

Statistiques de scan : théorie et application à l'épidémiologie

Mickaël Genin

▶ **To cite this version:**

Mickaël Genin. Statistiques de scan : théorie et application à l'épidémiologie. Médecine humaine et pathologie. Université du Droit et de la Santé - Lille II, 2013. Français. NNT : 2013LIL2S029 . tel-01004929

HAL Id: tel-01004929

https://theses.hal.science/tel-01004929

Submitted on 11 Jun 2014

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

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

Université Lille 2
Droit et Santé

Université Lille Nord de France
Pôle de recherche et d'Enseignement Supérieur

Numéro d'ordre : XXXX

Année 2013

THESE DE DOCTORAT

présentée par

Michaël GENIN

en vue de l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DROIT ET SANTÉ LILLE 2
DISCIPLINE : MATHÉMATIQUES APPLIQUÉES
SPÉCIALITÉ : STATISTIQUE

STATISTIQUES DE SCAN : THÉORIE ET APPLICATION À L'ÉPIDÉMIOLOGIE

présentée et soutenue publiquement le 03/12/2013 devant le jury composé de :

Président :	Gilbert SAPORTA	Professeur *Conservatoire National des Arts et Métiers de Paris*
Directeurs de thèse :	Alain DUHAMEL	Professeur des Universités - Praticien Hospitalier *Université de Lille 2*
	Cristian PREDA	Professeur des Universités *Université de Lille 1*
Rapporteurs :	Ali GANNOUN	Professeur des Universités *Université Montpellier 2*
	Mustapha RACHDI	Professeur des Universités *Université Pierre-Mendès-France, Grenoble*
Examinateurs :	Mohamed LEMDANI	Professeur des Universités *Université de Lille 2*

Table des matières

Introduction générale **13**

1 Les statistiques de scan **17**
1 Introduction . . . 17
1.1 Vocabulaire . . . 18
1.2 Contenu du chapitre . . . 18
2 Statistique de scan unidimensionnelle . . . 19
2.1 Statistique de scan unidimensionnelle discrète . . . 19
2.2 Statistique de scan unidimensionnelle continue . . . 19
3 Statistique de scan bidimensionnelle . . . 20
3.1 Statistique de scan bidimensionnelle discrète . . . 20
3.2 Statistique de scan bidimensionnelle continue . . . 21
3.3 Forme de la fenêtre . . . 22
4 Test statistique basé sur la statistique de scan . . . 23
4.1 Test basé sur la statistique de scan à fenêtre fixe . . . 23
4.2 Test de la statistique de scan à fenêtre variable . . . 27
5 Conclusion . . . 29

2 Approximation de la distribution des statistiques de scan **31**
1 Introduction . . . 31
2 Cas non-conditionnel . . . 32
2.1 Statistique de scan unidimensionnelle . . . 32
2.2 Statistique de scan bidimensionnelle . . . 36
3 Cas conditionnel . . . 44
3.1 Statistique de scan unidimensionnelle . . . 45
3.2 Statistique de scan bidimensionnelle . . . 49
4 Influence de la forme de la fenêtre . . . 54
4.1 Introduction . . . 54
4.2 Forme de la fenêtre de scan . . . 54
4.3 Applications numériques . . . 57
5 Conclusion . . . 61

3 Statistiques de scan spatiales **63**
1 Introduction . . . 63
2 Données spatiales . . . 64
2.1 Modèle général . . . 64
2.2 Données géostatistiques . . . 64
2.3 Données latticielles . . . 65
2.4 Données ponctuelles . . . 66
3 Cadre général de la statistique de scan spatiale . . . 69
3.1 Phase de détection . . . 70
3.2 Phase d'inférence statistique . . . 72
3.3 Notion de clusters secondaires . . . 73
4 Principaux modèles . . . 74

4.1 Modèle de Bernoulli . . . 74
4.2 Modèle de Poisson . . . 77
5 Limites des statistiques de scan spatiales . . . 81
5.1 Temps de calcul . . . 81
5.2 Forme de la fenêtre . . . 86
6 Une alternative à la méthode de Monte Carlo pour le TRVG . . . 86
6.1 Méthodologie . . . 86
6.2 Etude de simulation . . . 88
6.3 Résultats . . . 90
7 Conclusion . . . 92

4 Application à l'étude de la maladie de Crohn **93**
1 Introduction . . . 93
2 Matériels et méthodes . . . 94
2.1 Données . . . 94
2.2 Méthodes statistiques . . . 95
3 Résultats . . . 96
3.1 Variations spatiales et spatio-temporelles de l'incidence de MC . . . 96
3.2 Analyse de l'offre de soin en gastro-entérologie . . . 99
4 Discussion . . . 99
5 Conclusion . . . 100

Conclusion générale et perspectives **101**
1 Conclusion générale . . . 101
2 Perspectives . . . 102

Annexes **103**

A Test basé sur la statistique de scan : cas bidimensionnel **103**
1 Modèle binomial . . . 103
2 Modèle de Poisson . . . 104

B Formule d'interpolation linéaire **107**

C Forme de la fenêtre : applications numériques supplémentaires **109**

Bibliographie **117**

Table des figures

1.1 Statistique de scan continue . . . 18
1.2 Statistique de scan discrète . . . 18
1.3 Région rectangulaire discrète (a) de taille $[0, N_1] \times [0, N_2]$ et continue (b) de taille $[0, L] \times [0, K]$ . . . 20
1.4 (a) Processus de scan avec une fenêtre $m_1 \times m_2$ sur une région rectangulaire discrète de taille $[0, N_1] \times [0, N_2]$. (b) Processus de scan avec une fenêtre $u \times v$ sur une région rectangulaire continue de taille $[0, L] \times [0, K]$ . . . 21
1.5 Exemples de fenêtres de scan de formes discrètes différentes. . . . 22

2.1 Exemple de E_j . . . 32
2.2 Exemple de Z_j . . . 33
2.3 Exemple de Z_{l_1} . . . 38
2.4 Q_2 et Q_3 . . . 39
2.5 Q_{22}, Q_{23}, Q_{32} and Q_{33} . . . 39
2.6 Configurations possibles de formes rectangulaires pour $A = 16$. . . . 55
2.7 Cercle discret dans $\mathbb{Z}^2$. . . . 56
2.8 Modèle de Bernoulli $\mathcal{B}(1, 0.01)$: estimation de la puissance du test basé sur la statistique de scan. . . . 60
2.9 Modèle binomial $\mathcal{B}(5, 0.05)$: estimation de la puissance du test basé sur la statistique de scan. . . . 60
2.10 Modèle de Poisson $\mathcal{P}(0.25)$: estimation de la puissance du test basé sur la statistique de scan. . . . 60

3.1 (a) Cumuls pluviométriques sur le réseau météorologique suisse le 8 mai 1986 (passage du nuage de Chernobyl, données sic du package geoR du logiciel R, (b) Porosité d'un sol (données soil du package geoR) . . . 65
3.2 (a) Pourcentage de la population présentant le groupe sanguin A dans les 26 comtés d'Irlande, (b) Graphe de voisinage $\mathcal{G}$ des 26 comtés . . . 66
3.3 (a) Réalisation d'un processus de Poisson homogène sur $D = [0, 1]^2$, intensité $\lambda = 200$; (b) Réalisation d'un processus de Poisson hétérogène sur $D = [0, 1]^2$ d'intensité $\lambda(x, y) = 400 * (\sin(24x) + \sin(24y))$. . . . 69
3.4 Cas spatial - Ensemble $\mathcal{Z}$ de clusters potentiels . . . 71
3.5 Cas spatio-temporel - Ensemble $\mathcal{Z}$ de clusters potentiels. . . . 71
3.6 Deux modélisations possibles. . . . 87
3.7 Données simulées sur 245 comtés du Nord-Est des Etats-Unis. . . . 88
3.8 Croisement des probabilités issues de la méthode de Monte Carlo et celle issues de la modélisation 1. . . . 91

4.1 Départements couverts par le registre EPIMAD. . . . 94
4.2 Rapport standardisés d'incidence lissés de MC, ajustés sur le sexe et l'âge au diagnostic, de 1990 à 2006, au sein de la zone géographique couverte par le registre EPIMAD dans le nord de la France. . . . 96
4.3 Risques relatifs de MC dans le nord de la France pendant la période 1990 - 2006, clusters spatiaux constants pendant l'ensemble de la période étudiée. . . . 97

4.4 Risques relatifs de MC dans le nord de la France pendant la période 1990 - 2006, clusters spatiaux non-constants pendant l'ensemble de la période étudiée. 97

B.1 Approximation de p^* par interpolation linéaire. 107

Liste des tableaux

2.1 Approximations pour $\mathbb{P}(S(m,N) \leq k)$ par App. (2.6) de Haiman (2007) et App. (2.2) de Naus (1982), $X_i \sim \mathcal{B}(1,p)$, $p = 0.1$, $m = 30$. 35
2.2 Approximations pour $\mathbb{P}(S \leq k)$ par app. (2.6). $T = 1001$. 36
2.3 Approximation pour $\mathbb{P}(S \leq k)$: $X_{ij} \sim \mathcal{P}(0.25)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $N = 10^9$. 41
2.4 Approximation pour $\mathcal{P}(S \leq k)$: $X_{ij} \sim \mathcal{B}(5, 0.05)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $N = 10^9$. 41
2.5 Approximation de $\mathbb{P}(S \leq k)$ en utilisant l'approximation de Alm (2.25) et celle d'Haiman (2.27). $L = 500$, $K = 500$ et $\lambda = 0.01$ 44
2.6 Récapitulatif des différentes formules exactes et approximations de la distribution des différentes statistiques de scan. Il est important de préciser que les formules exactes citées dans ce tableau correspondent aux quantités Q_2 et Q_3 pour le cas unidimensionnel, qui sont utilisées par la suite dans les approximations. Par ailleurs, dans le cas bidimensionnel continu conditionnel, l'approximation proposée dans [Naus, 1966] est constituée du produit de deux distributions de la statistique de scan unidimensionnelle continue conditionnelle. 53
2.7 Comparaison de surface entre formes carrées et circulaires avec un rayon optimal R. 56
2.8 Approximations de $\mathbb{P}(S \leqslant n)$: Scan d'une région $N_1 \times N_2 = 42 \times 42$ avec différentes formes de fenêtres. 58

3.1 Temps de calcul du logiciel SaTScan© en fonction du nombre de centres et du nombre de simulations de Monte Carlo - *CPU : 2.2Ghz Intel Core I7 - RAM : 8 Go 1333 Mhz DDR3* 82
3.2 Valeur de $\mathbb{P}_{rejet\ \mathcal{H}_0}$ en fonction de π pour $\alpha = 0.05$ 84
3.3 Valeur de $\mathbb{P}_{rejet\ \mathcal{H}_0}$ en fonction de π pour $\alpha = 0.01$ 84
3.4 Valeurs de $e^D_{R,\alpha}$ pour différentes valeurs de α et de R 85
3.5 Comparaison de puissance entre l'approximation basée sur les simulations de Monte Carlo et l'approximation d'Haiman. 92

4.1 Description des clusters de MC issus des analyses spatio-temporelles, nord de la France, 1990-2006. 98

C.1 Approximations de $\mathbb{P}(S \leqslant n)$: Scan d'une région $N_1 \times N_2 = 42 \times 42$ avec différentes formes de fenêtres. 109

A Fany

Remerciements

Je voudrais tout d'abord exprimer mes sincères remerciements à Cristian Preda pour avoir accepter de co-diriger cette thèse. Successivement maître de stage de fin d'études d'ingénieur, collègue et directeur de thèse, il m'a indéniablement transmis le virus de la recherche. Il a su me faire partager ses connaissances scientifiques, son goût du détail et sa rigueur mathématique qui a indubitablement conduit à l'amélioration de mes connaissances dans ce domaine. Ses qualités humaines et scientifiques ont rendu nos sessions de travail hebdomadaires très agréables et fertiles sur plan statistique. Pour m'avoir fait confiance, je lui dit : *multumesc*.

Je tiens à remercier chaleureusement Alain Duhamel de m'avoir accueilli au sein de l'EA 2694 et d'avoir co-dirigé cette thèse. Je lui suis très reconnaissant de la confiance qu'il m'a accordée avant et pendant ces trois années. Ses précieux conseils en biostatistique et en épidémiologie ont contribué fortement à la qualité de cette thèse.

Je tiens ensuite à remercier Ali Gannoun et Mustapha Rachdi d'avoir accepté d'être les rapporteurs de cette thèse. Je suis flatté de l'intérêt qu'ils ont porté à ce travail.

Je voudrais exprimer mes remerciements à Gilbert Saporta d'avoir accepté de faire partie de mon jury. Sachant que ses ouvrages de statistiques m'accompagnent depuis le début de mes études en statistique, c'est un honneur de le compter parmi les membres du jury. Je tiens à remercier également Mohammed Lemdani pour avoir accepté de faire partie du jury ainsi que pour ses remarques aussi pertinentes que précises qui ont permis d'améliorer nettement la qualité de ce document.

Je souhaite remercier le professeur Régis Beuscart de m'avoir accueilli au sein du Centre d'Etudes et de Recherche en Informatique Médicale. Je tiens à remercier l'ensemble de l'équipe du CERIM et plus particulièrement Julien et Renaud, pour leur disponibilité, leur bonne humeur et leurs précieux conseils en informatique.

Je tiens également à remercier mes deux "colocataires" de bureau : Jean-Baptiste Beuscart et Mohammed Ben Hadj Yahia. Tous les trois issus de formations différentes, nos échanges ont été très instructifs. D'un point de vue scientifique, vous m'avez fait comprendre l'intérêt de la pluridisciplinarité en recherche médicale. D'un point de vue humain, ce fût un véritable plaisir de partager ce bureau avec vous.

Je souhaite adresser mes remerciements aux membres du Pôle de Santé Publique du CHRU de Lille notamment à Julia Salleron pour nos longs échanges en biostatistique et recherche clinique, à Brigitte Bonneau pour sa bonne humeur légendaire et au docteur Corinne Gower-Rousseau pour avoir mis à ma disposition les données du registre EPIMAD.

Je tiens à remercier Franck Gauzere, Armand Maul et Daniel Vagost, professeurs à l'Université de Metz, pour m'avoir transmis la passion de la statistique. Vos qualités scientifiques, pédagogiques et humaines ont fortement contribué à cette vocation.

Je souhaite également remercier mes amis qui ont su être présents dans les bons comme les mauvais moments : merci aux Ben, Freak, Julie, Davis, Elo, Mikado, Nicox et autres ...

Je voudrais remercier toute ma famille et belle-famille, et notamment mes parents qui m'ont toujours soutenu lors de mes études et m'ont permis de les réaliser dans les meilleures conditions possibles. Je vous dois énormément...

Enfin, je tiens à remercier tout particulièrement Fany pour tout ce qu'elle m'apporte. Ton Amour, ton soutien indéfectible et tes encouragements ont contribué pour beaucoup dans la réalisation de cette thèse. Malgré les difficultés, tu as su et sais toujours me faire voir la vie en *rose*. Je te dédie cette thèse.

Introduction générale

La notion de cluster désigne l'agrégation dans le temps et/ou l'espace d'évènements. Dans de nombreux domaines, les experts observent certaines agrégations d'évènements et la question se pose de savoir si ces agrégations peuvent être considérées comme normales (le fruit du hasard) ou non. D'un point de vue probabiliste, la normalité peut être décrite par une hypothèse nulle de répartition aléatoire des évènements. Par exemple, les experts en santé publique cherchent à déterminer les facteurs causaux de clusters temporels ou spatiaux de cancer. Sous l'hypothèse nulle stipulant que les cas de cancer sont répartis de manière uniforme dans une région, les chercheurs peuvent s'interroger sur la probabilité d'observation d'un cluster de cas de cancer. Si cette probabilité est faible, autrement dit, si le nombre de cas observés dans le cluster diffère nettement de celui qui serait observé dans une situation normale (décrite par l'hypothèse nulle) alors il est intéressant d'entreprendre des investigations. De surcroît, lorsque l'étiologie d'une maladie est peu connue voire inconnue, la mise en évidence de clusters de cas de maladie peut aider à identifier des facteurs d'expositions environnementaux et/ou des facteurs génétiques.

La détection de clusters d'évènements est un domaine de la statistique qui s'est particulièrement étendu au cours des dernières décennies. En premier lieu, la communauté scientifique s'est attachée à développer des méthodes dans le cadre unidimensionnel (ex : le temps) puis, par la suite, a étendu ces méthodes au cas multidimensionnel, et notamment bidimensionnel (l'espace). Parmi l'ensemble des méthodes de détection de clusters d'évènements, trois grands types de tests peuvent être distingués. Le premier concerne les tests globaux qui permettent de détecter une tendance globale à l'agrégation, sans pour autant localiser les clusters éventuels. Le deuxième type correspond aux tests focalisés qui sont utilisés lorsque des connaissances *a priori* permettent de définir un point source (date ou localisation spatiale) et de tester l'agrégation autour de ce dernier. Le troisième type englobe les tests de détection de cluster (ou sans point source défini) qui permettent la localisation, sans connaissance *a priori*, de clusters d'évènements et le test de leur significativité statistique. Au sein de cette thèse, nous nous sommes focalisés sur cette dernière catégorie et plus particulièrement aux méthodes basées sur les statistiques de scan (ou de balayage).

