leur définition de rangs multivariés associés aux marques $X_1, \cdots, X_n$ appartenant à $\mathbb{R}^p$ découle de la définition d'une fonction de signe spatial notée

$$S(x) = \begin{cases} ||x||^{-1}x & \text{si } x \neq 0, \\ 0 & \text{si } x = 0, \end{cases}$$

pour tout $x \in \mathbb{R}^p$, $||x||$ désignant la norme L_2. L'image de cette fonction est une direction (un point sur la p-sphère unité $\mathbb{S}^p$). Les signes multivariés associés à $X_1, \cdots, X_n$ sont

$$\forall i \in [\![1, n]\!], \quad S_i = S(A_x X_i),$$

où la matrice A_x, appelée matrice de transformation de Tyler (Tyler 1987), assure que la matrice de variance-covariance des signes multivariés est égale à $\frac{1}{p} I_p$, c'est-à-dire la matrice de variance-covariance de la distribution uniforme sur $\mathbb{S}^p$. Notons que la matrice de transformation de Tyler s'obtient facilement à l'aide d'une procédure itérative. Nous pouvons maintenant définir les rangs multivariés

$$\tilde{R}_i = \frac{1}{n} \sum_{j=1}^{n} S_{i,j}$$

où les $S_{i,j} = S\big(A_x(X_i - X_j)\big)$, avec $(i, j) \in [\![1, n]\!]^2$, sont les signes des différences transformées, obtenues encore via la transformation de Tyler. On peut remarquer que l'utilisation de cette transformation de Tyler permet aux rangs multivariés ainsi définis d'être totalement indépendants du système de coordonnées utilisé dans $\mathbb{R}^p$. Notons également que la notation $\tilde{R}_i$ est différente de la notation R_i utilisée dans le chapitre 2 pour définir les rangs univariés : en effet, lorsque $p = 1$, les rangs multivariés ne correspondent pas tout à fait aux rangs univariés, mais à leur version centrée (dont la somme est nulle) multipliée par $\frac{2}{n}$ et on a alors

$$\tilde{R}_i = \frac{2}{n}\Big(R_i - \frac{n+1}{2}\Big).$$

Version multivariée du test de Wilcoxon-Mann-Whitney

A partir de ces rangs multivariés $(\tilde{R}_1, \cdots, \tilde{R}_n) \in (\mathbb{R}^p)^n$, il est possible d'étendre au cas multivarié de nombreux tests (Kendall, Spearman, Mood...), et notamment celui de Wilcoxon-Mann-Whitney, évoqué dans le Chapitre 2. Nous cherchons à tester l'hypothèse H_0, correspondant à l'absence d'agrégat dans les marques multivariées : "les marques aléatoires $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées". Bien sûr, contrairement à l'approche paramétrique décrite précédemment, nous ne faisons aucune hypothèse sur la distribution des marques. Soit $Z \subset \mathcal{D}$ un agrégat potentiel. Pour tester une différence de distribution des marques entre Z et Z^c, Oja et Randles 2004 proposent d'utiliser une extension au cas multivarié de la statistique de Wilcoxon-Mann-Whitney, notée

$$U^2_{Z|Z^c} = \frac{p}{c_x^2}[n(Z)||\bar{R}_Z||^2 + n(Z^c)||\bar{R}_{Z^c}||^2$$

où

$$\begin{cases} \bar{R}_Z &= \dfrac{1}{n(Z)} \displaystyle\sum_{i=1}^{n} \tilde{R}_i \mathbb{1}_Z(s_i), \\ \\ c_x^2 &= \displaystyle\sum_{i=1}^{n} \tilde{R}_i^T \tilde{R}_i. \end{cases}$$

Sous H_0 et sous certaines conditions encore à définir (les propriétés statistiques de la matrice de Tyler étant inconnues), la loi limite de $U^2_{Z|Z^c}$ est la loi de Chi-deux à p degrés de liberté. Cette loi ne dépendant pas de $n(Z)$, nous considérons $U^2_{Z|Z^c}$ comme un indice adéquat pour mesurer le degré atypique des marques dans Z et pour comparer des agrégats potentiels de différentes tailles. L'indice de concentration multivarié non-paramétrique est donc

$$I_{MNP}(Z) = U^2_{Z|Z^c}.$$

Il est toujours positif et d'autant plus grand que la différence entre les marques dans Z et dans Z^c est marquée.

1.4 Comparaisons

Nous souhaitons comparer les trois indices de concentration multivariés introduits.

Qualités des différents indices de concentration

Concernant l'aptitude de ces indices de concentration à considérer les agrégats potentiels de manière équitable, il est à noter que les trois indices suivent, sous l'hypothèse nulle d'absence d'agrégats, des lois limites qui ne dépendent pas de $n(Z)$. Néanmoins, il nous est difficile de savoir si cette convergence vers la loi limite est plus ou moins rapide suivant les indices de concentration utilisés.

L'indice de concentration basé sur le modèle Gaussien indépendant $I_{MIG}(Z)$ peut être vu comme une restriction de celui basé sur le modèle Gaussien $I_{MG}(Z)$: il nous paraît donc avoir un handicap, surtout pour analyser des variables fortement dépendantes.

Enfin, la comparaison entre l'indice de concentration basé sur le modèle Gaussien $I_{MG}(Z)$ et celui basé sur les rangs multivariés $I_{MNP}(Z)$ ressemble fortement à celle entre $I_G(Z)$ et $I_R(Z)$ dans le cas univarié. On peut considérer que $I_{MG}(Z)$ sera plus efficace sur des données Gaussiennes alors que $I_{MNP}(Z)$ sera plus adapté à des distributions à queue plus lourde. En effet, Oja et Randles 2004 ont montré que leur version multivariée du test de Wilcoxon-Mann-Whitney devenait plus efficace (au sens de Pitman) que le test de Hotelling lorsque la distribution des données s'éloigne de la loi Gaussienne multivariée.

Etude de simulation

Nous simulons une marque indépendamment pour chacun des 94 départements français, toujours avec le même agrégat C présenté au Chapitre 2, en fixant le nombre de variables à $p = 5$ et selon trois modèles différents. Dans le modèle Gaussien, la marque X_i associée à la localisation géographique s_i suit la loi $\mathcal{N}_p(\mu_i, \Sigma)$ où

$$\mu_i = \begin{cases} (\mu_0 + \beta, \mu_0 + \beta, \mu_0 + \beta, \mu_0 + \beta, \mu_0 + \beta)^T & \text{si } s_i \in C, \\ (\mu_0, \mu_0, \mu_0, \mu_0, \mu_0)^T & \text{sinon.} \end{cases}$$

Dans le modèle Weibull, la marque X_i associée à la localisation géographique s_i suit la loi $\mathcal{W}_p(\lambda_i, k, \Sigma)$ où

$$\lambda_i = \begin{cases} \left(\frac{\mu_0+\beta}{\Gamma(1+\frac{1}{k_1})}, \frac{\mu_0+\beta}{\Gamma(1+\frac{1}{k_2})}, \frac{\mu_0+\beta}{\Gamma(1+\frac{1}{k_3})}, \frac{\mu_0+\beta}{\Gamma(1+\frac{1}{k_4})}, \frac{\mu_0+\beta}{\Gamma(1+\frac{1}{k_5})}\right)^T & \text{si } s_i \in C, \\ \left(\frac{\mu_0}{\Gamma(1+\frac{1}{k_1})}, \frac{\mu_0}{\Gamma(1+\frac{1}{k_2})}, \frac{\mu_0}{\Gamma(1+\frac{1}{k_3})}, \frac{\mu_0}{\Gamma(1+\frac{1}{k_4})}, \frac{\mu_0}{\Gamma(1+\frac{1}{k_5})}\right)^T & \text{sinon} \end{cases}$$

et $k_j = 2, 1 \leq j \leq 5$, correspond au $j^{\text{ème}}$ élément du vecteur des paramètres de forme. Dans le modèle exponentiel, la marque X_i associée à la localisation géographique s_i suit la loi $\mathcal{E}_p(\lambda_i, \Sigma)$ où

$$\lambda_i = \begin{cases} \left(\frac{1}{\mu_0+\beta}, \frac{1}{\mu_0+\beta}, \frac{1}{\mu_0+\beta}, \frac{1}{\mu_0+\beta}, \frac{1}{\mu_0+\beta}\right)^T & \text{si } s_i \in C, \\ \left(\left(\frac{1}{\mu_0}, \frac{1}{\mu_0}, \frac{1}{\mu_0}, \frac{1}{\mu_0}, \frac{1}{\mu_0}\right)^T \right. & \text{sinon.} \end{cases}$$

Dans chacun des trois modèles, la matrice Σ est fixée ainsi :

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ \rho & \rho & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{pmatrix}.$$

Ainsi, le groupe de variables (X^1, X^2, X^3) est indépendant du groupe de variables (X^4, X^5) et le paramètre ρ règle le niveau de dépendance dans chacun des deux groupes. Le paramètre β est la différence d'espérance entre l'intérieur et l'extérieur de l'agrégat, autrement dit l'intensité de l'agrégat. Le paramètre μ_0, c'est-à-dire l'espérance à l'extérieur de l'agrégat, est fixée à 10.