Apparues au début des années 1960 [Naus, 1965b], ces méthodes permettent de détecter des clusters d'évènements et de déterminer leur aspect "normal" (le fruit du hasard) ou "anormal". L'étape de détection est réalisée par le balayage (scan) par une fenêtre, dite fenêtre de scan, du domaine d'étude (discret ou continu) dans lequel sont observés les évènements (ex : le temps, l'espace, ...). Cette phase de détection conduit à un ensemble de fenêtres définissant chacune un cluster potentiel. Une statistique de scan est une variable aléatoire définie comme la fenêtre comportant le nombre maximum d'évènements observés.

Les statistiques de scan sont utilisées comme statistique de test pour vérifier l'indépendance et l'appartenance à une distribution donnée des observations, contre une hypothèse alternative privilégiant l'existence de cluster au sein d'une de la région étudiée. Par ailleurs, la principale difficulté réside dans la détermination de la distribution, sous l'hypothèse nulle, de la statistique de scan. En effet, puisqu'elle est définie comme le maximum d'une suite de variables aléatoires dépendantes, la dépendance étant due au recouvrement des différentes fenêtres de scan, il n'existe que dans de très rares cas de figure des solutions explicites. Aussi, un pan de la littérature est axé sur le développement de méthodes (formules exactes et surtout approximations) permettant de déterminer la distribution des statistiques de scan. Par ailleurs, dans le cadre bidimensionnel,

la fenêtre de scan peut prendre différentes formes géométriques (rectangulaire, circulaire, ...) qui pourraient avoir une influence sur l'approximation de la distribution de la statistique de scan. Cependant, à notre connaissance, aucune étude n'a évalué cette influence dans le cas discret.

Dans le cadre spatial, les statistiques de scan spatiales développées par [Kulldorff and Nagarwalla, 1995] et [Kulldorff, 1997] s'imposent comme étant, de loin, les méthodes les plus utilisées pour la détection de clusters spatiaux. Le principe de ces méthodes réside dans le fait de scanner le domaine d'étude avec des fenêtres de forme circulaire et de sélectionner le cluster le plus probable comme celui maximisant la statistique un test de rapport de vraisemblance. L'inférence statistique de ce cluster est réalisée par le biais de simulations de Monte Carlo. Or, dans le cas de bases de données de taille élevée et/ou lorsqu'une précision importante de la probabilité critique associée au cluster détecté est requise, les simulations de Monte Carlo conduisent à des temps de calculs extrêmement importants.

Ce manuscrit est divisé en 4 chapitres :

Chapitre 1. Ce chapitre s'attache, dans un premier temps, à définir les statistiques de scan unidimensionnelles et bidimensionnelles, discrètes et continues. Dans un deuxième temps, nous définissons le test statistique basé sur la statistique de scan afin de détecter un cluster d'évènements. Nous rappelons les résultats issus de [Naus, 1966] dans le cas continu. Ces derniers stipulent que la statistique de test d'un test de rapport de vraisemblance généralisé, visant à vérifier l'hypothèse nulle de répartition aléatoire des évènements contre un hypothèse alternative supportant l'existence de cluster, est une fonction monotone croissante de la statistique de scan. Ces résultats, considérés comme acquis dans le cas discret, n'ont, à notre connaissance, jamais été formalisés dans la littérature. Aussi, nous explicitons, sous forme de propositions, l'application de ces résultats dans les cas discrets uni et bidimensionnels.

Cette étude fait l'objet d'un article accepté dans *Romanian Journal of Pure and Applied Mathematics.*

Chapitre 2. Nous décrivons les différentes formules exactes et approximations de la distribution des statistiques de scan dans les cas non-conditionnel et conditionnel, faisant, pour chaque cas, la distinction entre les cas unidimensionnel et bidimensionnel, discret et continu. Par la suite nous évaluons l'influence de la forme de la fenêtre de scan sur la distribution des statistiques de scan bidimensionnelles discrètes. Nous avons réalisé une étude de simulation prenant en compte des fenêtres de scan de forme carrée, rectangulaire et circulaire (cercle discret). Cette étude a mis en évidence le fait que les distributions des statistiques de scan associées à ces formes sont très proches les unes des autres mais significativement différentes. Par ailleurs, nous mettons en évidence, par le biais d'une étude de simulation, que la puissance du test basé sur les statistiques de scan est liée à la forme de la fenêtre ainsi qu'à la forme du cluster existant sous l'hypothèse alternative.

Cette étude fait l'objet d'un article soumis à *Statistics and Probability Letters.*

Chapitre 3. Après une brève définition des différents types de données spatiales, ce chapitre décrit les méthodes de statistiques de scan spatiales, leur extension spatio-temporelle, ainsi que les principaux modèles (Bernoulli et Poisson). Par la suite, nous proposons une alternative à la méthode de Monte Carlo pour l'estimation de la distribution de la statistique de test. Notre méthodologie consiste à conserver la phase de détection qui conduit à la sélection d'un cluster le plus probable et de transposer le problème à l'approximation de la distribution de la statistique de scan unidimensionnelle discrète. Cette modélisation permet de réduire de manière importante les temps de calcul tout en proposant une erreur d'approximation de la distribution. Nous évaluons notre méthode par le biais de données simulées : les temps de calcul sont fortement réduits et la puissance statistique du test est supérieure à celle du test basé sur les simulations de Monte Carlo.

Chapitre 4. Dans ce chapitre, nous présentons une application des statistiques de scan spatio-temporelles à l'étude de la maladie de Crohn (MC) dans le nord de la France, de 1990 à 2006. Les données sont issues du registre Inserm-INVS EPIMAD qui répertorie de manière exhaustive

les cas de Maladies Inflammatoires Chroniques de l'Intestin (MICI), dont la MC fait partie. Cette pathologie chronique, dont l'étiologie est inconnue, n'est pas mortelle mais altère significativement la vie des patients. L'objectif de l'étude était de mettre en évidence des clusters significatifs de cas de MC dans le temps et l'espace afin de, par la suite, émettre des hypothèses sur les causes de MC, notamment environnementales. Nous isolons 18 clusters significatifs de deux types : 14 clusters spatiaux constants dans le temps et 4 clusters spatiaux non-constants dans le temps. Parmi les 14 clusters du premier type, 5 clusters de sur-incidence et 4 clusters de sous-incidence ont été détectés. Parmi les 4 clusters du deuxième type, 3 clusters de sur-incidence et 1 cluster de sous-incidence ont été détectés.

Cette étude a fait l'objet d'un article accepté dans *Journal of Public Health*.

Un conclusion générale ainsi que des perspectives viennent terminer ce manuscrit.

Chapitre 1

Les statistiques de scan

Sommaire

1 Introduction 17
1.1 Vocabulaire 18
1.2 Contenu du chapitre 18
2 Statistique de scan unidimensionnelle 19
2.1 Statistique de scan unidimensionnelle discrète 19
2.2 Statistique de scan unidimensionnelle continue 19
3 Statistique de scan bidimensionnelle 20
3.1 Statistique de scan bidimensionnelle discrète 20
3.2 Statistique de scan bidimensionnelle continue 21
3.3 Forme de la fenêtre 22
4 Test statistique basé sur la statistique de scan 23
4.1 Test basé sur la statistique de scan à fenêtre fixe 23
4.2 Test de la statistique de scan à fenêtre variable 27
5 Conclusion 29

1 Introduction

Dans beaucoup de domaines il faut décider si une certaine accumulation d'événements est "normale" ou pas. Dans l'affirmative, on peut supposer la présence d'un ensemble de facteurs de risque qui doivent par la suite être contrôlés. En santé publique, les services d'épidémiologie cherchent les facteurs pouvant expliquer des clusters de cancers ou d'anomalies de naissance. Les biologistes cherchent des clusters de palindromes dans les séquences d'ADN pour trouver des indices sur l'origine de la réplication de certains virus. En contrôle de qualité, on s'interroge sur les clusters de produits défectueux.

La décision est alors prise selon la grandeur de la probabilité d'observer un tel cluster, l'hypothèse nulle étant celle d'une situation (évolution) "normale". Si cette probabilité est petite, il est légitime de supposer la présence d'un écart par rapport à la situation normale et alors des décisions doivent être prises.

Les statistiques de scan (scan statistics) sont utilisées pour statuer du caractère exceptionnel ou non de l'observation d'un cluster d'événements. Plus précisément, ce sont des variables aléatoires utilisées comme statistiques de test pour vérifier l'hypothèse d'indépendance et l'appartenance à une distribution donnée des observations, contre une alternative privilégiant l'existence de clusters. De nombreux travaux sont consacrés à ce sujet. Mentionnons, par exemple, les monographies de [Glaz and Balakrishnan, 1999], [Balakrishnan and Koutras, 2001], [Glaz et al., 2001], [Fu and Lou, 2003] et plus récemment [Glaz et al., 2009].

1.1 Vocabulaire

Discret *vs* Continu. Il existe deux types de statistiques de scan : d'une part les statistiques de scan continues et d'autre part les statistiques de scan discrètes. La différence entre ces deux types réside dans la nature de la distribution régissant la survenue des évènements. En effet, si les évènements surviennent, par exemple, dans un intervalle de temps continu $[0, T]$, alors les statistiques de scan continues seront appropriées (Figure 1.1). *A contrario*, si les évènements surviennent dans un intervalle de temps discret $\{1, \ldots, T\}$ alors on parle de statistique de scan discrète (Figure 1.2).

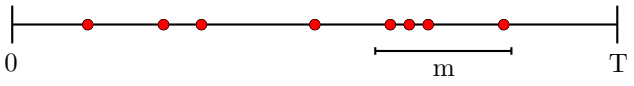


FIGURE 1.1 – Statistique de scan continue

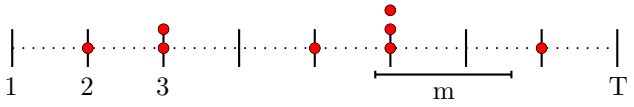


FIGURE 1.2 – Statistique de scan discrète

Conditionnel *vs* Non-conditionnel. Quelle que soit la nature de la statistique de scan (discrète ou continue), deux cas peuvent être distingués : le cas conditionnel et le cas non-conditionnel. Dans le premier, le nombre total d'évènements observés est connu et fixé. On parle de statistique de scan *conditionnelle* ou encore *statistique de scan rétrospective*. Par exemple, lors d'une étude rétrospective renseignant le nombre de cas de cancers dans une région donnée, les statistiques de scan conditionnelles seront appropriées.
Dans le deuxième cas le nombre total d'évènements observés est inconnu et considéré comme une variable aléatoire dont la loi de probabilité est connue. On parle de *statistique de scan non-conditionnelle* ou encore *statistique de scan prospective*. Ce type de statistique de scan peut, par exemple, être utilisé dans le cadre de système de surveillance journalier ou hebdomadaire de cas de grippe dans une région donnée. Le nombre de cas de grippe n'étant pas connu à l'avance, nous sommes bien en présence d'un système prospectif.

Différentes dimensions. Les statistiques de scan peuvent à la fois s'appliquer à une dimension ou à plusieurs dimensions. En effet, lorsque les évènements surviennent pendant une période de temps, la *statistique de scan unidimensionnelle* sera appropriée. Si les évènements sont distribués sur un sous-ensemble de $\mathbb{R}^2$ (discret ou continu) alors on parle de *statistique de scan bidimensionnelle*. Par ailleurs, il est possible d'appliquer la statistique de scan dans les cas où plus de 2 dimensions sont présentes (cas tri-dimensionnel ou plus). A ce propos, le cas tridimensionnel sera explicité, au sein du chapitre 3, dans le cadre des statistiques de scan spatio-temporelles.

1.2 Contenu du chapitre

Dans un premier temps, ce chapitre s'attache à définir les statistiques de scan, dans les cas unidimensionnel et bidimensionnel, discret et continu. De surcroît, seront distingués les cas conditionnel et non-conditionnel. Dans un second temps, ce chapitre décrit l'utilisation des statistiques de scan comme statistique de test permettant de vérifier l'indépendance et l'appartenance à un distribution des observations contre une hypothèse alternative supportant l'existence de cluster. A

ce propos, nous explicitons une série de propositions portant sur le rôle de la statistique de scan dans un test de rapport de vraisemblance généralisé dans les cas uni et bidimensionnels discrets.

2 Statistique de scan unidimensionnelle

2.1 Statistique de scan unidimensionnelle discrète

Soit $\{X_1, X_2, \ldots, X_N\}$ une suite de variables aléatoires indépendantes et identiquement distribuées à valeurs dans $\mathbb{N}$. La variable aléatoire X_i, $1 \leq i \leq N$, désigne le nombre d'évènements associés à l'expérience aléatoire indexée par i. Ces variables aléatoires peuvent être distribuées selon une loi de Bernoulli $\mathcal{B}(1,p)$, une loi binomiale $\mathcal{B}(n,p)$ ou encore une loi de Poisson $\mathcal{P}(\lambda)$.

Pour N fixé, $N \in \mathbb{N}^*$ et une fenêtre de taille $m \in \mathbb{N}^*$, $m < N$, définissons

$$\nu_t = \sum_{i=t}^{t+m-1} X_i,\ 1 \leq t \leq N-m+1. \tag{1.1}$$

La variable aléatoire ν_t correspond au nombre d'évènements observés à la suite de m expériences aléatoires consécutives de t à $t+m-1$.

[Naus, 1965b] a défini la statistique de scan unidimensionnelle discrète :

Définition 1.1. On appelle **statistique de scan unidimensionnelle discrète** associée à la suite de variables aléatoires *i.i.d.* $\{X_1, X_2, \ldots, X_N\}$ et à une fenêtre de taille m, la variable aléatoire définie par :

$$S(m,N) = \max_{1 \leq t \leq N-m+1} \nu_t. \tag{1.2}$$

$S(m,N)$ désigne le nombre maximal d'évènements de la suite $\{X_1, X_2, \ldots, X_N\}$ observé dans une fenêtre mobile de taille m.

Définition 1.2. On appelle **statistique de scan unidimensionnelle discrète conditionnelle** associée à la suite de variables aléatoires *i.i.d.* $\{X_1, X_2, \ldots, X_N\}$, à une fenêtre de taille m et au nombre total d'évènements observés $\sum_{i=1}^{N} X_i = a$, la variable aléatoire notée $S(m,N,a)$.

2.2 Statistique de scan unidimensionnelle continue

Soit $N = \{N_t\}_{t \geq 0}$ un processus ponctuel défini sur l'intervalle $[0,T]$, $T \in \mathbb{R}^+$ fixé. En pratique, la plupart du temps, N est un processus de Poisson d'intensité $\lambda > 0$. Considérons $m \in \mathbb{R}_+^*$, $m < T$ et ν_t la variable aléatoire définie par :

$$\nu_t = N(t+m) - N(t),\ t \in [0, T-m]. \tag{1.3}$$

La variable aléatoire ν_t correspond au nombre de sauts (évènements) du processus N dans l'intervalle $[t; t+m[$.

[Naus, 1966] a défini la statistique de scan unidimensionnelle continue :

Définition 1.3. On appelle **statistique de scan unidimensionnelle continue** associée au processus N, à l'intervalle $[0,T]$ et à une fenêtre de taille m, la variable aléatoire définie par :

$$S(m,T) = \max_{t \in [0, T-m]} \nu_t. \tag{1.4}$$

$S(m,T)$ désigne le nombre maximal d'évènements du processus N observé au sein d'une fenêtre mobile de taille m se déplaçant de manière continue sur l'intervalle $[0,T]$.

Définition 1.4. On appelle **statistique de scan unidimensionnelle continue conditionnelle** associée au processus N, à l'intervalle $[0,T]$, à une fenêtre de taille m et un nombre total d'évènements observés sur $[0,T]$ $N(T)=a$, la variable aléatoire notée $S(m,T,a)$.

L'étude de la distribution de la statistique de scan unidimensionnelle S fait l'objet de nombreux travaux. Dans [Huntington and Naus, 1975], les auteurs donnent une formule exacte pour $\mathbb{P}(S \leq k)$, $k \geq 0$, qui devient vite inexploitable (temps de calcul excessif) dès que T est grand par rapport à m. Dans [Neff and Naus, 1980], les auteurs établissent des tables pour la distribution de S pour plusieurs valeurs de λ et de T. [Janson, 1984] fournit des bornes inférieures et supérieures pour $\mathbb{P}(S \leq k)$. Ces méthodes font l'objet d'une étude approfondie au sein du chapitre 2.

3 Statistique de scan bidimensionnelle

3.1 Statistique de scan bidimensionnelle discrète

Soit une région rectangulaire $[0,N_1]\times[0,N_2]$ avec N_1, $N_2 \in \mathbb{N}$. Considérons $\{X_{i,j}\}$, $1 \leq i \leq N_1$ et $1 \leq j \leq N_2$ un ensemble de variables aléatoires indépendantes, identiquement distribuées et à valeurs dans $\mathbb{N}$. Chaque variable aléatoire $X_{i,j}$ désigne le nombre d'évènements observés dans une région élémentaire de dimension $[i-1,i]\times[j-1,j]$ (Figure 1.3(a)). Comme dans cas unidimensionnel discret, les $X_{i,j}$ peuvent être distribuées selon une loi de Bernoulli $\mathcal{B}(1,p)$, une loi binomiale $\mathcal{B}(n,p)$ ou une loi de Poisson $\mathcal{P}(\lambda)$.

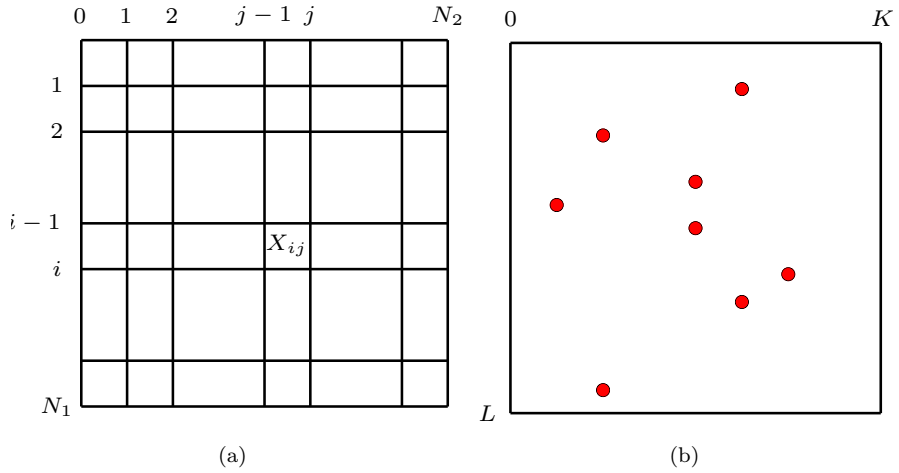


FIGURE 1.3 – Région rectangulaire discrète (a) de taille $[0,N_1]\times[0,N_2]$ et continue (b) de taille $[0,L]\times[0,K]$

Soient m_1, $m_2 \in \mathbb{N}^*$, $m_1 < N_1$, $m_2 < N_2$. Considérons une fenêtre rectangulaire de dimension $m_1 \times m_2$. Pour $1 \leq t \leq N_1 - m_1 + 1$ et $1 \leq s \leq N_2 - m_2 + 1$, définissons la variable aléatoire $\nu_{t,s}$ telle que

$$\nu_{t,s} = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} X_{ij}. \tag{1.5}$$

La variable aléatoire $\nu_{t,s}$ désigne le nombre d'évènements observés dans la fenêtre $[t;t+m_1]\times[s;s+m_2]$ (Figure 1.4(a)).

Définition 1.5. On appelle **statistique de scan bidimensionnelle discrète** associée à l'ensemble de variables aléatoires $\{X_{i,j}\}$, à une région rectangulaire $[0,N_1]\times[0,N_2]$ et à une fenêtre de dimension $m_1 \times m_2$, la variable aléatoire définie par

$$S(m_1,m_2,N_1,N_2) = \max_{\substack{1 \leq t \leq N_1 - m_1 + 1 \\ 1 \leq s \leq N_2 - m_2 + 1}} \nu_{t,s}. \tag{1.6}$$

$S(m_1, m_2, N_1, N_2)$ désigne le nombre maximal d'évènements de l'ensemble $\{X_{i,j}\}$ observé au sein d'une fenêtre de dimension $m_1 \times m_2$ se déplaçant dans la région rectangulaire discrète $[0, N_1] \times [0, N_2]$.

Définition 1.6. On appelle **statistique de scan bidimensionnelle discrète conditionnelle** associée à l'ensemble de variables aléatoires $\{X_{i,j}\}$, à une région rectangulaire $[0, N_1] \times [0, N_2]$, une fenêtre de dimension $m_1 \times m_2$ et au nombre total d'évènements observés sur $[0, N_1] \times [0, N_2]$ $\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} X_{ij} = a$, la variable aléatoire définie notée $S(m_1, m_2, N_1, N_2, a)$.

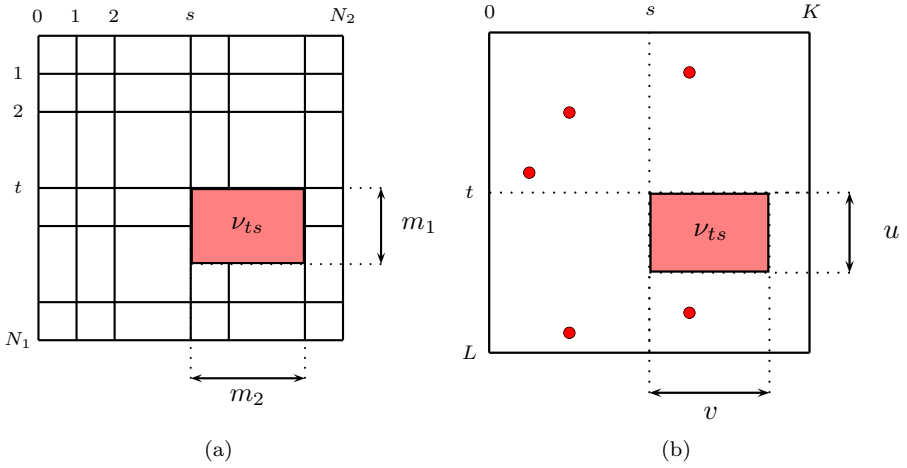


FIGURE 1.4 – (a) Processus de scan avec une fenêtre $m_1 \times m_2$ sur une région rectangulaire discrète de taille $[0, N_1] \times [0, N_2]$. (b) Processus de scan avec une fenêtre $u \times v$ sur une région rectangulaire continue de taille $[0, L] \times [0, K]$

3.2 Statistique de scan bidimensionnelle continue

Soit une région rectangulaire $[0, L] \times [0, K]$, $(L, K) \in \mathbb{R}^2$ et $L, K < \infty$ (Figure 1.3(b)). Soit $N = \{N_{(t,s)}, t \in [0, L], s \in [0, K]\}$ un processus spatial ponctuel. En pratique, et la plupart du temps, N est un processus ponctuel spatial de Poisson d'intensité $\lambda > 0$. Soient $\{u, v\} \in \mathbb{R}^{2*}$, $u < L$ et $v < K$. Considérons une fenêtre rectangulaire de dimension $u \times v$. Pour $t \in [0, L - u]$ et chaque $s \in [0, K - v]$, définissons la variable aléatoire $\nu_{t,s}$ telle que

$$\nu_{t,s} = N\left([t; t + u] \times [s; s + v]\right). \tag{1.7}$$

La variable aléatoire $\nu_{t,s}$ désigne le nombre d'évènements observés dans la fenêtre rectangulaire $[t; t + u] \times [s; s + v]$ (Figure 1.4(b)).

Définition 1.7. On appelle **statistique de scan bidimensionnelle continue** associée au processus N, à la région $[0, L] \times [0, K]$ et à une fenêtre de dimension $u \times v$, la variable aléatoire définie par

$$S(u, v, L, K) = \max_{\substack{0 \le t \le L - u \\ 0 \le s \le K - v}} \nu_{t,s}. \tag{1.8}$$

$S(u, v, L, K)$ désigne le nombre maximal d'évènements du processus N observé dans une fenêtre de taille $u \times v$ se déplaçant de manière continue dans la région $[0, L] \times [0, K]$.

Définition 1.8. On appelle **statistique de scan bidimensionnelle continue conditionnelle** associée au processus N, à la région $[0, L] \times [0, K]$, à une fenêtre de dimension $u \times v$ et au nombre total d'évènements observés sur $[0, L] \times [0, K]$, $N([0, L] \times [0, K]) = a$, la variable aléatoire notée $S(u, v, L, K, a)$.

La distribution de la statistique de scan bidimensionnelle S a fait l'objet de nombreux travaux, que ce soit dans le cas continu comme dans le cas discret, mais ne menant pas à des formules exactes. Dans le premier cas, des approximations basées sur des heuristiques sont proposées dans [Aldous, 1989] et [Alm, 1997]. Dans [Haiman and Preda, 2002], les auteurs proposent une approximation basée sur les propriétés du maximum partiel d'une suite stationnaire de variables aléatoires 1-dépendantes. Dans le cas discret, [Chen and Glaz, 1996] proposent des approximations de la distribution de S basées sur un comportement de type markovien et [Haiman and Preda, 2006] proposent une approximation également basée sur les propriétés du maximum partiel d'une suite de variables aléatoires 1-dépendantes. Ces méthodes d'approximation feront l'objet d'une étude approfondie dans le chapitre 2.

3.3 Forme de la fenêtre

Initialement, les statistiques de scan bidimensionnelles ont été définies avec une fenêtre de forme rectangulaire et de taille fixe [Naus, 1965a]. Il apparaît évident que cette restriction a une influence sur la capacité de détection du "vrai" cluster dans les données, surtout si ce dernier n'est soit pas de la même taille que la fenêtre soit pas de forme rectangulaire. Aussi, plusieurs auteurs ont proposé, dans le cas continu comme dans le cas discret, d'autres formes possibles de fenêtres, de tailles fixes ou variables. Nous distinguerons les méthodes paramétriques impliquant des formes géométriques connues (rectangles, cercles, ...) (Figure 1.5) des méthodes dites non-paramétriques permettant de détecter des clusters de formes irrégulières.

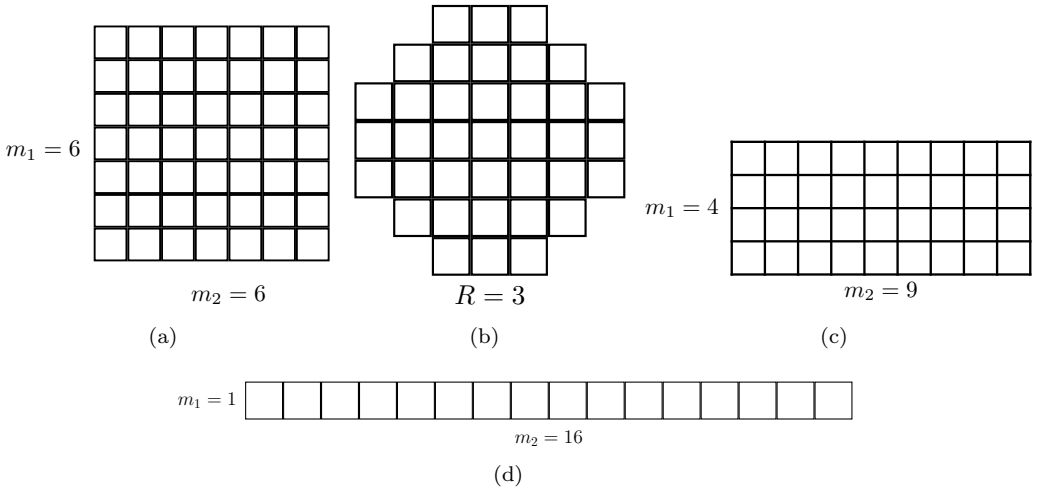


FIGURE 1.5 – Exemples de fenêtres de scan de formes discrètes différentes.

Méthodes paramétriques. Dans le cas continu, [Naus, 1965a] et [Loader, 1991] ont considéré des fenêtres de formes rectangulaires de taille fixe pour le premier et de taille variable pour le second. [Anderson and Titterington, 1997] ont utilisé des rectangles et des cercles de taille fixe. [Alm, 1997] est allé plus loin en proposant des fenêtres de formes rectangulaires, circulaires, triangulaires et elliptiques, toujours de surface fixe. Plus récemment, [Conley et al., 2005] et [Sahajpal et al., 2004] ont proposé des algorithmes génétiques de détection de cluster basés sur des fenêtres circulaires et elliptiques. Dans le cas discret, [Naus, 1965a, Chen and Glaz, 1996, Boutsikas and Koutras, 2003, Haiman and Preda, 2006] ont considéré des formes rectangulaires de taille fixe et [Chen and

Glaz, 2009] de taille variable. [Kulldorff and Nagarwalla, 1995, Kulldorff, 1997] a utilisé des fenêtres circulaires de tailles variables, puis a généralisé sa méthode à des formes elliptiques [Kulldorff, 2006].

Méthodes non-paramétriques. Dans le cas continu, [Demattei et al., 2007] ont proposé une méthode basée sur la construction de trajectoire pour la détection de cluster de forme arbitraire. Cette méthode débute par la détermination d'une trajectoire reliant les évènements. L'idée générale est que les points consécutifs dans un cluster présentent des distances associées plus faibles que les points en dehors du cluster, car la densité d'évènements dans un cluster est plus élevée. Le cluster est détecté en modélisant les changements structurels de distance au sein l'ordre de sélection des évènements, par le biais d'un modèle de régression. Dans le cas discret, [Patil and Taillie, 2004] ont proposé la statistique de scan ULS (*Upper Level Set*) qui se base sur un ensemble de clusters candidats définis par des ensembles de cellules connexes présentant une fréquence d'évènements supérieure à un seuil fixé. [Duczmal and Assuncao, 2004] ont développé la statistique de scan SA (Simulated Annealing) qui construit, dans un premier temps, l'ensemble des clusters potentiels par l'ensemble des sous-graphes connexes de cellules. Dans un deuxième temps, les sous-graphes d'intérêts sont sélectionnés par le biais d'un critère de rapport de vraisemblance, en utilisant un algorithme de recuit-simulé. [Tango and Takahashi, 2005, Tango and Takahashi, 2012] a développé la statistique de scan flexible dont le principe de fonctionnement est proche du scan SA, à la différence que l'ensemble des clusters potentiels est réduit à des zones comprenant un nombre maximum K de cellules adjacentes, permettant ainsi à la méthode de se passer d'un algorithme de recuit-simulé. [Assunção et al., 2006] ont proposé le scan MST (*Minimum Spanning Tree*), généralisation du scan ULS, utilisant le concept d'arbres couvrant de poids minimal pour la définition de l'ensemble des clusters candidats.

4 Test statistique basé sur la statistique de scan

La statistique de scan est une variable aléatoire utilisée comme statistique de test dont l'objectif est de tester la présence de cluster d'évènements au sein d'une région étudiée $\mathcal{R}$. De manière formelle, on souhaite tester l'hypothèse nulle, $\mathcal{H}_0$ d'indépendance et d'appartenance à une distribution donnée $\mathcal{F}$ des évènements observés $\{X_i\}_{i\geq 1}$:

$$\mathcal{H}_0 \ : \ X_i \sim \mathcal{F}(p_0),\ i \geq 0,$$

p_0 étant un paramètre d'intensité, contre une hypothèse alternative $\mathcal{H}_1$ privilégiant l'existence d'un cluster $\mathcal{C} \subset \mathcal{R}$ dans lequel les évènements sont i.i.d. selon une distribution identique $\mathcal{F}$ mais de paramètre d'intensité différent p_1, $p_1 > p_0$:

$$\begin{array}{rl} \mathcal{H}_1 \ : & X_i \sim \mathcal{F}(p_1),\ i \in \mathcal{C} \\ & X_i \sim \mathcal{F}(p_0),\ i \in \mathcal{R}\backslash\mathcal{C}, \end{array}$$

Il est important de distinguer le cas où la taille de la fenêtre est une constante connue (*i.e.* statistique de scan à fenêtre fixe) du cas où la taille de la fenêtre est variable (*i.e.* statistique de scan à fenêtre variable) (voir Section 4.2).

4.1 Test basé sur la statistique de scan à fenêtre fixe

Cette section rappelle, dans un premier temps, les résultats énoncés par [Naus, 1966] portant sur le cas unidimensionnel continu. Les auteurs ont montré que le test de rapport de vraisemblance testant $\mathcal{H}_0$ contre $\mathcal{H}_1$, a pour statistique de test une fonction monotone croissante de la statistique de scan à fenêtre de taille fixe.

Ce résultat est acquis dans le cas discret, cependant aucune formalisation rigoureuse n'a été, à notre connaissance, proposée dans la littérature. Nous formalisons ce résultat sous forme de propositions relatives au cas unidimensionnel discret dans le cadre des modèles binomiaux et Poisson. La démarche étant similaire dans le cas bidimensionnel, les propositions relatives à ce dernier sont disponibles en Annexe A.