Nous avons simulé 1000 jeux de données selon chaque fonction densité en faisant varier les valeurs des paramètres. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 999$ permutations aléatoires des marques. La table 3.1 donne les résultats obtenus pour les méthodes de balayage Λ_{MIG}, Λ_{MG} et Λ_{MNP}, utilisant respectivement les indices de concentration I_{MIG}, I_{MG} et I_{MNP}.

Suivant le modèle choisi, les valeurs fixées pour le paramètre β sont différentes : en effet, les différences d'espérance sont plus difficiles à détecter dans des modèles où la distribution est à queue épaisse, comme les modèles Weibull et exponentiel. Les performances des différentes méthodes de balayage augmentent avec l'intensité de l'agrégat β mais se dégradent lorsque la corrélation à l'intérieur des groupes ρ s'intensifie. Comme attendu, la méthode basée sur le modèle Gaussien indépendant est moins performante que celle prenant en compte les corrélations entre variables, et l'écart s'accroît avec le paramètre ρ. On constate également que la méthode basée sur les rangs multivariés prend le dessus en présence de données issues du modèle exponentiel, donc pouvant plus souvent contenir des valeurs extrêmes. Les taux de vrais positifs et faux positifs nous indiquent enfin que les agrégats exhibés par la méthode basée sur les rangs multivariés sont généralement de taille supérieure à ceux exhibés par les méthodes basées sur la vraisemblance. Ces observations sont similaires à celles qui ont pu être faites au Chapitre 2 pour des données univariées.

TABLE 3.1 – Tests appliqués à des marques multivariées

ρ		**modèle Gaussien**					**modèle Weibull**					**modèle exponentiel**			
	β		Λ_{MIG}	Λ_{MG}	Λ_{MNP}	β		Λ_{MIG}	Λ_{MG}	Λ_{MNP}	β		Λ_{MIG}	Λ_{MG}	Λ_{MNP}
0.0	1.0	Power	0.659	**0.841**	0.808	7	Power	0.870	**0.959**	0.887	12	Power	0.612	0.750	**0.863**
		%TP	0.825	0.898	**0.984**		%TP	0.970	0.980	**0.991**		%TP	0.940	0.951	**0.985**
		%FP	0.175	**0.102**	0.124		%FP	0.030	**0.020**	0.037		%FP	0.060	**0.049**	0.182
0.2	1.0	Power	0.586	**0.665**	0.609	7	Power	0.753	**0.884**	0.760	12	Power	0.603	0.706	**0.844**
		%TP	0.770	0.854	**0.922**		%TP	0.951	0.976	**0.987**		%TP	0.950	0.945	**0.985**
		%FP	0.230	**0.146**	0.162		%FP	0.049	**0.024**	0.047		%FP	**0.050**	0.055	0.092
0.5	1.0	Power	0.412	**0.447**	0.426	7	Power	0.602	**0.788**	0.523	12	Power	0.497	0.672	**0.711**
		%TP	0.697	0.790	**0.843**		%TP	0.922	0.958	**0.977**		%TP	0.933	0.935	**0.982**
		%FP	0.303	**0.210**	0.285		%FP	0.078	**0.042**	0.103		%FP	**0.067**	0.065	0.114
0.8	1.0	Power	0.304	**0.323**	0.280	7	Power	0.513	**0.728**	0.486	12	Power	0.428	0.658	**0.667**
		%TP	0.637	0.723	**0.754**		%TP	0.917	0.969	**0.985**		%TP	0.923	0.906	**0.956**
		%FP	0.363	**0.277**	0.298		%FP	0.083	**0.031**	0.128		%FP	**0.077**	0.094	0.138
0.0	1.5	Power	0.993	**1.000**	0.998	9	Power	0.946	**0.992**	0.972	14	Power	0.715	0.841	**0.914**
		%TP	0.957	0.978	**0.988**		%TP	0.982	0.991	**0.997**		%TP	0.978	0.978	**0.995**
		%FP	0.043	0.022	**0.019**		%FP	0.018	**0.009**	0.019		%FP	0.022	**0.021**	0.044
0.2	1.5	Power	0.968	**0.990**	0.977	9	Power	0.904	**0.971**	0.917	14	Power	0.668	0.790	**0.823**
		%TP	0.917	0.961	**0.981**		%TP	0.965	0.983	**0.991**		%TP	0.964	0.966	**0.988**
		%FP	0.083	**0.039**	0.042		%FP	0.035	**0.017**	0.019		%FP	0.036	**0.034**	0.053
0.5	1.5	Power	0.868	**0.956**	0.912	9	Power	0.796	**0.934**	0.802	14	Power	0.562	0.758	**0.814**
		%TP	0.854	0.931	**0.962**		%TP	0.958	0.983	**0.991**		%TP	0.958	0.958	**0.988**
		%FP	0.146	**0.069**	0.124		%FP	0.042	**0.017**	0.054		%FP	**0.041**	0.043	0.095
0.8	1.5	Power	0.737	**0.858**	0.794	9	Power	0.713	**0.886**	0.687	14	Power	0.471	0.710	**0.728**
		%TP	0.815	0.902	**0.933**		%TP	0.944	0.979	**0.990**		%TP	0.941	0.924	**0.952**
		%FP	0.185	**0.098**	0.143		%FP	0.056	**0.021**	0.74		%FP	**0.059**	0.076	0.088

Un exemple illustratif

Nous appliquons les méthodes de balayage basées sur Λ_{MG} et Λ_{MNP} au jeu de données environnementales évoqué au début du Chapitre. Ces données proviennent de l'analyse de différents lichens de type *Xanthoria parietina*, recueillis en 128 localisations réparties sur toute l'agglomération lilloise. Les lichens étant des organismes captant naturellement les polluants chimiques, la teneur de 14 de ces polluants a pu être mesurée pour chacun de ces lichens : aluminium (Al), antimoine (Sb), arsenic (As), cadmium (Cd), cobalt (Co), chrome (Cr), cuivre (Cu), plomb (Pb), manganèse (Mn), mercure (Hg), nickel (Ni), titane (Ti), zinc (Zn) et vanadium (V). A partir des valeurs observées pour ces 14 polluants, un indice de pollution globale synthétisant ces informations, nommé Ratio d'Imprégnation Moyen (RIM), a été créé : les détails sont donnés par OCCELLI, BAVDEK et al. 2016. La plupart de ces 14 polluants sont fortement corrélés aux autres, mais certains le sont moins, ce qu'atteste la Table 3.2.

TABLE 3.2 – Coefficients de corrélation entre variables

	Al	Cr	Cu	As	Hg	Cd	Mn	Co	Sb	Ni	V	Pb	Ti	Zn	RIM
Al	1	0,65	0,24	0,42	0,10	0,19	0,58	0,48	0,42	0,30	0,67	0,21	0,13	0,25	0,45
Cr	0,65	1	0,83	0,70	0,59	0,73	0,85	0,94	0,78	0,88	0,80	0,81	0,61	0,85	0,93
Cu	0,24	0,83	1	0,57	0,64	0,81	0,65	0,84	0,78	0,87	0,54	0,90	0,64	0,89	0,91
As	0,42	0,70	0,57	1	0,67	0,41	0,73	0,75	0,55	0,58	0,84	0,52	0,64	0,57	0,67
Hg	0,10	0,59	0,64	0,67	1	0,69	0,41	0,70	0,53	0,72	0,48	0,75	0,41	0,72	0,71
Cd	0,19	0,73	0,81	0,41	0,69	1	0,51	0,80	0,63	0,89	0,36	0,93	0,45	0,93	0,88
Mn	0,58	0,85	0,65	0,73	0,41	0,51	1	0,82	0,61	0,68	0,81	0,60	0,72	0,68	0,76
Co	0,48	0,94	0,84	0,75	0,70	0,80	0,82	1	0,72	0,95	0,74	0,89	0,65	0,91	0,95
Sb	0,42	0,78	0,78	0,55	0,53	0,63	0,61	0,72	1	0,69	0,59	0,70	0,53	0,75	0,86
Ni	0,30	0,88	0,87	0,58	0,72	0,89	0,68	0,95	0,69	1	0,54	0,97	0,59	0,96	0,94
V	0,67	0,80	0,54	0,84	0,48	0,36	0,81	0,74	0,59	0,54	1	0,45	0,54	0,52	0,66
Pb	0,21	0,81	0,90	0,52	0,75	0,93	0,60	0,89	0,70	0,97	0,45	1	0,57	0,96	0,93
Ti	0,13	0,61	0,64	0,64	0,41	0,45	0,72	0,65	0,53	0,59	0,54	0,57	1	0,63	0,65
Zn	0,25	0,85	0,89	0,57	0,72	0,93	0,68	0,91	0,75	0,96	0,52	0,96	0,63	1	0,96
RIM	0,45	0,93	0,91	0,67	0,71	0,88	0,76	0,95	0,86	0,94	0,66	0,93	0,65	0,96	1

Tous les polluants présentent des distributions à queue épaisse, caractérisées par la présence de valeurs extrêmes, comme le montre la Figure 3.1.