4.1.1 Cas continu unidimensionnel

Posons $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires indépendantes et identiquement distribuées définissant le temps d'occurrence de N évènements dans l'intervalle $[0, T]$. Afin de simplifier les expressions, nous considérerons par la suite $T = 1$. Soit f la fonction de densité des $X_i, 1 \leq i \leq N$ définie par

$$f(x) = \begin{cases} a, & \text{si } b \leq x \leq b + m \\ \frac{1-am}{1-m} & \text{si } 0 \leq x < b \text{ ou } b + m < x \leq 1, \end{cases} \tag{1.9}$$

où $a \in \mathbb{R}$, $b \in [0, 1]$ et $m \in]0, 1[$ sont des paramètres inconnus. Nous souhaitons tester l'hypothèse $\mathcal{H}_0$ selon laquelle les évènements sont distribués de manière uniforme sur $[0, 1]$ contre l'hypothèse alternative $\mathcal{H}_1$ selon laquelle il existe un cluster d'évènements dans un sous-intervalle de $[0, 1]$ de longueur m connue. Aussi, les hypothèses de test peuvent être définies par

$$\begin{cases} \mathcal{H}_0 : a = 1 \\ \mathcal{H}_1 : 1 < a \leq 1 - \frac{1}{m}. \end{cases} \tag{1.10}$$

Théorème 1.1 ([Naus, 1966]). *Le test de rapport de vraisemblance généralisé (TRVG) rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$ lorsque la statistique de scan unidimensionnelle continue à fenêtre de taille fixe m, $S(m, T)$, excède un seuil τ déterminé à partir de $\mathbb{P}(S(m, T) > \tau | \mathcal{H}_0) = \alpha$ où α correspond au risque de première espèce associé au test.*

Preuve. *Le TRVG pour (1.10) rejette $\mathcal{H}_0$ pour les grandes valeurs de la statistique de test suivante :*

$$\lambda = \lambda(X_1, \ldots, X_N) = \frac{\sup_{\Theta_1} \prod_{i=1}^{N} f(X_i)}{\sup_{\Theta_0} \prod_{i=1}^{N} f(X_i)}, \tag{1.11}$$

où Θ_0 et Θ_1 sont les sous-ensembles de l'espace des paramètres correspondant respectivement aux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$.

$$\Theta_0 = \{(a, b, m); a = 1\}$$

$$\Theta_1 = \{(a, b, m); 0 < m < 1; 1 < a \leq 1/m; 0 \leq b \leq 1 - m\}$$

Selon (1.9) et (1.10), nous avons

$$\sup_{\Theta_0} \prod_{i=1}^{N} f(X_i) = 1$$

(1.11) se réduit donc à

$$\lambda = \sup_{\Theta_1} \prod_{i=1}^{N} f(X_i) = \sup_{\Theta_1} a^{\nu_b} \left(\frac{1 - am}{1 - m} \right)^{N - \nu_b}, \tag{1.12}$$

où ν_b est le nombre de points contenu dans l'intervalle $[b; b + m]$. Or, comme m est une constante connue, le maximum de (1.12) est atteint pour $a = \nu_b / Nm$ si $\nu_b \leq Nm$ et pour $a = 1$ autrement. Comme Θ_1 nous restreint au premier cas, (1.12) se réduit à

$$G(m, \nu_b) = \left(\frac{\nu_b}{N} \right)^{\nu_b} \left(\frac{N - \nu_b}{N} \right)^{N - \nu_b} \left(\frac{1}{m} \right)^{\nu_b} \left(\frac{1}{1 - m} \right)^{N - \nu_b} \tag{1.13}$$

Lorsque m est fixée et $\nu_b > Nm$, $G(m, \nu_b)$ est une fonction monotone croissante en ν_b, donc le TRVG rejette $\mathcal{H}_0$ pour une valeur de ν_b aussi grande que possible, à savoir la statistique de scan $S(m, T)$.

4.1.2 Cas discret unidimensionnel

Modèle binomial. Soit $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires binomiales $\mathcal{B}(n, p_k)$, indépendantes avec

$$\mathbb{P}(X_i = x_i) = \binom{n}{x_i} p_k^{x_i}(1-p_k)^{n-x_i} \qquad \forall i \in \{1, 2, \ldots, N\},$$

où $\forall\, t, m \in \mathbb{N}$, $1 \leq m < N$, $1 \leq t \leq N - m + 1$,

$$k = \begin{cases} 0 & \text{si} \quad 1 \leq i < t \quad \text{ou} \quad t+m \leq i \leq N \\ 1 & \text{si} \quad t \leq i < t+m \end{cases}$$

On suppose la longueur de la fenêtre m connue et fixée, et les paramètres p_k connus. On souhaite tester l'hypothèse nulle $\mathcal{H}_0$ selon laquelle les X_i sont *i.i.d.* $\mathcal{B}(n, p_0)$:

$$\mathcal{H}_0 : p_0 = p_1 \tag{1.14}$$

Contre une hypothèse alternative $\mathcal{H}_1$ supportant un cluster d'évènements dans une fenêtre de taille m :

$$\mathcal{H}_1 : p_1 > p_0 \tag{1.15}$$

Proposition 1.1. *Le test de rapport de vraisemblance généralisé (TRVG) rejette $\mathcal{H}_0$ (1.14) au profit de $\mathcal{H}_1$ (1.15) lorsque la statistique de scan unidimensionnelle discrète à fenêtre de taille fixe m, $S(m, N)$, excède un seuil τ déterminé à partir de $\mathbb{P}(S(m,N) > \tau | \mathcal{H}_0) = \alpha$ où α correspond au risque de première espèce associé au test.*

Preuve. *La fonction de vraisemblance sous $\mathcal{H}_0$, $L_{\mathcal{H}_0}$, a pour expression*

$$L_{\mathcal{H}_0} = \prod_{i=1}^{N} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i}.$$

L'hypothèse $\mathcal{H}_1$ peut être exprimée en fonction de t :

$$\mathcal{H}_1 = \bigcup_{t=1}^{N-m+1} \mathcal{H}_1(t).$$

Ainsi, la fonction de vraisemblance $L_{\mathcal{H}_1}(t)$, a pour expression

$$L_{\mathcal{H}_1}(t) = \left(\prod_{i=1}^{t-1} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i} \right) \left(\prod_{i=t}^{t+m-1} \binom{n}{x_i} p_1^{x_i}(1-p_1)^{n-x_i} \right) \times \left(\prod_{i=t+m}^{N} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i} \right).$$

Le rapport de vraisemblance $LR(t, m)$ a donc pour expression

$$LR(t,m) = \frac{\left(\prod_{i=1}^{t-1} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i} \right) \left(\prod_{i=t}^{t+m-1} \binom{n}{x_i} p_1^{x_i}(1-p_1)^{n-x_i} \right)}{\prod_{i=1}^{N} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i}} \times \frac{\left(\prod_{i=t+m}^{N} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i} \right)}{\prod_{i=1}^{N} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i}},$$

qui se simplifie en

$$LR(t,m) = \frac{\prod_{i=t}^{t+m-1} \binom{n}{x_i} p_1^{x_i}(1-p_1)^{n-x_i}}{\prod_{i=t}^{t+m-1} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i}}.$$

Le logarithme du rapport de vraisemblance, $LLR(t,m)$ a donc pour expression

$$LLR(t,m) = \log \prod_{i=t}^{t+m-1} \binom{n}{x_i} p_1^{x_i}(1-p_1)^{n-x_i} - \log \prod_{i=t}^{t+m-1} \binom{n}{x_i} p_0^{x_i}(1-p_0)^{n-x_i},$$

se simplifiant en

$$LLR(t,m) = \sum_{i=t}^{t+m-1} \left[x_i \log\left(\frac{p_1}{p_0}\right) + (n-x_i)\log\left(\frac{1-p_1}{1-p_0}\right)\right].$$

Posons $C_1 = \log\left(\frac{p_1}{p_0}\right)$ et $C_2 = \log\left(\frac{1-p_1}{1-p_0}\right)$. Comme $p_1 > p_0$ alors $C_1 > 0$ et $C_2 < 0$.

$$LLR(t,m) = C_1 \sum_{i=t}^{t+m-1} x_i + C_2 \sum_{i=t}^{t+m-1} (n-x_i)$$

Posons

$$\nu_t = \sum_{i=t}^{t+m-1} x_i,$$

ν_t correspond au nombre d'évènements observés dans la fenêtre $[t, t+m-1]$. Le $LLR(t,m)$ s'écrit donc

$$LLR(t,m) = C_1\nu_t + C_2(mn - \nu_t).$$

Pour m fixé, et comme $C_1 > 0$ et $C_2 < 0$, le $LLR(t,m)$ est une fonction monotone croissante en ν_t. Par conséquent, le test de rapport de vraisemblance rejette $\mathcal{H}_0$ pour une valeur de ν_t suffisamment grande, à la savoir la statistique de scan unidimensionnelle discrète à fenêtre fixe de longueur m :

$$S(m,N) = \max_{1\le t\le N-m+1} \nu_t.$$

Modèle de Poisson. Soit $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires de Poisson $\mathcal{P}(\lambda_k)$, indépendantes avec

$$\mathbb{P}(X_i = x_i) = \frac{e^{-\lambda_k}\lambda_k^{x_i}}{x_i!} \qquad \forall i \in \{1,2,\ldots,N\},$$

où

$$\forall\, t,m \in \mathbb{N},\ 1 \le m < N,\ 1 \le t \le N-m+1$$

$$k = \begin{cases} 0 & \text{si} \quad 1 \le i < t \quad \text{ou} \quad t+m \le i \le N \\ 1 & \text{si} \quad t \le i < t+m. \end{cases}$$

On suppose la longueur de la fenêtre m connue et fixée, et les paramètres λ_k connus. On souhaite tester l'hypothèse nulle $\mathcal{H}_0$ selon laquelle les X_i sont *i.i.d.* $\mathcal{P}(\lambda_0)$:

$$\mathcal{H}_0 : \lambda_0 = \lambda_1 \tag{1.16}$$

Contre une hypothèse alternative $\mathcal{H}_1$ supportant un cluster d'évènements dans une fenêtre de taille m :

$$\mathcal{H}_1 : \lambda_1 > \lambda_0 \tag{1.17}$$

Proposition 1.2. *Le test de rapport de vraisemblance généralisé (TRVG) rejette $\mathcal{H}_0$ (1.16) au profit de $\mathcal{H}_1$ (1.17) lorsque la statistique de scan unidimensionnelle discrète à fenêtre de taille fixe m, $S(m,N)$, excède un seuil τ déterminé à partir de $\mathbb{P}(S(m,N) > \tau|\mathcal{H}_0) = \alpha$ où α correspond au risque de première espèce associé au test.*

Preuve. *La fonction de vraisemblance sous $\mathcal{H}_0$, $L_{\mathcal{H}_0}$, a pour expression*

$$L_{\mathcal{H}_0} = \prod_{i=1}^{N} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}.$$

Ainsi, la fonction de vraisemblance $L_{\mathcal{H}_1}(t)$, a pour expression

$$L_{\mathcal{H}_1}(t) = \left(\prod_{i=1}^{t-1} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}\right)\left(\prod_{i=t}^{t+m-1} \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}\right)\left(\prod_{i=t+m}^{N} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}\right).$$

Le rapport de vraisemblance $LR(t,m)$ a donc pour expression

$$LR(t,m) = \frac{\prod_{i=t}^{t+m-1} \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}}{\prod_{i=t}^{t+m-1} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}},$$

et son logarithme, $LLR(t,m)$

$$LLR(t,m) = \sum_{i=t}^{t+m-1} \left[(-\lambda_1 + x_i\log(\lambda_1) - \log(x_i!)) - (\lambda_0 + x_i\log(\lambda_0) - \log(x_i!))\right],$$

se simplifie en

$$LLR(t,m) = m(\lambda_0 - \lambda_1) + \sum_{i=t}^{t+m-1} x_i \log\left(\frac{\lambda_1}{\lambda_0}\right).$$

Posons $C = \log\left(\frac{\lambda_1}{\lambda_0}\right)$. Comme $\lambda_1 > \lambda_0$ alors $C > 0$ et $m(\lambda_0 - \lambda_1) < 0$.

$$LLR(t,m) = m(\lambda_0 - \lambda_1) + C\sum_{i=t}^{t+m-1} x_i.$$

Posons

$$\nu_t = \sum_{i=t}^{t+m-1} x_i$$

$$LLR(t,m) = m(\lambda_0 - \lambda_1) + C\nu_t.$$

Pour m fixé, et comme $C > 0$ et $m(\lambda_0 - \lambda_1) < 0$, le $LLR(t,m)$ est une fonction monotone croissante en ν_t. Par conséquent, le test de rapport de vraisemblance rejette $\mathcal{H}_0$ pour une valeur de ν_t suffisamment grande, à la savoir la statistique de scan unidimensionnelle discrète à fenêtre fixe de longueur m :

$$S(m,N) = \max_{1\leq t\leq N-m+1} \nu_t.$$

4.2 Test de la statistique de scan à fenêtre variable

Nous introduisons ici la notion de statistique de scan à fenêtre variable au sein d'une région étudiée $\mathcal{R}$. Nous désirons tester l'hypothèse nulle $\mathcal{H}_0$ stipulant que les évènements observés $\{X_i\}_{i\geq 1}$ sont i.i.d. selon une distribution $\mathcal{F}(p_0)$ contre l'hypothèse alternative supportant la présence d'un cluster $\mathcal{C} \subset \mathcal{R}$ dans lequel les évènements sont i.i.d. selon une distribution $\mathcal{F}(p_1)$, $p_1 > p_0$. A l'extérieur de $\mathcal{C}$ les évènements sont i.i.d. selon $\mathcal{F}(p_0)$. Lorsque la taille de la fenêtre de scan (*i.e.* $\mathcal{C}$) est variable, $\mathcal{C}_{min} \leq \mathcal{C} \leq \mathcal{C}_{max}$, on parle de **statistique de scan à fenêtre variable**.

Cette section présente le test basé sur la statistique de scan à fenêtre variable. Elle reprend essentiellement le résultat issu de [Nagarwalla, 1996]. Dans cet article, les auteurs ont montré que lorsque la taille de la fenêtre de scan est variable, le rapport de vraisemblance explicité *supra* n'est plus une fonction monotone croissante de la statistique de scan. Cette section décrit ce résultat uniquement dans le cas unidimensionnel continu. Cependant, les démonstrations dans les cas bidimensionnel continu et uni et bidimensionnel discret sont relativement similaires.

Posons $X_1, X_2, \dots X_N$ une suite de variables aléatoires indépendantes et identiquement distribuées dans l'intervalle $[0,1]$. Soit f la fonction de densité des $X_i, 1 \leq i \leq N$ définie par

$$f(x) = \begin{cases} a, & \text{si } b \leq x \leq b+m \\ \frac{1-am}{1-m} & \text{si } 0 \leq x < b \text{ ou } b+m < x \leq 1, \end{cases} \tag{1.18}$$

où a et b sont des paramètres inconnus.

Nous souhaitons tester l'hypothèse $\mathcal{H}_0$ selon laquelle les évènements sont distribués de manière uniforme sur $[0,1]$ contre l'hypothèse alternative $\mathcal{H}_1$ selon laquelle il existe un cluster d'évènements dans un sous-intervalle de $[0,1]$ de longueur m connue. Aussi, les hypothèses de test peuvent être définies par

$$\begin{cases} \mathcal{H}_0 : a = 1 \\ \mathcal{H}_1 : 1 < a \leq 1 - \frac{1}{m}. \end{cases} \tag{1.19}$$

Si m est une constante connue alors le théorème 1.1 stipule que le TRVG rejette $\mathcal{H}_0$ pour de grandes valeurs de la statistique de scan unidimensionnelle continue à fenêtre de taille m, $S(m,T)$. Plus précisément, la fonction $G(m,\nu_b)$ définie en (1.13) est une fonction monotone croissante de ν_b, le nombre d'évènements observés dans la fenêtre de taille $[b, b+m]$. Cependant, lorsque m est variable, il n'est plus possible de trouver une variable aléatoire de distribution connue telle que le rapport de vraisemblance λ défini en (1.12) soit une fonction monotone de cette variable.

Théorème 1.2 ([Nagarwalla, 1996])**.** *Le test de rapport de vraisemblance généralisé (TRGV) restreint pour le système d'hypothèses (1.19), a, b et m étant inconnus, rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$ lorsque la statistique de test*

$$\Lambda = \sup_{0<m<\nu_b/N,\, \nu_b \geq \nu_0} \left(\frac{\nu_b}{N}\right)^{\nu_b} \left(\frac{N-\nu_b}{N}\right)^{N-\nu_b} \left(\frac{1}{m}\right)^{\nu_b} \left(\frac{1}{1-m}\right)^{N-\nu_b} \tag{1.20}$$

excède un seuil τ déterminé à partir de $\mathbb{P}(\Lambda > \tau | \mathcal{H}_0) = \alpha$, α étant le risque de première espèce associé au test.

Preuve. *Pour une valeur fixée de $\nu_b = 1, \dots, N-1$, $G(m,\nu_m)$ est une fonction convexe de m car la dérivée de son logarithme est égale à*

$$\frac{Nm - \nu_b}{m(1-m)} = \begin{cases} < 0, & m < \nu_b/N \\ > 0, & m > \nu_b/N \end{cases}$$

Aussi, son minimum est atteint à ν_b/N et $\lim_{m\to 0} G(m,\nu_b) = \infty$. Intuitivement, cela signifie que chaque évènement ($\nu_b = 1$) peut être un cluster car il est compris dans un intervalle m de longueur infinitésimale. Aussi, en pratique, nous devons restreindre le domaine sur lequel on recherche le maximum de $G(m,\nu_b)$. Nous pouvons soit imposer une taille minimum de fenêtre $m \geq m_0 > 0$ ou uniquement considérer les clusters potentiels comprenant $\nu_b \geq \nu_0$ évènements, où $2 \leq \nu_0 \leq N-2$. Cette dernière affirmation, choisie dans [Nagarwalla, 1996], implique que $m \geq m_{min} = \min_{1\leq i\leq N-\nu_0+1}(x_{(i+\nu_0-1)} - x_{(i)}) > 0$, où $x_{(1)} < x_{(2)} < \ldots < x_{(N)}$ sont les observations ordonnées par ordre croissant sur $[0,1]$. Aussi, pour de grandes valeurs de la statistique

$$\Lambda = \sup_{0<m<\nu_b/N,\, \nu_b \geq \nu_0} \left(\frac{\nu_b}{N}\right)^{\nu_b} \left(\frac{N-\nu_b}{N}\right)^{N-\nu_b} \left(\frac{1}{m}\right)^{\nu_b} \left(\frac{1}{1-m}\right)^{N-\nu_b}$$

le TRVG restreint rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$. La classe d'alternatives pour laquelle le test est dérivé correspond à des sous-intervalles de $[0,1]$ de taille inconnue et variable, dans lesquels il y a un risque uniforme élevé. Nous utilisons l'appellation "restreint" pour le TRVG car ce dernier n'évalue que les espaces d'ordre $\nu_b \geq \nu_0$, $2 \leq \nu_0 \leq N-2$ où ν_0 est fixé par l'utilisateur.

Le calcul de Λ est donné par l'algorithme 1. Aussi, le cluster le plus probable débute à i^*ème observation et est constitué de ν_b^* évènements. λ^* correspond à la valeur observée du rapport de vraisemblance.

Algorithme 1 - Calcul de Λ

Pour tout ν_b, $\nu_0 \leq \nu_b \leq N-2$ **Faire**
  Pour tout i, $1 \leq i \leq N - \nu_b + 1$ **Faire**
    Calculer $x_{(i+\nu_b-1)} - x_{(i)}$
    Déterminer $m_{min}(\nu_b) = \min_{1 \leq i \leq N-\nu_b+1}(x_{(i+\nu_b-1)} - x_{(i)})$
    Déterminer $i_{min}(\nu_b)$ la position telle que $m_{min}(\nu_b) = x_{(i+\nu_b-1)} - x_{(i)}$
    Calculer $\lambda(n) = \left(\frac{\nu_b}{N}\right)^{\nu_b} \left(\frac{N-\nu_b}{N}\right)^{N-\nu_b} \left(\frac{1}{m_{min}}\right)^{\nu_b} \left(\frac{1}{1-m_{min}}\right)^{N-\nu_b}$
  fin Pour
fin Pour
Déterminer $\lambda^* = \max_{\nu_0 \leq \nu_b \leq N-2} \lambda(\nu_b)$
Conserver ν_b^* et i^*, les valeurs de ν_b et i pour lesquelles le maximum est atteint.

5 Conclusion

Les sections 2 et 3 ont défini les statistiques de scan dans les cas uni et bidimensionnel, discret et continu. La section 4 explicite le rôle de la statistique au sein d'un TRVG permettant de tester l'existence de cluster d'évènements dans une fenêtre. Dans le cas où la taille de la fenêtre de scan est fixe, la statistique de test du TRVG est une fonction monotone croissante de la statistique de scan. Nous l'avons formalisé sous forme de propositions dans les cas uni et bidimensionnel discrets. Ainsi, dans ce cas de figure, déterminer la distribution, sous $\mathcal{H}_0$ de la statistique de test du TRVG revient à déterminer la distribution de la statistique de scan à fenêtre de taille fixe, ce qui fera l'objet d'une étude approfondie au sein du chapitre 2.

Dans le cas où la taille de la fenêtre de scan est variable, la statistique de test du TRVG n'est plus une fonction monotone croissante de la statistique de scan. Dans [Nagarwalla, 1996], les auteurs ont proposé un TRVG restreint dont la statistique de test, Λ, ne dispose pas de distribution de forme analytique. En conséquence, les auteurs proposent d'utiliser des simulations de Monte Carlo afin d'estimer sa distribution sous $\mathcal{H}_0$. L'extension de cette méthode au cas spatial sera explicitée dans le cadre des statistiques de scan spatiales, au sein du chapitre 3.

Chapitre 2

Approximation de la distribution des statistiques de scan

Sommaire

1 Introduction 31
2 Cas non-conditionnel 32
2.1 Statistique de scan unidimensionnelle 32
2.2 Statistique de scan bidimensionnelle 36
3 Cas conditionnel 44
3.1 Statistique de scan unidimensionnelle 45
3.2 Statistique de scan bidimensionnelle 49
4 Influence de la forme de la fenêtre 54
4.1 Introduction 54
4.2 Forme de la fenêtre de scan 54
4.3 Applications numériques 57
5 Conclusion 61

1 Introduction

Dans le chapitre 1, nous avons montré que les statistiques de scan sont utilisées comme statistique de test pour vérifier l'hypothèse $\mathcal{H}_0$ l'indépendance et l'appartenance à une distribution donnée des observations contre une hypothèse alternative $\mathcal{H}_1$ supportant l'existence de cluster. Aussi, la connaissance de la distribution de probabilités de la statistique de scan, sous $\mathcal{H}_0$, est nécessaire afin de pouvoir rejeter, ou non, cette dernière hypothèse.

Par définition, la statistique de scan est le maximum d'une suite de variables aléatoires dépendantes. En effet, les ν_t définis en (1.1) sont dépendants (recouvrement). Aussi, la détermination de sa loi de probabilité d'un point de vue analytique s'avère être une tâche ardue voire impossible dans de nombreux cas. Néanmoins, plusieurs auteurs ont développé des formules exactes, dans des cas très particuliers, et plus largement des méthodes d'approximations probabilistes. Ce chapitre a pour objectif de présenter les différentes méthodes existantes permettant de déterminer la distribution des statistiques de scan.

Une attention toute particulière doit être portée sur la différenciation entre les statistiques de scan conditionnelles et non-conditionnelles. En effet, les formules exactes ainsi que les approximations reposent sur des concepts mathématiques qui diffèrent en fonction des deux cas de figure.

Par ailleurs, les différentes techniques d'approximation présentées dans le présent chapitre sont fonction du type de modèle probabiliste régissant la distribution des observations. Nous distinguerons les principaux modèles : Bernoulli, binomial et Poisson.

Dans un premier temps, ce chapitre va s'attacher à décrire les différentes méthodes d'approximation de la distribution de la statistique de scan sous $\mathcal{H}_0$ dans les cas non conditionnel et conditionnel, uni et bidimensionnel, discret et continu. Nous proposons une série de démonstrations concernant des théorèmes énoncés dans la littérature. Dans un second temps, nous avons étudié, par le biais d'une étude de simulations, l'influence de la forme de la fenêtre de scan sur l'approximation de la distribution des statistiques de scan bidimensionnelles discrètes ainsi que sur la puissance de détection.

2 Cas non-conditionnel

2.1 Statistique de scan unidimensionnelle

2.1.1 Statistique de scan unidimensionnelle discrète

Techniques d'approximation. Soit $\{X_1, X_2, \ldots, X_N\}$ une suite de variables aléatoires indépendantes et identiquement distribuées à valeurs dans $\mathbb{N}$. En considérant $S(m, N)$ définie en (1.2), nous cherchons à déterminer

$$\mathbb{P}(S(m,N) \leq k),\ \forall k \geq 0.$$

Posons

$$Q_L = \mathbb{P}(S(m, Lm) \leq k), \tag{2.1}$$

où $L \in \mathbb{N}$ et $L = N/m$.

[Naus, 1982] a fourni une approximation de Q_L basée sur un comportement markovien :

Théorème 2.1. *Pour $N = Lm$, $L \geq 3$ et $k \geq 0$, Q_L peut être approximée par*

$$Q_L \approx Q_2 \left(\frac{Q_3}{Q_2}\right)^{L-2}. \tag{2.2}$$

Preuve. *Posons E_j, $1 \leq j \leq L-1$ tel que*

$$E_j = \left\{ \max_{(j-1)m+1 \leqslant t \leqslant jm+1} \nu_t \leq k \right\}, \tag{2.3}$$

avec ν_t défini en (1.1). Remarquons que E_i et E_j, $1 \leq i, j \leq L-1$ sont indépendants si $|i-j| > 1$ (Figure 2.1).

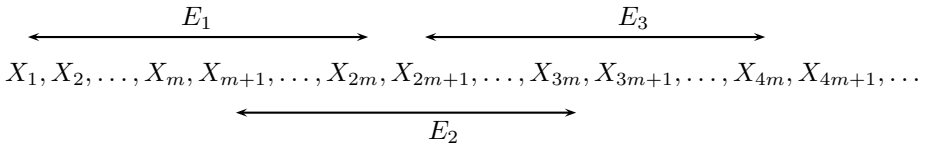


FIGURE 2.1 – Exemple de E_j

Q_L peut être alors exprimée de la façon suivante :

$$Q_L = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\ldots\mathbb{P}\left(E_{L-1}|\bigcap_{i=1}^{L-2} E_i\right).$$

En utilisant un comportement de type markovien pour la suite $\{E_j\}$:

$$\mathbb{P}\left(E_l|\bigcap_{i=1}^{l-1} E_i\right) \underbrace{\approx \mathbb{P}(E_l|E_{l-1})}_{\text{Absence de mémoire}} \underbrace{\approx \mathbb{P}(E_2|E_1)}_{\text{Echangeabilité}} = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} = \frac{Q_3}{Q_2}$$

Q_L peut donc être approximée par

$$Q_L \approx Q_2 \left(\frac{Q_3}{Q_2}\right)^{L-2}.$$

Remarque 2.1. *Ce type d'approximation est souvent dénommée "Product type" dans la littérature.*

L'approximation (2.2) a été évaluée au travers de nombreuses études de simulations et s'avère être très précise [Glaz et al., 2001]. Néanmoins, nous ne disposons pas d'une erreur associée à l'approximation.

[Haiman, 2007] a proposé une approximation de Q_L basée les propriétés du maximum partiel d'une suite stationnaire de variables aléatoires 1-dépendantes, la notion d'1-dépendance étant définie par

Définition 2.1. Soit $\{Z_n\}_{n\geq 1}$ une suite stationnaire de variables aléatoires. On dit que les Z_n sont 1-dépendantes si $\forall k \geq 1$, $\sigma(\ldots, Z_k)$ et $\sigma(Z_{k+2}, \ldots)$ sont indépendantes. En d'autres termes, $\forall\, i \neq j$, Z_i est indépendante à Z_j si $|i-j| \geq 2$.

Cette approximation se base sur un résultat énoncé par [Haiman, 1999] portant sur le maximum partiel d'une suite stationnaire de variable aléatoires 1-dépendantes. Soit $\{Z_n\}_{n\geq 1}$ une suite stationnaire de variables aléatoires 1-dépendantes. Pour $x < \omega$, $\omega = \sup\{u|\mathbb{P}(Z_1 \leq u) < 1\}$, posons

$$q_n = q_n(x) = \mathbb{P}(\max\{Z_1, Z_2, \ldots, Z_n\} \leq x).$$

Théorème 2.2 ([Haiman, 1999])**.** *Pour tout x tel que $\mathbb{P}(Z_1 > x) = 1 - q_1 \leq 0.025$ et pour tout entier $n > 3$ tel que $3.3n(1-q_1)^2 \leq 1$, nous avons*

$$\frac{\left|q_n - \frac{2q_1 - q_2}{(1-q_1+q_2+2(q_1-q_2)^2)^n}\right|}{q_n} \leqslant 3.3n(1-q_1)^2. \tag{2.4}$$

Ce résultat exprime le fait que q_n peut être approximée par . Pour une démonstration de ce théorème, voir [Haiman, 1999].

Faisons maintenant le lien entre ce théorème et la statistique de scan. Pour $L = N/m$ et ν_t défini en (1.1), posons

$$Z_j = \left\{ \max_{(j-1)m+1 \leqslant t \leqslant jm+1} \nu_t \right\},\ 1 \leq j \leq L-1. \tag{2.5}$$

Il est aisé de voir de que les Z_j forment une suite stationnaire de variables aléatoires 1-dépendantes (Figure 2.2).

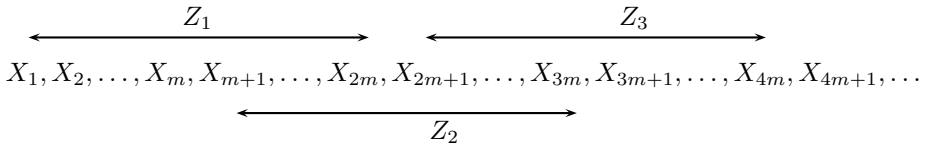


FIGURE 2.2 – Exemple de Z_j

Aussi, $S(m, Lm)$ peut être vue comme le maximum d'une suite stationnaire de variables aléatoires 1-dépendantes :

$$S(m, Lm) = \max_{1 \leq j \leq L-1} \{Z_j\}.$$

Pour $i \in \{2, 3\}$, posons

$$Q_i\,(= Q_i(k)) = \mathbb{P}\left(\bigcap_{j=1}^{i-1} \{Z_j \leq k\}\right) = \mathbb{P}\left(\max_{1 \leq j \leq i-1} \{Z_j\}\right) = \mathbb{P}(S(m, im) \leq k).$$

En prenant les notations de l'équation (2.4), remarquons que $Q_i = q_{i-1}$. Aussi, $1 - Q_2 \leq 0.025$, le théorème (2.2) peut être appliqué à la suite $\{Z_1, Z_2, \ldots, Z_{L-1}\}$ afin d'obtenir l'approximation suivante de la distribution de $S(m, Lm)$:

$$\mathbb{P}(S(m, Lm) \leq k) \approx \frac{2Q_2 - Q_3}{(1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2)^{L-1}}, \tag{2.6}$$

avec une erreur d'approximation de l'ordre de $3.3L(1 - Q_2)^2$.

Remarque 2.2 (***Comparaison des deux techniques***). *L'approximation développée par [Naus, 1982] présente la particularité de ne pas nécessiter de conditions d'applications. Aussi, elle permet d'approximer l'ensemble de la distribution de $S(m, N)$. Cependant, elle ne propose pas d'erreur d'approximation.* A contrario, *l'approximation proposée par [Haiman, 2007] a des conditions d'applications. Aussi, elle ne permet pas d'approximer l'ensemble de la distribution de $S(m, N)$ mais seulement les queues de distribution, ce qui est le plus pertinent dans le cas du test statistique associé. Cependant, elle présente l'avantage d'associer une erreur à l'approximation de la distribution. En outre, les deux approximations, lorsqu'elles sont applicables, donnent des résultats similaires puisqu'on peut montrer que*

$$\frac{2Q_2 - Q_3}{(1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2)^{L-1}} = Q_2 \left(\frac{Q_3}{Q_2}\right)^{L-2} (1 + \mathcal{O}(L(1 - Q_2)^2)).$$

Remarque 2.3. *Les approximations (2.2) et (2.6) présentent comme caractéristique principale de dépendre uniquement de la connaissance préalable de la distribution de $S(m, 2m)$ et $S(m, 3m)$, à savoir Q_2 et Q_3. De surcroît, elles sont applicables quelle que soit la distribution des X_i.*

Remarque 2.4. *Considérons les approximations (2.2) et (2.6). Lorsque N n'est pas un multiple de m, considérons $L = \lfloor \frac{N}{m} \rfloor$ et en utilisant ces inégalités*

$$\mathbb{P}(S(m, (L+1)m) \leq k) \leq \mathbb{P}(S(m, N) \leq k) \leq \mathbb{P}(S(m, Lm) \leq k),$$

on peut raisonnablement approximer $\mathbb{P}(S(m, N) \leq k)$ par interpolation linéaire.

Comme nous l'avons dit précédemment, les approximations (2.2) et (2.6) font appel aux quantités Q_2 et Q_3. Pour certaines distributions des X_i, ces quantités sont calculables par le biais de formules exactes. C'est le cas du modèle de Bernoulli. Pour les autres distributions des X_i, le calcul des quantités Q_2 et Q_3 est réalisé par des approximations utilisant des simulations de Monte Carlo.

Modèle de Bernoulli. Soit $\{X_1, X_2, \ldots, X_N\}$ une suite de variables aléatoires indépendantes de Bernoulli $\mathcal{B}(1, p)$.