Afin d'identifier les zones présentant un niveau de pollution significativement

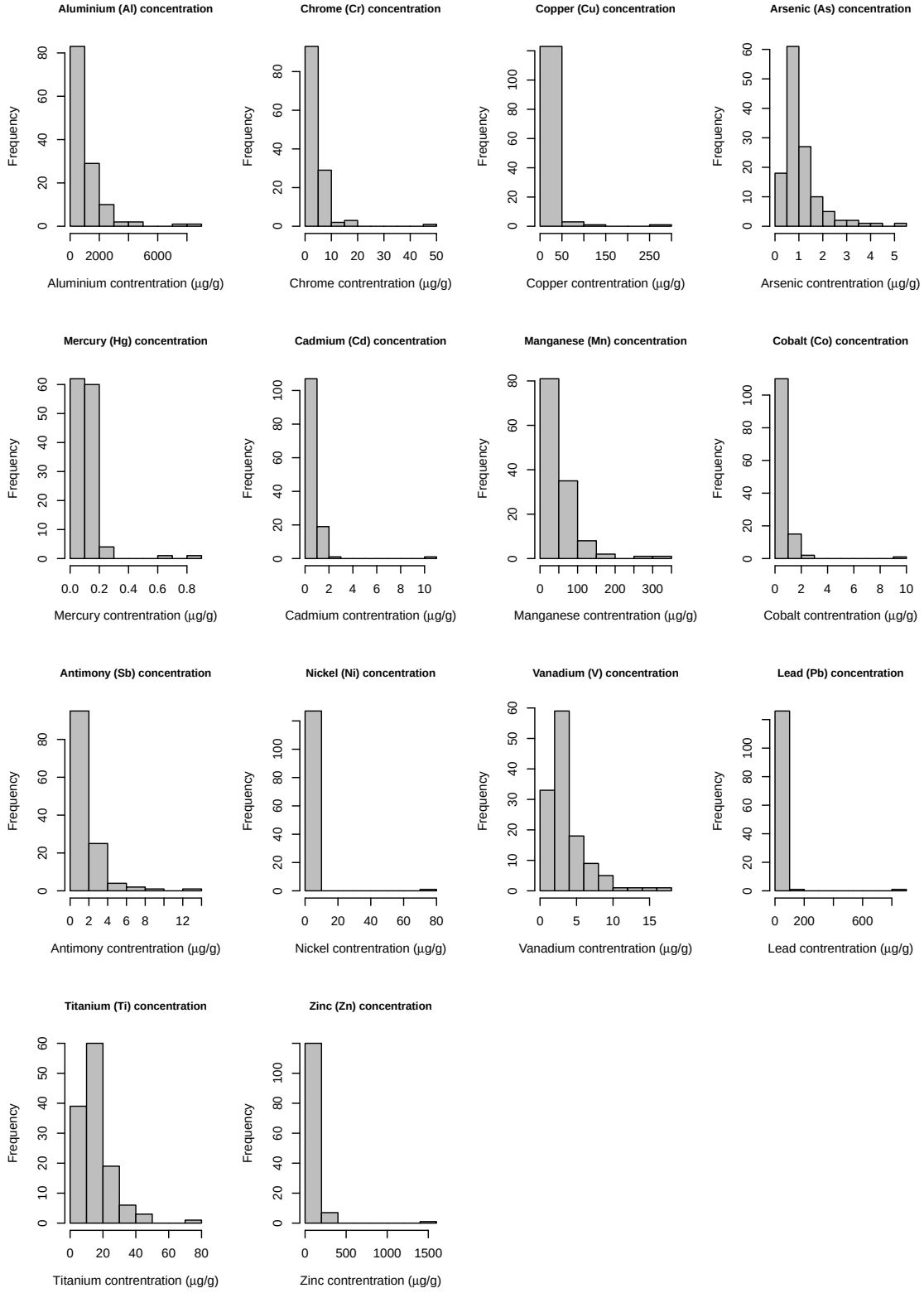

FIGURE 3.1 – Distributions des polluants, Métropole de Lille.

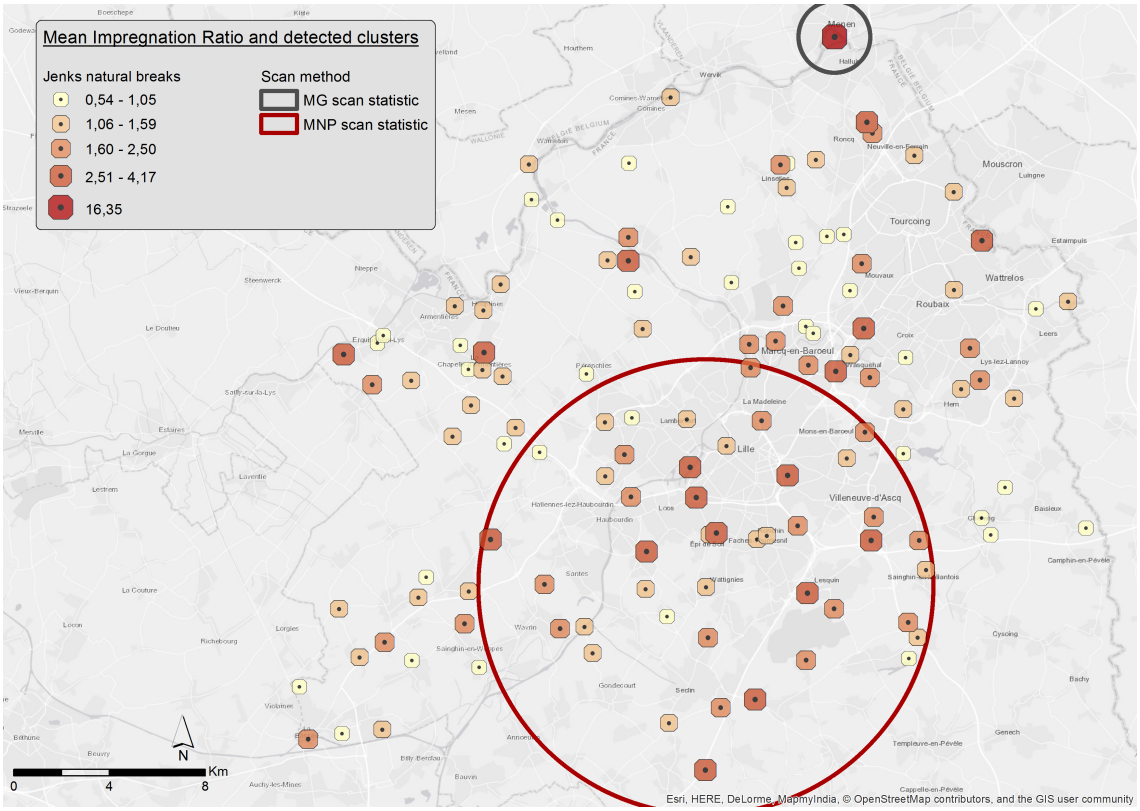

FIGURE 3.2 – Valeurs du RIM et agrégats détectés par les méthodes de balayage.

plus élevé, nous avons appliqué les méthodes de balayage basées sur Λ_{MG} et λ_{MNP} à ce jeu de données, en nous basant sur $T = 999$ permutations aléatoires pour estimer la significativité. Les deux méthodes exhibent des agrégats très significatifs ($p = 0.001$ pour Λ_{MG} et $p = 0.002$ pour Λ_{MNP}) illustrés par la Figure 3.2.

La médiane [écart interquartile] de chaque polluant à l'intérieur et à l'extérieur de l'agrégat le plus probable est donnée par la Table 3.3. L'agrégat associé à λ_{MG} ne comporte qu'une seule localisation, située au Nord de la zone. Cette localisation présente le taux le plus élevé pour 12 des 14 polluants, et parmi les plus élevés pour les 2 polluants restants. Suivant les polluants, le taux observé en cette localisation est 4 à 48 fois supérieur à la médiane des autres observations : nous sommes donc en présence d'une donnée très extrême et l'indice de concentration basé sur le rapport de vraisemblance Gaussien y accorde beaucoup d'importance. A contrario, la méthode basée sur λ_{MNP} identifie un agrégat contenant 42 localisations centrées sur la ville de Lille (le cercle rouge), et ne contenant pas celle détectée par Λ_{MG}. Les valeurs du RIM sont 1.74 [1.28-2.46] à l'intérieur et 1.18 [0.91-1.66] à l'extérieur de l'agrégat, et les médianes des polluants sont toujours plus élevées à l'intérieur qu'à l'extérieur. On retrouve là un comportement classique pour les indices de concentration basés sur les rangs (multivariés ou non) : ils sont plus sensibles à des groupes importants de valeurs plus élevées dans un voisinage qu'à un petit de groupe de valeurs extrêmes.

TABLE 3.3 – Comparaison des statistiques de balayage Gaussienne (Λ_{MG}) et non-paramétrique (λ_{MNP}). Pour chaque méthode de balayage, les niveaux de polluants sont décrits (médiane [premier quartile ; troisième quartile]) à l'intérieur et à l'extérieur de l'agrégat le plus probable.

	Métropole de Lille			
	λ_{MG} ($p = 0.001$)		λ_{MNP} ($p = 0.002$)	
Polluants	Out ($n = 127$)	In ($n = 1$)	Out ($n = 84$)	In ($n = 44$)
Aluminium	690 [496 ;1162]	3057	648 [458 ;1119]	761 [576 ;1260]
Chrome	3.21 [2.26 ;5.17]	45.1	3.00 [2.06 ;5.03]	3.68 [2.76 ;6.54]
Cuivre	12.6 [8.83 ;19.2]	259	10.9 [7.69 ;17.4]	16.9 [11.8 ;21.8]
Arsenic	0.78 [0.61 ;1.21]	5.10	0.77 [0.56 ;1.18]	0.85 [0.67 ;1.41]
Mercure	0.11 [0.09 ;0.13]	0.81	0.10 [0.09 ;0.11]	0.12 [0.10 ;0.16]
Cadmium	0.46 [0.30 ;0.83]	10.6	0.35 [0.25 ;0.72]	0.80 [0.46 ;1.34]
Manganèse	44.0 [33.0 ;55.9]	309	39.5 [30.8 ;52.5]	50.2 [39.0 ;58.0]
Cobalt	0.52 [0.36 ;0.76]	9.62	0.48 [0.31 ;0.74]	0.57 [0.45 ;0.84]
Antimoine	1.09 [0.66 ;2.00]	14.0	0.87 [0.55 ;1.42]	1.54 [1.08 ;2.62]
Nickel	2.33 [1.58 ;3.21]	72.9	2.15 [1.44 ;2.98]	2.52 [1.97 ;3.71]
Vanadium	2.79 [2.00 ;4.29]	15.0	2.62 [1.90 ;4.29]	3.00 [2.14 ;4.41]
Plomb	18.0 [9.89 ;33.0]	873	13.0 [8.00 ;24.5]	34.5 [18.0 ;46.5]
Titane	13.0 [10.0 ;19.5]	76.0	12.0 [9.00 ;19.2]	14.5 [12.0 ;20.0]
Zinc	87.1 [59.5 ;124]	1583	72.7 [52.3 ;111]	113 [83.5 ;168]

2 Marques fonctionnelles

À partir d'Octobre 2018, j'ai co-encadré la thèse de Zaineb Smida en compagnie d'Ali Gannoun. Zaineb a commencé par s'intéresser à des tests non-paramétriques de comparaisons d'échantillons pour des données fonctionnelles : elle a introduit un test de la médiane dans ce cadre-là. Dans un second temps, la relation entre ce type de tests et les statistiques de balayage étant évidente, il nous a paru naturel d'utiliser l'un de ces tests pour proposer une statistique de balayage spatial permettant d'analyser des processus ponctuels spatiaux dont les marques sont fonctionnelles. Il est à noter que les données fonctionnelles sont de plus en plus présentes dans de nombreuses études (environnementales, médicales, économétriques ...) en raison notamment de la multiplication des capteurs. Pour plus de détails, consulter RAMSAY et SILVERMAN 2005. Ces données sont de plus en plus souvent géo-localisées (DELICADO et al. 2010) et, même si quelques études ont été menées pour modéliser (CRONIE et al. 2021) ou partitionner (GAETAN, GIRARDI et PASTRES 2017) de tels jeux de données, il n'existe, à notre connaissance, aucune méthode de détection d'agrégat particulièrement adaptée.

Dans cette section, on considèrera une variable aléatoire X à valeurs dans un espace fonctionnel χ, associée à chaque événement du processus ponctuel. Par souci de simplicité, nous supposerons que χ est un espace hilbertien, par exemple $L^2([0,1],\mathbb{R})$. À partir du jeu de données $\{(s_i, x_i) : i \in [\![1, n]\!]\}$, où $x_i \in \chi$ est la marque fonctionnelle associée à la localisation s_i, l'indice de concentration $I(Z)$ doit nous permettre de déterminer dans quel agrégat potentiel $Z \subset \mathcal{C}$ les observations de X sont les plus atypiques par rapport à l'ensemble des observations de X dans W.

2.1 Un indice de concentration non-paramétrique

Remarquons d'abord que, pour construire un indice de concentration pour cette statistique de balayage, la démarche classique qui consiste à s'appuyer sur un rapport de vraisemblance généralisé n'est pas évidente dans la cadre fonctionnel. En effet, comme précisé par Ferraty 2011, la notion de densité de probabilité, et donc de vraisemblance, n'est pas clairement définie pour des variables aléatoires fonctionnelles. Bien que certaines approximations, comme celle de Jacques et Preda 2013, aient été proposées, nous avons préféré utiliser une approche non-paramétrique en nous basant sur la généralisation du test de Wilcoxon-Mann-Whitney au cadre fonctionnel introduite par Chakraborty et Chaudhuri 2015.

Version fonctionnelle du test de Wilcoxon-Mann-Whitney

Pour tester une différence de distribution des marques fonctionnelles entre Z et Z^c, Chakraborty et Chaudhuri 2015 proposent d'utiliser une extension au cas fonctionnel de la statistique de Wilcoxon-Mann-Whitney, notée

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

où $\|.\|_\chi$ est une norme associée à l'espace fonctionnel χ. On remarque que cette formule fait à nouveau intervenir la fonction signe, cette fois-ci dans le cadre fonctionnel. Afin de standardiser cette statistique, nous introduisons

$$\tilde{T}_{\text{WMW}}(Z) = \sqrt{\frac{n(Z)n(Z^c)}{n}} T_{\text{WMW}}(Z)$$

et l'indice de concentration

$$\text{U}(Z) := \|\tilde{T}_{\text{WMW}}(Z)\|_\chi.$$

En effet, Chakraborty et Chaudhuri ont montré que, sous H_0, si $\dfrac{n(Z)}{n} \to \gamma \in [0,1]$ lorsque $n(Z), n(Z^c) \to \infty$,

$$\tilde{T}_{\text{WMW}}(Z) \text{ converge faiblement vers } G(0,\Gamma), \tag{3.1}$$

où $G(m,C)$ est la distribution d'un processus Gaussien dans χ de fonction moyenne $m \in \chi$ et de fonction de covariance C. Cette fonction de covariance Γ ne dépendant pas de $n(Z)$ et $n(Z^c)$, la loi asymptotique de $\text{U}(Z)$ est la même quelle que soit la taille de l'agrégat potentiel Z.

Astuces de calcul

D'un point de vue calculatoire, évaluer cet indice de concentration peut être assez intensif car cela nécessite, pour chaque agrégat potentiel Z, la somme de $n(Z)\big(n - n(Z)\big)$ termes. Néanmoins, Frévent et al. 2021 ont montré que la statistique de Wilcoxon-Mann-Whitney pouvait se réécrire

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

Ainsi, si l'on calcule au préalable, pour tout $i \in [\![1, n]\!]$, les fonctions aléatoires

$$W_i = \sum_{j=1}^{n} \frac{X_j - X_i}{\|X_j - X_i\|_\chi},$$

le calcul de $\mathrm{U}(Z)$ ne nécessite plus que de sommer n_Z de ces termes.