[Naus, 1982] a fourni des formules exactes dans les cas très restreints où $N = 2m$ et $N = 3m$

Théorème 2.3. *Soient $b(k; m; p) = \binom{m}{k} p^k (1-p)^{m-k}$ la loi de probabilité de la loi binomiale $\mathcal{B}(m, p)$ et $F_b(r; s; p) = \sum_{i=0}^{r} b(i; s; p)$ $r = 0, 1, 2, ..., s$ sa fonction de répartition. Pour $2 < k < N$ et $0 < p < 1$ nous avons*

$$\begin{aligned} Q_2 = &(F_b(k-1; m; p))^2 - (k-1)b(k; m, p)F_b(k-2; m; p) \\ &+ mpb(k; m; p)Fb(k-3; m-1; p), \end{aligned} \tag{2.7}$$

$$Q_3 = (F_b(k-1; m; p))^3 - A_1 + A_2 + A_3 - A_4, \tag{2.8}$$

où A_1, A_2, A_3 et A_4 sont respectivement égaux à

$$A_1 = 2b(k; m; p)F_b(k-1; m; p)\{(k-1)F_b(k-2; m; p) - mpFb(k-3; m-1; p)\}$$

$$\begin{aligned} A_2 = 0.5((b(k; m; p))^2\{&(k-1)(k-2)F_b(k-3; m; p) - 2(k-2)mpF_b(k-4; m-1; p) \\ &+ m(m-1)p^2F_b(k-5; m-2; p)\} \end{aligned}$$

$$A_3 = \sum_{r=1}^{k-1} b(2k-r; m; p) \left(F_b(r-1; m; p)\right)^2$$

$$A_4 = \sum_{r=2}^{k-1} b(2k-r; m; p)b(r; m; p)\{(r-1)F_b(r-2; m; p) - mpF_b(r-3; m-1; p)\}$$

Exemples numériques. La Table 2.1 compare les approximations (2.2) et (2.6) dans le cas discret lorsque les évènements sont distribués selon une loi de Bernoulli $\mathcal{B}(1,p)$.

k	8	9	10	11
$\mathbb{P}(S(30,256\times 30)\leq k)$:				
App. (2.6)	0.0794	0.5161	0.85979	0.970613
Erreur	0.12	0.008	0.0023	10^{-6}
App. (2.2)	0.0797	0.5172	0.86028	0.970726
$\mathbb{P}(S(30,512\times 30)\leq k)$:				
App. (2.6)		0.2658	0.73888	0.941997
Erreur		0.017	0.00046	0.000017
App. (2.2)		0.2663	0.739295	0.9421067
$\mathbb{P}(S(30,1024\times 30)\leq k)$:				
App. (2.6)		0.07052	0.5456789	0.8872712
Erreur		0.033	0.0009	0.000034
App. (2.2)		0.07060	0.54596	0.887373

TABLE 2.1 – Approximations pour $\mathbb{P}(S(m,N)\leq k)$ par App. (2.6) de Haiman (2007) et App. (2.2) de Naus (1982), $X_i\sim\mathcal{B}(1,p)$, $p=0.1$, $m=30$.

2.1.2 Statistique de scan unidimensionnelle continue

Techniques d'approximation. Soit $N=\{N_t\}_{t\geq 0}$ un processus ponctuel défini sur l'intervalle $[0,T]$, $T\in\mathbb{R}^+$ fixé. En considérant $S(m,T)$ définie en (1.4), nous cherchons à déterminer

$$\mathbb{P}(S(m,T)\leq k),\ \forall k\geq 0.$$

Posons

$$Q_L=\mathbb{P}(S(m,Lm)\leq k),$$

où $L\in\mathbb{N}$ et $L=T/m$.

Les approximations (2.2) et (2.4) définies dans le cas discret sont également applicables pour Q_L dans le cas continu [Naus, 1982, Haiman, 2000]. En effet, par analogie au cas discret, pour $1\leq j\leq L-1$ et ν_t défini en (1.3), posons

$$E_j=\left\{\max_{t\in[(j-1)m,jm]}\nu_t\leq k\right\}=\{S(m,[(j-1)m,(j+1)m])\leq k\},$$
$$Z_j=\max_{t\in[(j-1)m,jm]}\nu_t=S(m,[(j-1)m,(j+1)m]).$$

Processus de Poisson. Soit $N=\{N_t\}_{t\geq 0}$ un processus de Poisson d'intensité λ et défini sur l'intervalle $[0,T]$, $T\in\mathbb{R}^+$ fixé. [Huntington and Naus, 1975] ont proposé une formule exacte pour $\mathbb{P}(S(m,T)\leq k)$ qui conduit à des temps de calculs extrêmement importants lorsque T est grand par rapport à m. Lorsque $T=2m$ et $T=3m$, [Neff and Naus, 1980] ont fourni des formules exactes pour Q_L :

Théorème 2.4. *Soient $\Psi=\lambda m$, $\forall k\geq 0$, $p(j;\Psi)=\frac{e^{-\Psi}\Psi^j}{j!}$ la loi de probabilité de la loi de Poisson $\mathcal{P}(\Psi)$ et $F_p(k-1;\Psi)=\sum_{j=0}^{k-1}p(j;\Psi)$ sa fonction de répartition. Pour $k>2$ et $\Psi>0$ nous avons*

$$Q_2=(F_p(k-1;\Psi))^2-(k-1)p(k;\Psi)p(k-2;\Psi)-(k-1-\Psi)p(k;\Psi)F_p(k-3;\Psi),\tag{2.9}$$

$$Q_3=(F_p(k-1;\Psi))^3-A_1+A_2+A_3+A_4,\tag{2.10}$$

avec

$$A_1 = 2p(k;\Psi)F_p(k-1;\Psi)\{(k-1)F_p(k-2;\Psi) - \Psi F_p(k-3;\Psi)\},$$
$$A_2 = .5(p(k;\Psi))^2\{(k-1)(k-2)F_p(k-3;\Psi) - 2(k-2)\Psi F_p(k-4;\Psi) + \Psi^2 F_p(k-5;\Psi)\},$$
$$A_3 = \sum_{r=1}^{k-1} p(2k-r;\Psi)(F_p(r-1;\Psi))^2,$$
$$A_4 = \sum_{r=2}^{k-1} p(2k-r;\Psi)p(r;\Psi)\big\{(r-1)F_p(r-2;\Psi) - \Psi F_p(r-3;\Psi)\big\}.$$

Par ailleurs, [Alm, 1983] a proposé

$$\mathbb{P}(S(m,T) \leq k) \approx F_p(k-1;\Psi)e^{\{((k-\Psi)/k)\lambda(T-m)p(k-1;\Psi)\}}. \tag{2.11}$$

Exemples numériques. La table 2.2 présente quelques valeurs numériques résultant de l'application de l'approximation (2.6) utilisant les formules exactes de Neff et Naus définies en (2.9) et (2.10) pour le calcul de Q_2 et Q_3

k	λ	App. (2.6)	Erreur
4	0.1	0.9854	2×10^{-6}
6	0.5	0.9302	2.5×10^{-5}
9	1.3	0.9405	1.7×10^{-5}

TABLE 2.2 – Approximations pour $\mathbb{P}(S \leq k)$ par app. (2.6). $T = 1001$.

2.2 Statistique de scan bidimensionnelle

2.2.1 Statistique de scan bidimensionnelle discrète

Soit une région rectangulaire $[0, N_1] \times [0, N_2]$ avec N_1, $N_2 \in \mathbb{N}$. Considérons $\{X_{i,j}\}$, $1 \leq i \leq N_1$ et $1 \leq j \leq N_2$ un ensemble de variables aléatoires indépendantes, identiquement distribuées et à valeurs dans $\mathbb{N}$. En considérant $S(m_1, m_2, N_1, N_2)$ définie en (1.6), nous cherchons à déterminer

$$\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k),\ k \geq 0.$$

Une approximation de type "Product type" a été proposée par [Glaz and Naus, 1991], puis améliorée par [Chen and Glaz, 1996].

Théorème 2.5 ([Glaz and Naus, 1991])**.** *Soit $\{X_{ij}\}$ une suite de variables aléatoires i.i.d. selon une loi binomiale de paramètres n et $0 < p_0 < 1$ ou une loi de Poisson de paramètre λ_0. La distribution de $S(m_1, m_2, N_1, N_2)$ est approximée par*

$$\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k) \approx Q_{m_1,2m_2-1}\left(\frac{Q_{m_1,2m_2}}{Q_{m_1,2m_2-1}}\right)^{(N_1-2m_1+1)(N_2-m_2+1)}, \forall k \geq 0, \tag{2.12}$$

avec

$$Q_{m_1,2m_2-1} = \mathbb{P}(S(m_1, m_2, m_1, 2m_2 - 1) \leq k),$$

$$Q_{m_1,2m_2} = \mathbb{P}(S(m_1, m_2, m_1, 2m_2) \leq k).$$

Preuve. *Pour $1 \leq i_1 \leq N_1 - m_1 + 1$ et $1 \leq i_2 \leq N_2 - m_2 + 1$ définissons les évènements*

$$A_{i_1,i_2} = \left\{ \sum_{i=i_1}^{i_1+m_1-1} \sum_{i=i_2}^{i_2+m_2-1} X_{ij} \leq k \right\}.$$

Aussi

$$\mathbb{P} = \mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k) = \mathbb{P}\left(\bigcap_{i_1=1}^{N_1-m_1+1} \bigcap_{i_2=1}^{N_2-m_2+1} A_{i_1,i_2}\right).$$

Afin de simplifier la présentation des résultats, prenons le cas où $N_1 = N_2 = N$ *et* $m_1 = m_2 = m$. *Pour une valeur fixée de* i_1, $1 \leq i_1 \leq N_1 - m_1 + 1$, *[Glaz and Naus, 1991] ont montré que l'on peut approximer de manière précise*

$$\mathbb{P}\left(\bigcap_{i_2=1}^{N-m+1} A_{i_1,i_2}\right) = \mathbb{P}\left(\prod_{i_2=1}^{m+1} A_{i_1,i_2}\right) \prod_{i_2=m+2}^{N-m+1} \left[\frac{\mathbb{P}\left(\bigcap_{j=1}^{i_2} A_{i_1,j}\right)}{\mathbb{P}\left(\bigcap_{j=1}^{i_2-1} A_{i_1,j}\right)}\right].$$

En utilisant une hypothèse de comportement markovien, notamment d'absence de mémoire sur la dépendance entre les A_{i_1,i_2}

$$\mathbb{P}\left(\bigcap_{i_2=1}^{N-m+1} A_{i_1,i_2}\right) \approx \mathbb{P}\left(\prod_{i_2=1}^{m+1} A_{i_1,i_2}\right) \prod_{i_2=m+2}^{N-m+1} \left[\frac{\mathbb{P}\left(\bigcap_{j=i_2-m}^{i_2} A_{i_1,j}\right)}{\mathbb{P}\left(\bigcap_{j=i_2-m}^{i_2-1} A_{i_1,j}\right)}\right],$$

et par stationnarité des A_{i_1,i_2}

$$\mathbb{P}\left(\bigcap_{i_2=1}^{N-m+1} A_{i_1,i_2}\right) \approx Q_{m,2m}\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)^{N-2m}, \tag{2.13}$$

où

$$Q_{m,m+l-1} = \mathbb{P}\left(A_{1,1} \cap A_{1,2} \cap \ldots A_{1,l}\right),\ 1 \leq l \leq N - m + 1,$$

sont estimés par simulations de Monte Carlo. Etant donné que nous devons scanner $N_m + 1$ *régions rectangulaires adjacentes de taille* $m \times N$, *l'approximation de (2.12) devrait être*

$$\mathbb{P} \approx \left[Q_{m,2m}\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)^{N-2m}\right]\left[Q_{m,2m}\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)^{N-2m}\right]^{N-m},$$

or [Chen and Glaz, 1996] ont remplacé dans le second membre $Q_{m,2m}$ *par* $Q_{m,2m}/Q_{m,2m-1}$ *pour prendre en compte la dépendance entre les* $\{A_{i_1,i_2}, 1 \leq i_2 \leq N + m - 1\}$ *pour les différentes valeurs de* i_1 *:*

$$\mathbb{P} \approx \left[Q_{m,2m}\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)^{N-2m}\right]\left[\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)^{N-2m}\right]^{N-m}$$

$$\mathbb{P} \approx Q_{m,2m-1}\left(\frac{Q_{m,2m}}{Q_{m,2m-1}}\right)^{(N-2m+1)(N-m+1)}. \tag{2.14}$$

L'équation (2.14) utilize l'approximation (2.13) pour $i_1 = 1$ *et pour* $2 \leq i_1 \leq N - m + 1$.

L'approximation (2.14) a été évaluée au travers d'études de simulations et se révèle être très précise [Glaz et al., 2001]. Néanmoins, nous ne disposons pas d'une erreur associée à l'approximation.

[Haiman and Preda, 2006] ont proposé une approximation basée sur le théorème 2.2 portant sur le maximum d'une suite stationnaire de variables aléatoires 1-dépendantes. Cette méthode d'approximation est décomposée en 2 étapes.

Première étape. Pour $L_1, L_2 \in \mathbb{N}^*$ tels que $N_1 = L_1 m_1$ et $N_2 = L_2 m_2$, posons

$$Z_{l_1} = \max_{\substack{(l_1-1)m_1+1 \leqslant t \leqslant l_1 m_1+1 \\ 1 \leqslant s \leqslant (L_2-1)m_2+1}} \nu_{t,s},\ 1 \leq l_1 \leq L_1 - 1.$$

La suite $\{Z_1, Z_2, \ldots, Z_{L_1-1}\}$ (Figure 2.3) forme une suite stationnaire de variables aléatoires 1-dépendantes et

$$\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k) = \mathbb{P}\left(\max_{1 \leq l_1 \leq L_1-1} \{Z_{l_1}\} \leq k\right).$$

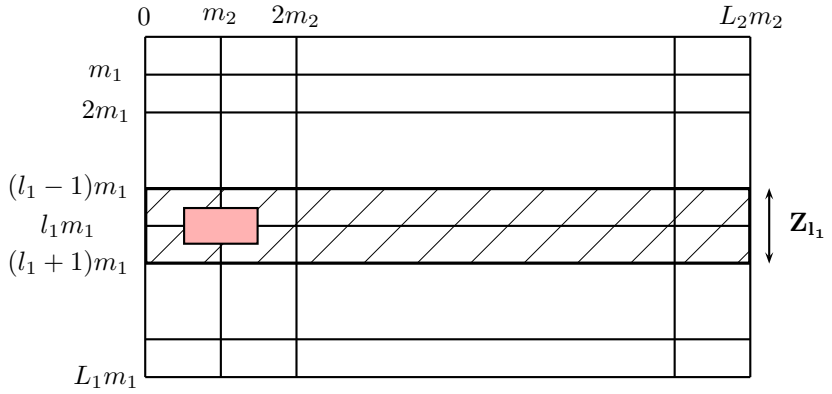


FIGURE 2.3 – Exemple de Z_{l_1}

Pour $i \in \{2, 3\}$ posons

$$Q_i = Q_i(k) = \mathbb{P}\left(\bigcap_{l_1=1}^{i-1} \{Z_{l_1} \leq k\}\right) = \mathbb{P}\left(\max_{1 \leq l_1 \leq i-1} \{Z_{l_1}\} \leq k\right). \tag{2.15}$$

Ainsi, si $1 - Q_2 \leq 0.025$ alors le théorème (2.2) peut être appliqué à la suite $\{Z_1, Z_2, \ldots, Z_{L_1-1}\}$ pour donner l'approximation suivante :

$$\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k) \approx \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{L_1-1}}, \tag{2.16}$$

avec une erreur relative de l'ordre de

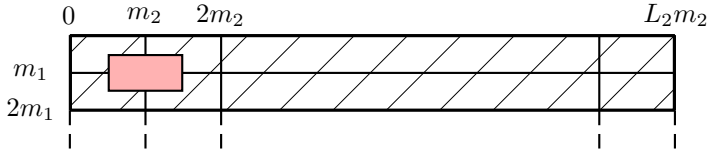
$$3.3(L_1 - 1)(1 - Q_2)^2. \tag{2.17}$$

Nous pouvons remarquer que Q_2 et Q_3 représentent, respectivement, la distribution de la statistique de scan à fenêtre de taille $m_1 \times m_2$ sur les régions rectangulaires $[0, 2m_1] \times [0, N_2]$ et $[0, 3m_1] \times [0, N_2]$ (Figure 2.4)

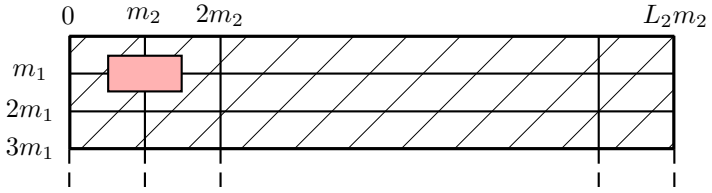
Ensuite, afin d'évaluer l'équation (2.16), il est nécessaire d'obtenir des approximations pour Q_2 et Q_3. Cela mène à la deuxième étape de la méthode.

Deuxième étape. Pour $i \in \{2, 3\}$, posons

$$Z_{i,l_2} = \max_{\substack{1 \leqslant t \leqslant (i-1)m_1+1 \\ (l_2-1)m_2+1 \leqslant s \leqslant l_2 m_2+1}} \nu_{t,s} \quad , 1 \leqslant l_2 \leqslant L_2 - 1.$$



(a) $Q_2 = \mathbb{P}(S(m_1, m_2, 2m_1, L_2m_2) \leqslant k)$



(b) $Q_3 = \mathbb{P}(S(m_1, m_2, 3m_1, L_2m_2) \leqslant k)$

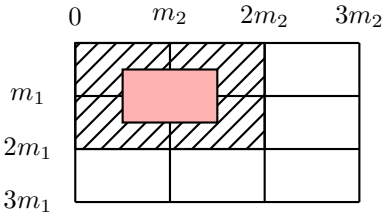
FIGURE 2.4 – Q_2 et Q_3

Nous pouvons remarquer que les suites $\{Z_{2,1}, Z_{2,2}, \ldots, Z_{2,L_2-1}\}$ et $\{Z_{3,1}, Z_{3,2}, \ldots, Z_{3,L_2-1}\}$ forment des suites stationnaires de variables aléatoires 1-dépendantes.