De plus, il est également utile, lorsque l'on cherche à maximiser l'indice de concentration $U(Z)$ sur l'ensemble des agrégats potentiels $\mathcal{C}$, de parcourir cet ensemble de manière intelligente. En effet, si deux agrégats potentiels successifs Z et Z' ne diffèrent que d'une localisation, i.e. $Z' = Z \cup s_k$, on a

$$\begin{aligned} n(Z')n(Z'^c)T_{\mathrm{WMW}}(Z') &= \sum_{\{i:s_i \in Z'\}} \sum_{j=1}^{n} \frac{X_j - X_i}{\|X_j - X_i\|_\chi} \\ &= \sum_{\{i:s_i \in Z\}} W_i + W_k \\ &= n(Z)n(Z^c)T_{\mathrm{WMW}}(Z) + W_k \end{aligned}$$

donc la mise à jour de l'indice de concentration se fait encore plus rapidement. Il est à noter que cette dernière astuce n'est pas limitée au cadre des marques fonctionnelles et peut s'appliquer à la plupart des indices de concentration introduits dans ce mémoire.

2.2 Applications

Etude de simulation

Nous simulons une marque fonctionnelle indépendamment pour chacun des 94 départements français, toujours avec le même agrégat C présenté au Chapitre 2. Ces marques fonctionnelles appartiennent à $\chi = L^2([0, 1], \mathbb{R})$ et sont définies ainsi :

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

où, pour tout $k \geq 1$, $e_k(t) = \sqrt{2}\sin(t/\sigma_k)$ est une base orthonormale de χ, $\sigma_k = ((k - 0.5)\pi)^{-1}$ et les $Z_{i,k}$ sont des variables aléatoires indépendantes correspondant à la projection de X_i sur la base de Karhunen-Loève (Karhunen 1947, Lévy et Loève 1948).
Nous avons examiné deux cas de figure, soit un mouvement Brownien standard (lorsque les $Z_{i,k}/\sigma_k$ suivent une loi $\mathcal{N}(0, 1)$), soit un processus de Student centré à cinq degrés de liberté (lorsque les $Z_{i,k}/\sigma_k$ suivent une loi $t(5)$).
Les distributions des marques à l'intérieur et à l'extérieur de l'agrégat C diffèrent d'un décalage Δ. Trois types de décalage ont été envisagés : pour tout $t \in [0, 1]$, $\Delta_1(t) = ct$, $\Delta_2(t) = ct(1 - t)$ et $\Delta_3(t) = c\sin(2\pi t)$, avec $c > 0$.

La statistique de balayage fonctionnelle basée sur l'indice de concentration de type Wilcoxon-Mann-Whitney introduit précédemment est appelée $\Lambda_{\mathrm{WMWFSS}}$. Nous la comparons à des statistiques de balayage univariées appliquées à des résumés des marques fonctionnelles :

— la première se base sur la moyenne des marques, à savoir

$$\forall i \in [\![1, n]\!], \quad \bar{X}_i = \int_0^1 X_i(t)dt.$$

Chacune de ces valeurs moyennes est associée à sa localisation et on calcule alors la statistique de balayage basée sur les rangs introduite dans le second paragraphe du mémoire, en utilisant les mêmes agrégats potentiels et les mêmes permutations aléatoires que pour la statistique de balayage fonctionnelle. On note cette statistique de balayage univariée basée sur les moyennes Λ_{MBUSS}.

— la seconde, inspirée par la fonction LISA définie par Mateu, Lorenzo et Porcu 2007, s'appuie sur les déviations à la fonction moyenne des marques, à savoir

$$\forall i \in [\![1, n]\!], \quad D_i = \int_0^1 (X_i(t) - \bar{X}(t))^2 dt,$$

où

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

est la fonction moyenne associée aux marques fonctionnelles. Chacune de ces déviations est associée à sa localisation et on calcule de nouveau la statistique de balayage basée sur les rangs. On note cette statistique de balayage univariée basée sur les déviations Λ_{DBUSS}.

Nous avons simulé 100 jeux de données selon chaque distribution en faisant varier les valeurs des paramètres. Le niveau du test est fixé à $\alpha = 5\%$. Les p-valeurs sont calculées à partir de $T = 99$ permutations aléatoires des marques. Les tableaux 3.4, 3.5 et 3.6 donnent les résultats obtenus pour différents décalages.

Comme attendu, les performances des trois méthodes de balayage s'améliorent lorsque l'intensité de l'agrégat c augmente. On peut noter que :

- les puissances des trois méthodes augmentent avec c mais cette augmentation est plus lente lorsque le processus est distribué selon la loi de Student que dans le cas Gaussien : cela peut s'expliquer par le fait que la loi de Student est à queue plus lourde donc des valeurs éloignées de l'espérance sont moins atypiques. La méthode fonctionnelle Λ_{WMWFSS}, qui s'appuie sur l'intégralité de l'information disponible, est plus puissante que les deux autres. La différence entre Λ_{WMWFSS} et Λ_{MBUSS} est faible lorsque le décalage $\Delta(t)$ est linéaire, mais elle s'accroît lorsque celui-ci est quadratique. Quant au décalage sinusoïdal, il est totalement indétectable par la méthode Λ_{MBUSS} car ce décalage ne produit aucun changement dans la valeur moyenne. La méthode univariée basée sur les déviations Λ_{DBUSS} est un peu plus adaptée pour le décalage sinusoïdal mais sa puissance reste largement inférieure à la méthode fonctionnelle.
- Les taux de vrais positifs et de faux positifs s'améliorent également lorsque c augmente. Comme pour sa détection, la localisation de l'agrégat le plus significatif est plus facile lorsque la distribution sous-jacente est Gaussienne que de Student. L'utilisation de la totalité de l'information est importante pour retrouver le "vrai" agrégat donc il est logique de constater que les taux de vrais positifs et faux positifs obtenus par la méthode fonctionnelle Λ_{WMWFSS} sont globalement meilleurs que ceux obtenus par les méthodes univariées. La

TABLE 3.4 – Simulations avec $\Delta_1(t) = ct$.

Normal distribution					**Student distribution**				
c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{MBUSS}	Λ_{DBUSS}
1.25	Power	**0.310**	0.260	0.110	1.5	Power	**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	Power	**0.380**	0.340	0.150	2.0	Power	**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	Power	**0.590**	0.450	0.160	2.5	Power	**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	Power	**0.730**	0.660	0.300	3.0	Power	**0.920**	0.870	0.520
	nTP	**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	Power	**0.920**	0.890	0.570	3.5	Power	**0.980**	0.940	0.800
	%TP	**0.978**	0.961	0.879		%TP	**9.974**	0.973	0.956
	%FP	**0.056**	0.070	0.077		%FP	**0.035**	0.050	0.048
3.0	Power	**1.000**	1.000	0.870	4.0	Power	**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	Power	**1.000**	**1.000**	0.910	4.5	Power	**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 – Simulations avec $\Delta_2(t) = ct(1-t)$.

Normal distribution					**Student distribution**				
c		Λ_{WMWFSS}	Λ_{WMWUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{WMWUSS}	Λ_{DBUSS}
4.0	Power	**0.410**	0.330	0.120	4.5	Power	**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	Power	**0.460**	0.320	0.160	5.5	Power	**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	Power	**0.560**	0.470	0.210	6.5	Power	**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	Power	**0.700**	0.530	0.260	7.5	Power	**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	Power	**0.830**	0.590	0.290	8.5	Power	**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	Power	**0.870**	0.760	0.460	9.5	Power	**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	Power	**0.960**	0.810	0.530	10.5	Power	**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 – Simulations avec $\Delta_3(t) = c\,\sin(2\pi t)$.

Normal distribution					**Student distribution**				
c		Λ_{WMWFSS}	Λ_{WMWUSS}	Λ_{DBUSS}	c		Λ_{WMWFSS}	Λ_{WMWUSS}	Λ_{DBUSS}
1.0	Power	**0.310**	0.080	0.170	1.0	Power	**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	Power	**0.660**	0.040	0.350	1.25	Power	**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	Power	**0.960**	0.060	0.660	1.5	Power	**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	Power	**1.000**	0.070	0.940	1.75	Power	**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	Power	**1.000**	0.060	1.000	2.0	Power	**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	Power	**1.000**	0.020	1.000	2.25	Power	**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	Power	**1.000**	0.050	1.000	2.5	Power	**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

différence est plus notable concernant les taux de faux positifs : les méthodes univariées ont tendance à sélectionner comme agrégat le plus significatif des zones plus grandes que le "vrai" agrégat.