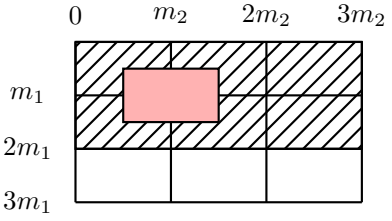
Pour $i, j \in \{2, 3\}$, posons

$$Q_{ij} = Q_{ij}(k) = \mathbb{P}\left(\bigcap_{l_2=1}^{j-1} \{Z_{i,l_2}\} \leq k\right) = \mathbb{P}\left(\max_{1 \leq l_2 \leq L_2-1} \{Z_{i,l_2}\} \leq k\right). \tag{2.18}$$

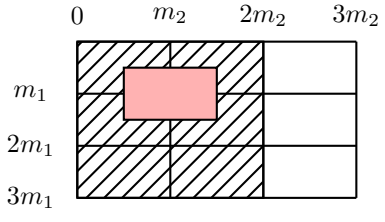
Nous pouvons observer que $Q_{22} = \mathbb{P}(Z_{2,1} \leqslant k) = \mathbb{P}(S(m_1, m_2, 2m_1, 2m_2) \leqslant k)$, $Q_{23} = \mathbb{P}(Z_{2,1}, Z_{2,2} \leqslant k) = \mathbb{P}(S(m_1, m_2, 2m_1, 3m_2) \leqslant k)$, $Q_{32} = \mathbb{P}(Z_{3,1} \leqslant k) = \mathbb{P}(S(m_1, m_2, 3m_1, 2m_2) \leqslant k)$ et $Q_{33} = \mathbb{P}(Z_{3,1}, Z_{3,2} \leqslant k) = \mathbb{P}(S(m_1, m_2, 3m_1, 3m_2) \leqslant k)$ (Figure 2.5).



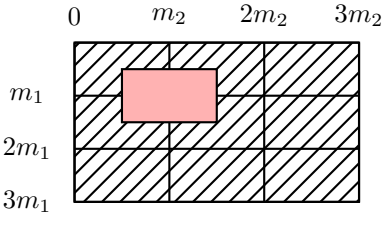
(a) Q_{22}



(b) Q_{23}



(c) Q_{32}



(d) Q_{33}

FIGURE 2.5 – Q_{22}, Q_{23}, Q_{32} and Q_{33}

La deuxième étape de la méthode consiste à appliquer une deuxième fois le théorème (2.2) afin de donner une approximation de Q_2 et Q_3 définis en (2.15). En effet, si $1 - Q_{22} \leq 0.025$ et

$1 - Q_{32} \leq 0.025$, le théorème (2.2) nous donne les approximations suivantes :

$$Q_2 \approx \frac{2Q_{22} - Q_{23}}{[1 + Q_{22} - Q_{23} + 2(Q_{22} - Q_{23})^2]^{L_2 - 1}} \tag{2.19}$$

avec une erreur d'approximation de l'ordre de $3.3(L_2 - 1)(1 - Q_{22})^2$,

$$Q_3 \approx \frac{2Q_{32} - Q_{33}}{[1 + Q_{32} - Q_{33} + 2(Q_{32} - Q_{33})^2]^{L_2 - 1}} \tag{2.20}$$

avec une erreur d'approximation de l'ordre de $3.3(L_2 - 1)(1 - Q_{32})^2$.

L'erreur relative à l'approximation de $\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k)$ en utilisant l'équation (2.16) est de l'ordre de

$$E_{app} = 3.3(L_2 - 1)(L_1 - 1)\left((1 - Q_{22})^2 + (1 - Q_{32})^2 + (L_2 - 1)(Q_{22} - Q_{23})^2\right). \tag{2.21}$$

Comme il n'existe pas de formules exactes pour Q_{22}, Q_{23}, Q_{32} and Q_{33}, leurs estimations sont réalisées au moyen de N simulations indépendantes. Pour $i, j \in \{2, 3\}$, posons $\widehat{Q}_{ij}$ l'estimation de Q_{ij} par le biais de simulations de Monte Carlo. L'erreur relative aux simulations est de l'ordre de

$$E_{sim} \approx (L_1 - 1) \times (L_2 - 1) \times 1.96\sqrt{\frac{(\widehat{Q}_{22} - \widehat{Q}_{33}) - (\widehat{Q}_{22} - 2\widehat{Q}_{23} + \widehat{Q}_{33})^2}{N}}. \tag{2.22}$$

Ainsi, l'erreur totale relative à l'approximation de $\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k)$ est

$$E_{tot} = E_{app} + E_{sim}. \tag{2.23}$$

Il est important de remarquer que E_{app} tend rapidement vers zéro lorsque $\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k)$ est proche de 1. De surcroît, pour des valeurs élevées de L_1 et L_2, la contribution de E_{app} à l'erreur totale, E_{tot}, est négligeable par rapport à E_{sim}. Aussi, la précision de l'approximation de $\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq k)$ dépend essentiellement du nombre de simulations, N, utilisées pour estimer Q_{22}, Q_{23}, Q_{32} and Q_{33}.

Remarque 2.5. *Si N_1 and N_2 ne sont pas multiples de m_1 et m_2 alors considérons $K = \lfloor \frac{N_1}{m_1} \rfloor$ and $L = \lfloor \frac{N_2}{m_2} \rfloor$. En utilisant*

$$\begin{aligned} \mathbb{P}(S \leq n) &\geq \mathbb{P}\left(S(m_1, m_2, (K+1)m_1, (L+1)m_2) \leqslant n\right) \\ \mathbb{P}(S \leq n) &\leq \mathbb{P}\left(S(m_1, m_2, Km_1, Lm_2) \leqslant n\right), \end{aligned}$$

on peut estimer $\mathbb{P}(S \leq n)$ par interpolation linéaire. Plus précisément, si la région scannée $\mathcal{R}$ est rectangulaire de taille $N_1 \times N_2$ alors l'approximation est donnée par interpolation linéaire entre la valeur v_1 obtenue en scannant la plus grande région rectangulaire $\mathcal{R}_1 = N_1^1 \times N_2^1$ contenue dans $\mathcal{R}$ et la valeur v_2 correspondant à la plus petite région rectangulaire $\mathcal{R}_2 = N_1^2 \times N_2^2$ contenant $\mathcal{R}$ avec $\mathcal{R}_1$ et $\mathcal{R}_2$ des régions de tailles multiples de $m_1 \times m_2$. La formule d'interpolation est donnée en Annexe B.

Exemples numériques. Les Tables 2.4 et 2.4 présentent des valeurs numériques résultant de l'approximation de $\mathbb{P}(S \leq k)$ par la méthode de Glaz (2.14) et celle d'Haiman (2.16) pour des évènements distribués selon un modèle binomial et Poisson.

k	$\mathbb{P}(S \leq k)$	App. (2.14)	App. (2.16)	Erreur
15	0.8596	0.8374	0.860427482	0.067409646
16	0.9402	0.9351	0.940749305	0.010867255
17	0.9783	0.9764	0.977260378	0.001546897
18	0.9930	0.9920	0.991966851	0.000217233

TABLE 2.3 – Approximation pour $\mathbb{P}(S \leq k)$: $X_{ij} \sim \mathcal{P}(0.25)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $N = 10^9$.

k	$\mathbb{P}(S \leq x)$	App. (2.14)	App. (2.16)	Erreur
15	0.8932	0.8830	0.896135764	0.035108915
16	0.9617	0.9577	0.960112719	0.004770939
17	0.9868	0.9862	0.986256278	0.000584065
18	0.9948	0.9958	0.995633424	8.08015E-05

TABLE 2.4 – Approximation pour $\mathcal{P}(S \leq k)$: $X_{ij} \sim \mathcal{B}(5, 0.05)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $N = 10^9$.

2.2.2 Statistique de scan bidimensionnelle continue

Modèle de Poisson. Soit une région rectangulaire $[0, L] \times [0, K]$, $\{L, K\} \in \mathbb{R}^2$ et $K, L < \infty$. Soit $N = \{N_{(t,s)}, t \in [0, L], s \in [0, K]\}$ un processus de Poisson homogène bidimensionnel d'intensité λ. Soient $\{u, v\} \in \mathbb{R}^{2*}$, $u < L$ et $v < K$. En considérant $S(u, v, K, L)$ définie en (1.8), nous cherchons à déterminer

$$\mathbb{P}(S(u, v, L, K) \leq k),\ k \geq 0,$$

ou encore

$$\mathbb{P}(S(u, v, L, K, \lambda) \leq k),\ k \geq 0,$$

dans le cadre du modèle de Poisson. Remarquons que

$$\mathbb{P}(S(u, v, L, K, \lambda) \leq k) = \mathbb{P}(S(1, 1, \frac{L}{u}, \frac{K}{v}, \lambda uv) \leq k),\ k \geq 0. \tag{2.24}$$

Afin de déterminer la distribution de $S(u, v, L, K, \lambda)$, [Alm, 1997] a proposé une première approche. Pour une valeur fixée de s, $0 < s < K$, remarquons que $N^{(1)}_{(t,s)} = N_{(t,s)}$ est un processus de Poisson homogène unidimensionnel d'intensité λv. Posons

$$S^{(1)}_u(s) = S(u, v, L, s : s + v) = \max_{0 \leq t \leq L-u} \left(N^{(1)}_{(t,s)}(t + u) - N^{(1)}_{(t,s)}(t) \right) = \max_{0 \leq t \leq L-u} \nu_{t,s},$$

à savoir la statistique de scan unidimensionnelle continue à fenêtre de taille $u \times v$ associée au processus $N^{(1)}_{(t,s)}$ sur $[0, L] \times [s, s + v]$. Pour s fixé, $\mathbb{P}(S^{(1)}_u(s) \leq k)$ peut être déterminée par (2.11). On en déduit que

$$\max_{0 \leq s \leq K-v} S^{(1)}_u(s) = S(u, v, L, K, \lambda).$$

[Alm, 1997] a proposé l'approximation suivante

$$\mathbb{P}(S(u, v, L, K, \lambda) \leq k) \approx \mathbb{P}(S^{(1)}_u(s) \leq k) e^{-\gamma k} \tag{2.25}$$

où

$$e^{-\gamma k} \approx \left(1 - \frac{\lambda uv}{k} \lambda uv \left(\frac{K}{v} - 1 \right) \mathbb{P}(S^{(1)}_u = k - 1) \right),$$

$\mathbb{P}(S_u^{(1)}(s) \leq k)$ et $\mathbb{P}(S_u^{(1)} = k-1)$ pouvant être approchées en utilisant (2.11), à savoir l'approximation de la distribution de la statistique de scan unidimensionnelle continue.

Comme dans le cas discret, cette formule d'approximation n'apporte pas d'erreur d'approximation.

Dans [Haiman and Preda, 2002], les auteurs ont proposé une approximation basée sur le maximum partiel d'une suite stationnaire de variables aléatoires 1-dépendantes. La méthodologie d'approximation est semblable à celle utilisée dans le cas discret (Section 2.2.1), à savoir deux utilisations successives du théorème (2.2).

Compte tenu de (2.24), supposons par la suite que $u = v = 1$ et $L, K \in \mathbb{N}$. Pour ν_{ts} défini en (1.7), posons

$$Q_{L,K} = Q_{L,K}(k,\lambda) = \mathbb{P}\left(\max_{\substack{0 \leqslant t \leqslant L-1 \\ 0 \leqslant s \leqslant K-1}} \nu_{ts} \leq k \right) = \mathbb{P}(S(1,1,L,K,\lambda) \leq k).$$

Observons que

$$X_j = \max_{\substack{0 \leq t \leq L-1 \\ j-1 \leq s \leq j}} \nu_{t,s},\ 1 \leq j \leq K-1,$$

forme une suite de variables aléatoires 1-dépendantes et

$$Q_{L,K} = \mathbb{P}\left(\max_{1 \leq j \leq K-1} X_j \leq k \right).$$

Pour $i \in \{2,3\}$ posons

$$Q_i = Q_i(k) = \mathbb{P}\left(\bigcap_{j=1}^{i-1} \{X_j \leq k\} \right) = \mathbb{P}\left(\max_{1 \leq j \leq i-1} \{X_j\} \leq k \right). \tag{2.26}$$

Aussi, si $1 - Q_2 \leq 0.025$ alors le théorème (2.2) peut être appliqué :

$$Q_{L,K} \approx \frac{2Q_2 - Q_3}{\left(1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2\right)^{K-1}}, \tag{2.27}$$

avec une erreur relative de l'ordre de $3.3(K-1)(1-Q_2)^2$. Ensuite, afin d'approximer Q_2 et Q_3, il suffit, comme dans le cas discret, d'appliquer une nouvelle fois le théorème (2.2). En effet, remarquons que

$$Y_i = \max_{\substack{i-1 \leq t \leq i \\ 0 \leq s \leq 1}} \nu_{ts},\ 1 \leq i \leq L-1,$$

forme une suite stationnaire de variables aléatoires 1-dépendantes et que

$$Q_2 = \mathbb{P}\left(\max_{1 \leq i \leq L-1} Y_i \leq k \right).$$

De manière analogue, remarquons que

$$Z_i = \max_{\substack{i-1 \leq t \leq i \\ 0 \leq s \leq 2}} \nu_{ts},\ 1 \leq i \leq L-1,$$

forme également une suite stationnaire de variables aléatoires 1-dépendantes et que

$$Q_3 = \mathbb{P}\left(\max_{1 \leq i \leq L-1} Z_i \leq k \right).$$

Si $1-Q_{2,2} \leq 0.025$ et $1-Q_{3,2} \leq 0.025$, le théorème (2.2) nous donne les approximations suivantes :

$$Q_2 \approx \frac{2Q_{2,2} - Q_{2,3}}{[1 + Q_{2,2} - Q_{2,3} + 2(Q_{2,2} - Q_{2,3})^2]^{L-1}} \tag{2.28}$$

avec une erreur d'approximation de l'ordre de $3.3(L-1)(1-Q_{2,2})^2$,

$$Q_3 \approx \frac{2Q_{3,2} - Q_{3,3}}{[1 + Q_{3,2} - Q_{3,3} + 2(Q_{3,2} - Q_{3,3})^2]^{L-1}} \tag{2.29}$$

avec une erreur d'approximation de l'ordre de $3.3(L-1)(1-Q_{3,2})^2$. En substituant Q_2 et Q_3 dans l'équation (2.27) par leur approximations (2.28) et (2.29), on peut facilement vérifier que l'erreur totale résultant de l'approximation de $Q_{L,K}$ est bornée par

$$E_{app} = 3.3(L-1)(K-1)\left((1-Q_{2,2})^2 + (1-Q_{2,3})^2 + (L-1)(Q_{2,2} - Q_{3,2})^2\right).$$

Afin d'appliquer cette méthode d'approximation, nous avons besoin d'évaluer de manière précise les $Q_{i,j}$, $(i,j) \in \{(2,2),(2,3),(3,3)\}$, car $Q_{2,3} = Q_{3,2}$. Cela peut être effectué par des simulations. Or, la simulation d'un processus de Poisson bidimensionnelle homogène se base souvent sur la propriété d'uniformité conditionnelle qui stipule que conditionnellement à $N([0,L]\times[0,K]) = n$, les évènements sont distribués selon une loi uniforme sur $[0,L]\times[0,K]$. Cependant, du fait du conditionnement, cette approche ne permet pas de simuler directement les $Q_{i,j}$ car nous sommes en présence d'une statistique de scan conditionnelle. Néanmoins, il est possible de faire le lien les cas conditionnel et non-conditionnel. Aussi, considérons $Q^n_{L,K}$ la statistique de scan bidimensionnelle continue conditionnelle telle que

$$Q^n_{L,K} = Q^n_{L,K}(k,\lambda) = \mathbb{P}\left(S(u,v,L,K,\lambda) \leq k / N([0,L]\times[0,K]) = n\right),\ 1 \leq k \leq n.$$

Le théorème des probabilités totales nous permet d'exprimer $Q_{L,K}$ en fonction de $Q^n_{L,K}$:

$$Q_{L,K}(k) = \sum_{n=0}^{\infty} Q^n_{L,K}(k)\mathbb{P}(N([0,L]\times[0,K]) = n).$$

Or pour $0 \leq n \leq k$, $Q^m_{L,K} = 1$ et pour $n > kLK$, $Q^n_{L,K} = 0$, donc

$$Q_{L,K}(k) = \sum_{n=0}^{k} \mathbb{P}(N([0,L]\times[0,K]) = n) + \sum_{n=k+1}^{kLK} Q^n_{L,K}(k)\mathbb{P}(N([0,L]\times[0,K]) = n),$$

$$Q_{L,K}(k) = e^{-\lambda LK}\left(\sum_{n=0}^{k} \frac{(\lambda KL)^n}{n!} + \sum_{n=k+1}^{kLK} Q^n_{L,K}(k)\frac{(\lambda KL)^n}{n!}\right). \tag{2.30}$$

Ainsi, la simulation des Q_{ij} est indirectement réalisée par la simulation de Monte Carlo des Q^n_{ij}. Nous allons décrire la manière de simuler ces derniers. Remarquons tout d'abord si $\pi_1,\ldots,\pi_n$ sont des points répartis de manière uniforme dans le rectangle $[0,i]\times[0,j]$, $(i,j) \in \{(2,2),(2,3),(3,3)\}$ et

$$\nu^{n,(i,j)}_{ts} = \#\{k | \pi_k \in [s,s+1]\times[t,t+1]\},\ (t,s) \in [0,i-1]\times[0,j-1],$$

alors

$$\mathbb{P}\left(\max_{\substack{0\leq t\leq i-1\\ 0\leq s\leq j-1}} \nu^{n,(i,j)}_{ts} \leq k\right) = \mathbb{P}\left(S^n_{i,j} \leq k\right) = Q^n_{i,j}(k),\ 1 \leq k \leq n,\ i,j = 2,3.$$

Afin d'approximer les Q_{ij} posons $M = M(\lambda) \in \mathbb{N}^*$.