Un exemple illustratif

Nous appliquons maintenant ces méthodes de balayage à un jeu de données réelles portant sur l'évolution de la population dans 47 provinces espagnoles lors des 20 dernières années. Les données de population sont fournies par l'Institut Espagnol de la Statistique (`www.ine.es`) et les données géographiques des provinces espagnoles (contours et centres : voir figure 3.3) par la bibliothèque *R raster* . Pour des raisons géographiques, nous avons exclu de l'étude les îles Baléares et Canaries, ainsi que les cités autonomes de Ceuta et Melilla situées en Afrique du Nord. À chaque centre de province s_i, i variant de 1 à 47, on associe l'évolution démographique entre les années 1998 à 2019 (voir figure 3.4). L'évolution démographique est définie comme la population entre les années 1998 et 2019 divisée par la population en 1998.

Notre objectif est d'identifier une possible zone spatiale où l'évolution démographique serait différente. Pour cela, nous avons calculé la statistique de balayage fonctionnelle : $\Lambda_{\text{WMWFSS}} = 2.72025$. Par rapport aux statistiques calculées sur $T = 999$ permutations aléatoires, la valeur de la statistique observée est très significative ($p_{value} = 0.001$) et l'agrégat le plus significatif $\hat{C}$ est représenté sur la figure 3.5. Cet agrégat comprend 13 provinces de l'Ouest de l'Espagne (*Asturias*, *Galicia*, *Extremadura* et l'Ouest de *Castilla y León*) dans lesquelles les marques fonctionnelles

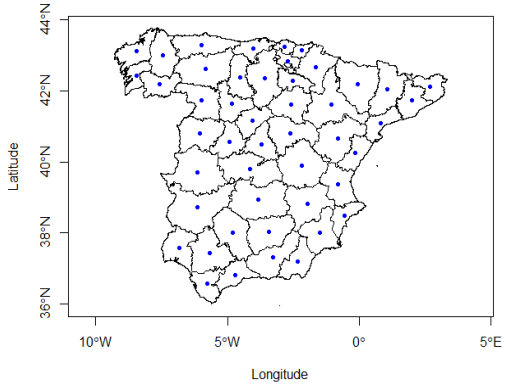

FIGURE 3.3 – Les 47 provinces espagnoles et leurs centres.

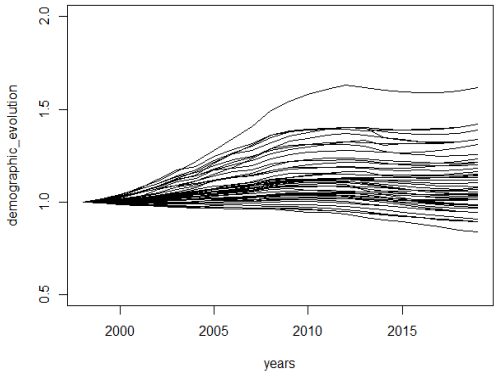

FIGURE 3.4 – Évolution démographique des 47 provinces de 1998 à 2019.

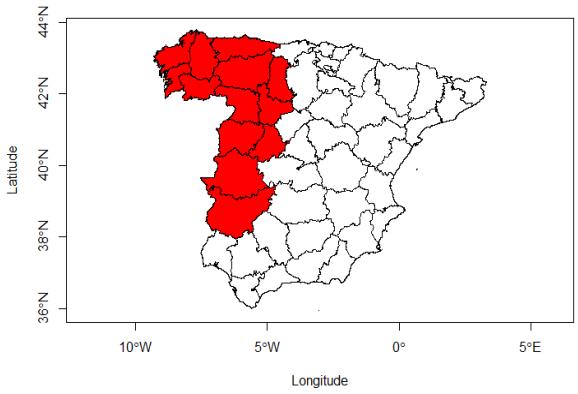

FIGURE 3.5 – L'agrégat le plus significatif détecté par la méthode de balayage fonctionnelle.

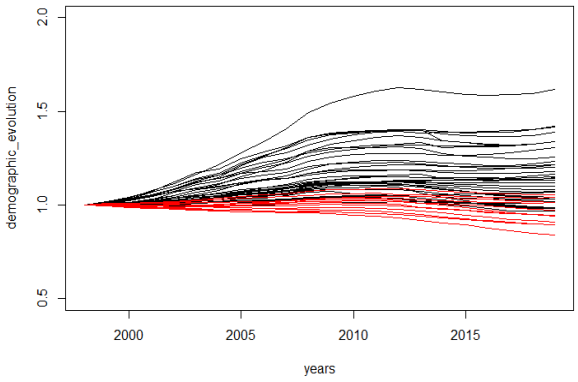

FIGURE 3.6 – Les 13 courbes en rouge représentent les provinces incluses dans l'agrégat le plus significatif.

sont de niveau significativement plus faible qu'ailleurs, comme le montre la figure 3.6. D'après l'Institut Espagnol de la Statistique, ces 13 provinces sont parmi celles qui présentent les taux de natalité les plus faibles, et les taux de mortalité les plus forts. Cela s'explique notamment par la forte émigration des populations jeunes de ces provinces vers les régions où les emplois sont plus nombreux comme Madrid et la Catalogne, voire vers les pays étrangers comme l'Allemagne.

Chapitre 4

Perspectives

Publications : [13], [14], [15], [16].

Dans cette dernière partie du mémoire, je décris les pistes de recherche que je suis en train de creuser en ce moment. Pour certaines, le travail est bien engagé ; pour d'autres, il reste encore beaucoup de chemin. Le premier paragraphe, qui s'appuie sur [15], décrit des alternatives à la statistique de Wilcoxon-Mann-Whitney pour données fonctionnelles introduite dans le Chapitre 3. Dans le second paragraphe, j'explique comment prendre en compte l'autocorrélation spatiale dans les méthodes de balayage, comme nous avons commencé à le faire dans [14]. Le troisième paragraphe s'intéresse à une utilisation alternative des indices de concentration introduits précédemment, afin de comparer la distribution spatiale de deux semis de points, comme dans [13]. Enfin, nous nous demanderons comment identifier à la fois une intensité d'événements plus importante et des marques associées atypiques comme dans [16].

1 Une panoplie de statistiques de balayage pour données fonctionnelles

Dans la deuxième section du Chapitre 3, nous avons vu comment définir une statistique de balayage pour données fonctionnelles en se basant sur le test de type Wilcoxon-Mann-Whitney introduit par Chakraborty et Chaudhuri 2015. Néanmoins, il existe de nombreux autres tests pour l'égalité de deux populations dans le cadre fonctionnel. Frévent et al. 2021 se sont notamment appuyés sur un test de type ANOVA introduit par Cuevas, Febrero-Bande et Fraiman 2004 pour définir une nouvelle statistique de balayage pour données fonctionnelles. Sous l'impulsion de Camille Frévent et de ses collègues lillois, nous avons développé une bibliothèque *R* permettant de donner facilement accès à ces différentes statistiques de balayage fonctionnelles.

Dans cette direction, Zaineb Smida et moi avons l'intention de développer deux autres statistiques de balayage fonctionnelles et de les comparer aux existantes. La première est basée sur une généralisation du test de Student introduite par Horváth, Kokoszka et Reeder 2013. La seconde s'appuie sur un test de la médiane pour données fonctionnelles développé par Smida, Cucala et Gannoun 2022.

2 Des statistiques de balayage prenant en compte l'autocorrélation spatiale

Dans la littérature des méthodes de balayage, on considère généralement que les observations (que ce soit les localisations des événements ou les marques associées) sont indépendantes. Cette hypothèse simplifie énormément la construction d'indices de concentration mais elle n'est pas vraiment réaliste car des données spatiales sont généralement sujettes à de l'auto-corrélation spatiale (Cressie 1993). Par exemple, on constate généralement une corrélation positive entre les mesures d'un polluant à des sites voisins.
Il existe peu de travaux qui prennent en compte cette corrélation spatiale dans la détection d'agrégats. Pour des processus ponctuels non marqués, Loh et Z. Zhu 2007 ont montré que, d'un point de vue théorique comme pratique, les méthodes de détection d'agrégat classiques identifient des agrégats significatifs trop larges en présence de corrélation spatiale. Pour des processus ponctuels marqués par une variable continue X, Lee, Gangnon et J. Zhu 2017 identifient des agrégats spatiaux sur les coefficients d'une régression de X sur une variable explicative.

Avec mes collègues lillois Michaël Genin et Mohamed-Ahmed Salem, nous réfléchissons depuis quelques temps aux moyens de modifier les statistiques de balayage classiques pour données continues afin de prendre en compte cette autocorrélation spatiale. Dans un article publié récemment, nous avons introduit une nouvelle statistique de balayage issue d'un modèle classique en économétrie, le modèle SAR (Spatial AutoRegressive) (Cliff et Ord 1973). Lorsque des observations sont issues d'un tel modèle, il est facile de les rendre indépendantes par un simple produit matriciel. Néanmoins, cela nécessite de connaître les deux paramètres du modèle SAR : la matrice de voisinage W et le niveau d'autocorrélation spatiale ρ. Dans notre article, nous avons considéré que W était connue et proposé une méthode d'estimation originale de ρ : en effet, les méthodes classiques ne fonctionnent que si aucun agrégat spatial n'est présent dans le jeu de données, et c'est précisément ce que l'on cherche à tester ! Les premiers résultats sont encourageants mais l'efficacité de cette nouvelle statistique de balayage est fortement liée au fait que les données sont bien issues d'un modèle SAR. Nous envisageons donc de nous appuyer sur d'autres modèles classiques comme le modèle SEM (Spatial Error Model).

3 Des méthodes de balayage globales pour comparer deux semis de points

Mon collègue Florent Bonneu et moi avons commencé en 2019 à réfléchir à l'utilisation des statistiques de balayage pour comparer la distribution spatiale de deux semis de points. Comment, par exemple, tester si les feux de forêt accidentels sont répartis de la même manière que les feux de forêt criminels ? Il existe quelques tests globaux (i.e. analysant les différences de distribution sur la globalité du domaine d'observation) s'appuyant notamment sur des estimateurs de l'intensité des deux processus observés (Fuentes-Santos, González-Manteiga et Mateu 2017) ou de la fonction K de Ripley (Diggle et Chetwynd 1991).

Néanmoins, les statistiques de balayage, bien qu'entrant dans la catégorie des

tests locaux (i.e. pointant sur la différence locale la plus significative) peuvent également être utilisées. De plus, nous avons pensé qu'une légère modification de ces statistiques de balayage pouvait permettre de capter plus facilement la globalité des différences entre les deux distributions spatiales : il s'agit de remplacer la recherche du maximum de l'indice de concentration par le calcul de la variance empirique de ce même indice de concentration. En effet, plus cet indice varie sur l'ensemble des agrégats potentiels, plus les différences de distribution spatiale sont marquées.

Dans un premier article, nous avons mené une étude de simulation afin de comparer les performances des méthodes déjà existantes avec les méthodes de balayage classiques et leur version globale, telle qu'expliquée précédemment. La conclusion de cette étude est double :

— les méthodes de balayage classiques, bien que conçues pour mener des tests locaux, ont des performances tout à fait comparables aux autres.
— l'adaptation des méthodes de balayage aux tests globaux n'est pas aussi efficace que prévu.

Nous réfléchissons donc à une autre manière de globaliser les statistiques de balayage pour obtenir de meilleurs résultats.

4 Des statistiques de balayage pour identifier des agrégats de localisations et de marques atypiques

L'idée développée ici est de mettre en place une ou des méthodes de balayage afin d'identifier, à partir de l'observation d'un processus ponctuel marqué par une variable réelle, la zone où la somme totale des marques est la plus significativement élevée. Cela diffère des études classiques où on cherche soit la zone où le nombre d'événements est le plus significativement élevé (les marques ne sont pas utilisées, seul le volume de la zone compte), soit la zone où la moyenne des marques est la plus significativement élevée (le volume de la zone n'est pas utilisé, seules les marques comptent). Ici, les deux informations vont compter, voilà pourquoi je parlerai de balayage double. On peut imaginer de nombreux cas de figure où cette problématique est pertinente : par exemple, un promeneur peut chercher les zones où les cèpes sont les plus nombreux et les plus gros (en supposant que le goût est indépendant de la taille !).

Mathématiquement, on se place dans le même cadre que dans le Chapitre 2 mais l'on va chercher à mettre en défaut l'hypothèse nulle H_0 : "les localisations $s_1, \cdots, s_n$ sont la réalisation d'un processus de Poisson inhomogène d'intensité $\psi(s) = k\phi(s)$ et les marques $X_1, \cdots, X_n$ sont i.i.d. et indépendantes des localisations". On peut remarquer que cette hypothèse nulle combine celles utilisées dans les sections 1 et 2 du Chapitre 2. La mise en place d'une statistique de balayage adaptée à ce problème paraît donc pertinente.

Concernant la famille d'agrégats potentiels, rien ne change par rapport à ce que l'on a vu précédemment au Chapitre 1. Par contre, il est nécessaire d'introduire un indice de concentration qui dépende à la fois des marques et des localisations. Là encore, on peut choisir de construire cet indice de deux manières :

— soit de manière paramétrique, en précisant la distribution des marques sous H_0 et en construisant une hypothèse alternative $H_{1,Z}$ pour chaque agrégat

potentiel Z, afin de pouvoir calculer un rapport de vraisemblance ;
— soit de manière non-paramétrique, par exemple en standardisant la somme des marques (ou la somme des rangs des marques) dans Z par son espérance et sa variance sous H_0.

A priori, la manière paramétrique ne pose pas de problème mathématique insoluble, à partir du moment où les hypothèses H_0 et $H_{1,Z}$ sont correctement spécifiées. Les calculs des estimateurs de vraisemblance sont similaires à ceux rencontrés dans le Chapitre 2. La manière non-paramétrique s'avère plus complexe : le calcul de l'espérance sous H_0 de la somme des marques dans Z nécessite de conditionner par le nombre d'événements qui se produisent dans Z. Idem pour la variance. Une fois ces difficultés surmontées, il faudra comparer les deux approches par une étude de simulation la plus complète possible.

Publications

Voici la liste de mes articles publiés ou soumis auxquels je fais référence dans ce document.

[1] Cucala, L. (2008). A hypothesis-free multiple scan statistic with variable window. Biometrical Journal, 50, p. 299-310.

[2] Cucala, L. (2009). A flexible spatial scan test for case event data. Computational Statistics and Data Analysis, 53, p. 2843-2850.

[3] Dematteï, C. et Cucala, L. (2011). Multiple spatio-temporal cluster detection for case event data : an ordering-based approach. Communications in Statistics – Theory and Methods, 40, p. 358-372.

[4] Cucala, L., Dematteï, C., Lopes, P. et Ribeiro, A. (2013). Spatial scan statistics for case event data based on connected components. Computational Statistics, 28, p. 357-369.

[5] Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes. Spatial Statistics, 10, p. 117-125.

[6] Cucala, L. (2016). A Mann-Whitney scan statistic for continuous data. Communications in Statistics – Theory and Methods, 45, p. 321-329.

[7] Cucala, L. (2016). Scan statistics for detecting high-variance clusters. Journal of Probability and Statistics, Article ID 7591680.

[8] Cucala, L. et Dematteï, C. (2016). Spatial cluster detection for socio-economic data. CSBIGS, 6, p. 1-9.

[9] Cucala, L., Genin, M., Lanier, C. et Occelli, F. (2017). A multivariate Gaussian scan statistic for spatial data. Spatial Statistics, 21, p. 66-74.

[10] 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.

[11] Cucala, L., Genin, M., Occelli, F. et Soula, J. (2019). A multivariate nonparametric scan statistic for spatial data. Spatial Statistics, 29, p. 1-14.

[12] Smida, Z. Cucala, L., Durif, G. et Gannoun, A. (2021). A Wilcoxon-Mann-Whitney spatial scan statistic for functional data. Computational Statistics and Data Analysis, 167, 107378.

[13] Bonneu, F. et Cucala, L. (2021). Global scan methods for comparing two spatial point processes. In : Daouia A., Ruiz-Gazen A. (eds) Advances in Contemporary Statistics and Econometrics. Springer, Cham.

[14] Ahmed, M.S. Cucala, L. et Genin, M. (2021). Spatial autoregressive models for scan statistic. Journal of Spatial Econometrics.

[15] Frévent, C., Ahmed, M.S., Soula, J., Smida, Z., Cucala, L., Dabo-Niang,

S. et Genin, M. (2021). HDSpatialScan : Multivariate and Functional Spatial Scan Statistics.

[16] Cucala, L. (2021). A double scan statistic for continuous data.

Bibliographie

Allard, D. et C. Fraley (1997). "Non parametric maximum likelihood estimation of features in spatial point processes using Voronoï tesselation". In : *Journal of the American Statistical Association* 92, p. 1485-1493.

Anderson, T.W. (2003). *An introduction to multivariate statistical analysis, third edition.* Wiley, New York.

Bar-Hen, A., M. Koskas et N. Picard (2007). *Spatial cluster detection using the number of connected components of a graph.* Rapp. tech. 17. MAP5.

Chakraborty, A. et P. Chaudhuri (2015). "A Wilcoxon-Mann-Whitney type test for infinite-dimensional data". In : *Biometrika* 102, p. 239-246.

Cliff, A. et K. Ord (1973). *Spatial autocorrelation.* Pion Limited, London.

Conover, W. (1980). *Practical nonparametric statistics.* Wiley, New York.

Cressie, N. (1993). *Statistics for spatial data, revised edition.* Wiley, New York.

Cronie, O. et al. (2021). "Functional marked point processes ; a natural structure to unify spatio-temporal frameworks and to analyse dependent functional data". In : *Test.*

Cuevas, A., M. Febrero-Bande et R. Fraiman (2004). "An anova test for functional data". In : *Computational Statistics and Data Analysis* 47, p. 111-122.

David, H.A. (1981). *Order statistics, second edition.* Wiley, New York.

Delicado, P. et al. (2010). "Statistics for spatial functional data : some recent contributions". In : *Environmetrics* 21, p. 224-239.

Dematteï, C., N. Molinari et J.P. Daurès (2007). "Arbitrarily shaped multiple spatial cluster detection for case event data". In : *Computational Statistics and Data Analysis* 51, p. 3931-3945.

Diggle, P.J. et A.G. Chetwynd (1991). "Second-Order analysis of spatial clustering for inhomogeneous populations". In : *Biometrics* 47, p. 1155-1163.

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

Duczmal, L., A. Cançado et al. (2007). "A genetic algorithm for irregularly shaped spatial scan statistics". In : *Computational Statistics and Data Analysis* 52, p. 43-52.

Duczmal, L., M. Kulldorff et L. Huang (2006). "Evaluation of spatial scan statistics for irregularly shaped clusters". In : *Journal of Computational and Graphical Statistics* 15, p. 428-442.

Duczmal, L., G. Moreira et al. (2011). "Voronoi distance based prospective space-time scans for point data sets : a dengue fever cluster analysis in a southeast Brazilian town". In : *International Journal of Health Geographics* 10.29.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses". In : *Annals of Mathematical Statistics* 28, p. 181-187.

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

Frévent, C. et al. (2021). "Detecting spatial clusters on functional data : new scan statistic approaches". In : *Spatial Statistics* 46.100550.

Fuentes-Santos, I., W. González-Manteiga et J. Mateu (2017). "A nonparametric test for the comparison of first-order structures of spatial point processes". In : *Spatial Statistics* 22, p. 240-260.

Gaetan, C., P. Girardi et R. Pastres (2017). "Spatial clustering of curves with an application of satellite data". In : *Spatial Statistics* 20, p. 110-124.

Glaz, J., J. Naus et S. Wallenstein (2001). *Scan statistics*. Springer-Verlag, New York.

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

Hotelling, H. (1931). "The generalization of Student's ratio". In : *Annals of Mathematical Statistics* 2, p. 360-378.

Huffer, F. et C. Lin (1999). "An approach to computations involving spacings with applications to the scan Statistic". In : *Scan statistics and applications (Editors : Glaz J. and Balakrishnan N)*. Birkhäuser, Boston.

— (2001). "Computing the joint distribution of general linear combinations of spacings or exponential variates". In : *Statistica Sinica* 11, p. 1141-1157.

Jacobson, N. (2009). *Basic algebra*. Dover.

Jacques, J. et C. Preda (2013). "Funclust : a curves clustering method using functional random variables density approximation". In : *Neurocomputing* 112, p. 164-171.

Jung, I., M. Kulldorff et A. Klassen (2007). "A spatial scan statistic for ordinal data". In : *Statistics in Medicine* 26, p. 1594-1607.

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

Kulldorff, M. (1997). "A spatial scan statistic". In : *Communications in Statistics. Theory and Methods* 26, p. 1481-1496.

Kulldorff, M., W. Athas et al. (1998). "Evaluating cluster alarms : a space-time scan statistic and brain cancer in Los Alamos". In : *American Journal of Public Health* 88, p. 1377-1380.

Kulldorff, M., R. Heffernan et al. (2005). "A space-time permutation scan statistic for the early detection of disease outbreaks". In : *PLoS Medicine* 2, p. 216-224.

Kulldorff, M., L. Huang et K. Konty (2009). "A scan statistic for continuous data based on the normal probability model". In : *International Journal of Health Geographics* 8.58.

Kulldorff, M., L. Huang, L. Pickle et al. (2006). "An elliptic spatial scan statistic". In : *Statistics in Medicine* 25, p. 3929-3943.

Kulldorff, M., F. Mostashari et al. (2007). "Multivariate spatial scan statistics for disease surveillance". In : *Statistics in Medicine* 26, p. 1824-1833.

Lawson, A. et D. Denison (2002). *Spatial cluster modelling*. Chapman et Hall/CRC, London.

Lee, J., R. Gangnon et J. Zhu (2017). "Cluster detection of spatial regression coefficients". In : *Statistics in medicine* 36, p. 1118-1133.

Lehamnn, E. (1999). *Elements of large sample theory*. Springer.

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

Loh, J. et Z. Zhu (2007). "Accounting for spatial correlation in the scan statistic". In : *Annals of Applied Statistics* 1, p. 560-584.

Mann, H. et D. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other". In : *Annals of Mathematical Statistics* 18, p. 50-60.

Mardia, K., J. Kent et J. Bibby (1979). *Multivariate analysis*. Academic Press.

Mateu, J., G. Lorenzo et E. Porcu (2007). "Detecting features in spatial point processes with clutter via local indicators of spatial association". In : *Journal of Computational and Graphical Statistics* 16, p. 968-990.

Møller, J. et R. Waagepetersen (2003). *Statistical inference and simulation for spatial point processes*. Chapman et Hall/CRC, London.

Nagarwalla, N. (1996). "A scan statistic with a variable window". In : *Statistics in Medicine* 15, p. 845-850.

Naus, J. (1963). "Clustering of random points in the line and plane". Thèse de doct. Harvard University.

Occelli, F., R. Bavdek et al. (2016). "Using lichen biomonitoring to assess environmental justice at a neighbourhood level in an industrial area of Northern France". In : *Ecological Indicators* 60, p. 781-788.

Occelli, F., M.-A. Cuny et al. (2014). "Étude de l'imprégnation de l'environnement de trois bassins de vie de la région Nord-Pas-de-Calais par les éléments traces métalliques. Vers une nouvelle utilisation des données de biosurveillance lichénique." In : *Pollution Atmosphérique* 220, p. 2268-3798.

Oja, H. et R. Randles (2004). "Multivariate nonparametric tests". In : *Statistical Science* 19, p. 598-605.

Patil, G. et C. Taillie (2004). "Upper level set scan statistic for detecting arbitrarily shaped hotspots". In : *Environmental and Ecological Statistics* 11, p. 183-197.

Pyke, R. (1965). "Spacings (with discussion)". In : *Journal of the Royal Statistical Society, Series B* 27, p. 395-449.

Ramsay, J.O. et B.W. Silverman (2005). *Functional data analysis, second edition*. Springer-Verlag, New York.

Randles, R. (1989). "A distribution-free multivariate sign test based on interdirections". In : *Journal of the American Statistical Association* 84, p. 1045-1050.

Saporta, G. (2011). *Probabilités, analyse des données et statistique*. Technip, Paris.

Shelnutt, J. et V. Yao (2005). "A spatial analysis of income inequality in Arkansas at the county level : evidence from tax and commuting data". In : *Regional Economic Development* 1, p. 52-65.

Smida, Z., L. Cucala et A. Gannoun (2022). "A median test for functional data". In : *Journal of Nonparametric Statistics* 34, p. 520-553.

Speakman, S., E. McFowland et D. Neill (2015). "Scalable detection of anomalous patterns with connectivity constraints". In : *Journal of Computational and Graphical Statistics* 24, p. 1014-1033.

Tango, T. et K. Takahashi (2005). "A flexibly shaped spatial scan statistic for detecting clusters". In : *International Journal of Health Geographics* 4.11.

Tyler, D. (1987). "A distribution-free M-estimator of multivariate scatter". In : *Annals of Statistics* 15, p. 234-251.

West, D. (2000). *Introduction to graph theory, second edition*. Prentice-Hall, London.

Wilcoxon, F. (1945). "Individual comparisons by ranking methods". In : *Biometrics Bulletin* 1, p. 80-83.

Wilks, S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In : *Annals of Mathematical Statistics* 9, p. 60-62.

Zhang, Z., M. Kulldorff et R. Assunção (2010). "Spatial scan statistics adjusted for multiple clusters". In : *Journal of Probability and Statistics* 642379.

Zuo, Y. et R. Serfling (2000). "General notions of statistical depth function". In : *Annals of Statistics* 28, p. 461-482.