1. Pour $2 \leq m \leq M$, utiliser des simulations de Monte Carlo pour générer R réalisations indépendantes des S_{ij}^m et poser $\widehat{Q_{ij}^m}(k)$, correspondant à la fonction de distribution empirique des $Q_{ij}^m(k)$, $1 \leq k \leq m$.
2. $Q_{ij}^m(k)$ définie en (2.30) est estimée par

$$\widehat{Q_{ij}(k)} = e^{-\lambda ij}\left(\sum_{m=0}^{k}\frac{(\lambda ij)^m}{m!} + \sum_{k+1}^{\min\{M,kij\}}\widehat{Q_{ij}^m}(k)\frac{(\lambda ij)^m}{m!}\right),\ i,j=2,3,\ 1\leq k\leq m. \quad (2.31)$$

Remarque 2.6. *$M = M(\lambda)$ est un quantile de la loi de Poisson tel que $\mathbb{P}(N([0,L]\times[0,K]) \leq M(\lambda)) = \tau$, avec une valeur de τ très faible (ex : 0.00001). Cette technique permet de réduire le nombre d'estimations des $\widehat{Q_{ij}^m}(k)$, $2 \leq m \leq M(\lambda)$ et ainsi réduire le temps de calcul.*

De par le théorème central-limite, l'erreur d'estimation des Q_{ij}^m, ϵ_{ij}^m, est bornée, au niveau de confiance de 95%, par

$$\epsilon_{ij}^m = 1.96\sqrt{\frac{Q_{ij}^m(1-Q_{ij}^m)}{R}}.$$

Ainsi, par la formule (2.31), l'erreur d'approximation de $Q_{ij}(k)$, notée $\epsilon_{ij}(k)$, au niveau de confiance de 95%, est bornée par

$$\epsilon_{ij}(k) = e^{-\lambda ij}\sum_{k+1}^{\min\{M,kij\}}\epsilon_{ij}^m(k)\frac{(\lambda ij)^m}{m!} + \widehat{Q_{ij}^M}(k)e^{-\lambda ij}\sum_{m=\min\{M,kij\}+1}^{kij}\frac{(\lambda ij)^m}{m!}, \quad (2.32)$$

le second terme en (2.32) disparaissant si $M \geq kij$. En prenant désormais en compte les erreurs ϵ_{ij}, l'erreur relative à $Q_{L,K}$ est de l'ordre de

$$E_{tot} = E_{app} + LK(\epsilon_{22} + \epsilon_{23} + \epsilon_{33}).$$

Exemples numériques. La Table 2.5 présente des approximations de $\mathbb{P}(S \leq k)$ en utilisant l'approximation de Alm (2.25) et celle d'Haiman (2.27) dans un cas particulier.

k	App. (2.27)	Erreur	App. (2.25)
2	0.69318103	0.008570775	0.7839302629
3	0.998401542	6.37679E-05	0.9987785770
4	0.999994866	4.39515E-07	0.9999959179

TABLE 2.5 – Approximation de $\mathbb{P}(S \leq k)$ en utilisant l'approximation de Alm (2.25) et celle d'Haiman (2.27). $L = 500$, $K = 500$ et $\lambda = 0.01$

3 Cas conditionnel

Lorsque le nombre total d'évènements observés sur la région étudiée est connu, nous nous plaçons dans le cadre de la statistique de scan conditionnelle. Ce conditionnement introduit une dépendance entre les observations qui ne rend plus applicables les formules d'approximation basées sur les propriétés du maximum partiel d'une suite stationnaire de variables aléatoires 1-dépendantes. Cependant, les formules d'approximation basées sur des raisonnements markoviens peuvent être adaptées au cas conditionnel. C'est ce que nous présentons dans cette section.

3.1 Statistique de scan unidimensionnelle

3.1.1 Statistique de scan unidimensionnelle discrète

Modèle de Bernoulli. Considérons $X_1, \ldots, X_N$ une suite de variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli, $\mathcal{B}(1,p)$. Supposons que nous savons que a succès et $N-a$ échecs ont été observés. Nous sommes intéressés par le calcul de

$$\mathbb{P}(S(m,N,a) \leq k) = \mathbb{P}\left(S(m,N) \leq k \Big/ \sum_{i=1}^{N} X_i = a \right). \tag{2.33}$$

Dans [Naus, 1974, Théorème 1], les auteurs ont proposé une formule exacte pour (2.33) :

Théorème 2.6. *Pour k, m, $L = N/m \in \mathbb{N}$ et $2 \leq k \leq a$. Soient $(n_1, n_2, \ldots, n_L)$ une partition de a en L entiers positifs ou nuls et Θ_k l'ensemble des permutations de toutes les partitions de a en L entiers positifs inférieurs ou égaux à k, nous avons*

$$\mathbb{P}\left(S(m,N) \leq k \Big/ \sum_{i=1}^{N} X_i = a \right) = \frac{(m!)^L}{\binom{N}{a}} \sum_{\sigma \in \Theta_k} det|d_{ij}|, \tag{2.34}$$

où

$$d_{ij} = \begin{cases} 0 & si\ c_{ij} < 0\ ou\ c_{ij} > m \\ \frac{1}{c_{ij}!(m-c_{ij})!} & sinonn \end{cases}$$

avec

$$\begin{aligned} c_{ij} = &\quad (j-i)k - \textstyle\sum_{r=1}^{j-1} n_r + n_i \quad pour\ i < j \\ = &\quad (j-i)k + \textstyle\sum_{r=j}^{i} n_r \quad pour\ i \geq j. \end{aligned}$$

$|d_{ij}|$ correspond à une matrice de dimension $L \times L$.

Dans le cas particulier où $k > a/2$, une formule plus simple a été proposée par [Naus, 1974, Corollaire 2] :

$$\mathbb{P}\left(S(m,N) \geq k \Big/ \sum_{i=1}^{N} X_i = a \right) = \frac{2\sum_{s=k}^{a} \binom{m}{a}\binom{N-m}{a-s}}{\binom{N}{a}} + (Lk - a - 1)\frac{\binom{m}{k}\binom{N-m}{a-k}}{\binom{N}{a}}$$

Remarque 2.7. *Lorsque N, m et L sont grands et $k \leq a/2$ le calcul de (2.33) en utilisant l'équation (2.34) devient extrêmement consommateur en termes de temps de calcul, un déterminant d'une matrice de dimension $L \times L$ devant être calculé pour chaque élément de l'ensemble Θ_k.*

Dans [Chen and Glaz, 1999], les auteurs ont proposé une approximation de type *Product-type* pour (2.33) :

Théorème 2.7. *Pour $N, m, k \in \mathbb{N}$, $L = N/m, k \geq 0$ et $L \geq 2$, la distribution de $S(m,N,a)$ peut être approximée par*

$$\mathbb{P}\left(S(m,N) \leq k \Big/ \sum_{i=1}^{N} X_i = a \right) = q_{3m}(a) \left[\frac{q_{3m}(a)}{q_{2m}(a)} \right]^{L-3}, \tag{2.35}$$

avec

$$q_{rm}(a) = \sum_{j=0}^{\min rk-r,a} q(rw|j) \frac{\binom{rm}{j}\binom{N-rm}{a-j}}{\binom{N}{a}},\ r \in 2,3,$$

et

$$q(rm|j) = \mathbb{P}\left(S(m,rm) \leq k \Big/ \sum_{i=1}^{N} X_i = j \right),$$

pouvant être évalué en utilisant (2.34).

Preuve. *La démonstration repose sur la même technique que celle utilisée dans le cas non-conditionnel. Pour $N = Lm$, $L \geq 2$ et $1 \leq i \leq L-1$ et ν_t défini en (1.1), posons*

$$E_i = \left\{ \max_{(i-1)m+1 \leq t \leq im+1} \nu_t \leq k \right).$$

$$\mathbb{P}(S(m,N,a) \leq k) = \mathbb{P}\left(\bigcap_{i=1}^{L-1} E_j \right) = \mathbb{P}(E_1) \prod_{i=2}^{L-1} \mathbb{P}\left(E_j \middle| \bigcap_{j=1}^{i-1} E_j \right)$$

En utilisant un raisonnement de type markovien sur la dépendance entre les E_i nous avons

$$\mathbb{P}(S(m,N,a) \leq k) \approx \mathbb{P}(E_1) \prod_{i=2}^{L-1} \mathbb{P}(E_i | E_{i-1}) \approx \mathbb{P}(E_1 \cap E_2) \prod_{i=2}^{L-2} \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)}.$$

Comme $\mathbb{P}(E_1) = q_{2m}(a)$ et $\mathbb{P}(E_1 \cap E_2) = q_{3m}(a)$, nous avons

$$\mathbb{P}(S(m,N,a) \leq k) \approx q_{3m}(a) \left[\frac{q_{3m}(a)}{q_{2m}(a)} \right]^{L-3}.$$

Modèle Binomial. Soient $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires i.i.d. selon une loi binomiale $\mathcal{B}(n,p)$. Supposons que le nombre total d'évènements $\sum_{i=1}^N X_i = a$ ait été observé. Dans ce cas, la distribution jointe des X_i, $1 \leq i \leq N$, conditionnellement à l'évènement $\{\sum_{i=1}^N X_i = a\}$, est une loi hypergéométrique multivariée

$$\mathbb{P}\left(X_i = x_i,\ 1 \leq i \leq N \middle/ \sum_{i=1}^N X_i = a \right) = \frac{\binom{n}{x_1}\binom{n}{x_2}\cdots\binom{n}{a-\sum_{i=1}^{N-1} x_i}}{\binom{nN}{a}}. \tag{2.36}$$

Dans [Glaz et al., 2001], les auteurs ont proposé d'utiliser l'approximation (2.35). Cependant, à l'inverse du modèle de Bernoulli, il n'existe pas de formule exacte pour $q_{rm|j}$ défini en (2.36). Aussi, les auteurs ont proposé d'utiliser l'algorithme de simulation de loi hypergéométrique multivariée développé dans [Patefield, 1981] pour simuler les valeurs de $q_{2m}(a)$ et $q_{3m}(a)$, estimées par $\widehat{q_{2m}}(a)$ et $\widehat{q_{3m}}(a)$. Soient $V_1, V_2, \ldots, V_N$ une suite de variables aléatoires distribuées selon une loi hypergéométrique multivariée définie en (2.36). Pour $r \in \{2,3\}$, nous avons

$$q_{rm}(a) = \mathbb{P}\left(\bigcap_{s=1}^{r-1} E_s \right) = \mathbb{P}\left[\bigcap_{s=1}^{r-1} \left(\sum_{i=(s-1)m+1}^{(s+1)m} V_i \leq k \right) \right].$$

Aussi, il suffit de réaliser R simulations des variables $V_1, V_2, \ldots V_{3m}$ de distribution jointe

$$\mathbb{P}(V_1 = v_1, V_2 = v_2, \ldots, V_{3m} = v_{3m}) = \frac{\binom{n}{v_1}\binom{n}{v_2}\cdots\binom{n}{a-\sum_{i=1}^{3m} V_i}}{\binom{nN}{a}}.$$

L'estimation de $q_{rm}(a)$, $2 \leq r \leq 3$ est donnée par

$$\widehat{q_{rm}}(a) = \frac{1}{R} \sum_{i=1}^{R} \mathbb{1}_{\{S(m,rm,a) \leq k\}}. \tag{2.37}$$

Modèle de Poisson. Soient $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires i.i.d. selon une loi de Poisson $\mathcal{P}(\lambda)$. Supposons que le nombre total d'évènements $\sum_{i=1}^N X_i = a$ ait été observé. Dans ce cas, la distribution jointe des X_i, $1 \leq i \leq N$, conditionnellement à l'évènement $\{\sum_{i=1}^N X_i = a\}$, est une loi multinomiale

$$\mathbb{P}\left(X_i = x_i,\ 1 \leq i \leq N \middle/ \sum_{i=1}^N X_i = a \right) = \binom{a}{x_1, x_2, \ldots, x_N} \left(\frac{1}{N} \right)^a. \tag{2.38}$$

Dans [Glaz et al., 2001], les auteurs ont proposé d'utiliser l'approximation (2.35). Comme dans le cas du modèle binomial, il n'existe pas de formules exactes pour $q_{2m}(a)$ et $q_{3m}(a)$. Aussi, leur valeur est estimée par $\widehat{q_{2m}}(a)$ et $\widehat{q_{3m}}(a)$, obtenus par simulations. Soient $W_1, W_2, \dots W_N$ une suite de variables aléatoires distribuée selon une loi multinomiale définie en (2.38). Pour $r \in \{2, 3\}$, nous avons

$$q_{rm}(a) = \mathbb{P}\left(\bigcap_{s=1}^{r-1} E_s\right) = \mathbb{P}\left[\bigcap_{s=1}^{r-1}\left(\sum_{i=(s-1)m+1}^{(s+1)m} W_i \leq k\right)\right].$$

Aussi, il suffit de réaliser R simulations des variables $W_1, W_2, \dots W_{3m}$ de distribution jointe

$$\mathbb{P}(W_1 = w_1, W_2 = w_2, \dots, W_{3m} = w_{3m}) = \binom{a}{w_1, w_2, \dots, w_{3m}}\left(\frac{1}{N}\right)^a.$$

L'estimation de $q_{rm}(a)$, $2 \leq r \leq 3$, est identique à (2.37).

3.1.2 Statistique de scan unidimensionnelle continue

Modèle de Poisson. Soit $N = \{N_t\}_{t\geq 0}$ un processus de Poisson homogène d'intensité $\lambda, \lambda > 0$, défini sur l'intervalle $[0, T]$, $T \in \mathbb{R}^+$ fixé. Supposons que $N(T) = a$ évènements aient été observés sur $[0, T]$. En considérant $S(m, T, a)$ définie en (1.4), nous cherchons à déterminer

$$\mathbb{P}(S(m, T, a) \leq k),\ k \geq 0. \tag{2.39}$$

En prenant en compte le conditionnement, considérons $X_1, \dots, X_a$ une suite de variables aléatoires i.i.d. selon une loi uniforme sur $[0, T]$, $\mathcal{U}[0, T]$, qui correspond aux temps d'occurrence des a évènements. Soient $X_{(1)} < X_{(2)} < \dots < X_{(a)}$ les statistiques d'ordre associées telles que

$$X_{(1)} = \min_{1\leq i\leq a} X_i \ \text{ et } \ X_{(a)} = \max_{1\leq i\leq a} X_i.$$

Déterminer $\mathbb{P}(S(m, T, a) \leq k)$ est équivalent à

$$\mathbb{P}(S(m, T, a) \leq k) = \mathbb{P}\left(\min_{1\leq i\leq a-k}\left(X_{(i+k)} - X_{(i)}\right) \geq m\right).$$

L'évènement $\left\{\min_{1\leq i\leq a-k} X_{(i+k)} - X_{(i)} \geq m\right\}$ signifie que la distance minimale entre k évènements est supérieure à la taille de la fenêtre de scan, m. Autrement dit, il n'existe pas de fenêtre m contenant plus de k évènements.

Dans [Naus, 1965b], les auteurs ont proposé une formule exacte pour (2.39)

Théorème 2.8. *Pour $k, m, L = T/m \in \mathbb{N}$, $2 \leq k \leq a$ et $L \geq 2$. Soient $(n_1, n_2, \dots, n_L)$ une partition de a en L entiers positifs ou nuls et Θ_k l'ensemble des permutations de toutes les partitions de a en L entiers positifs inférieurs ou égaux à k, nous avons*

$$\mathbb{P}(S(m, T, a) \leq k) = a!L^{-a}\sum_{\sigma\in\Theta_k} \det[1/c_{ij}!], \tag{2.40}$$

où

$$\begin{aligned} c_{ij} &= (j-i)k - \textstyle\sum_{r=1}^{j-1} n_r + n_i \quad \textit{pour } i < j \\ &= (j-i)k + \textstyle\sum_{r=j}^{i} n_r \quad \textit{pour } i \geq j. \end{aligned}$$

Remarque 2.8. *A l'instar du cas unidimensionnel discret conditionnel, le calcul de (2.39) en utilisant (2.40) devient extrêmement consommateur en temps de calcul, voire irréalisable, pour des valeurs élevées de a, ou lorsque L est très important face à m.*

Dans [Naus, 1982], les auteurs ont proposé une approximation de (2.39) basée sur un raisonnement markovien. Compte tenu du conditionnement en a évènements observés sur $[0,T]$, le principe réside dans le fait d'approximer la distribution de $S(m,N)$ définie en (1.2), où $N=a$, sur la suite de variables aléatoires X_i, $1 \leq i \leq a$.

Théorème 2.9. *Soient $X_1,\ldots,X_a$ une suite variables aléatoires i.i.d. selon une loi uniforme sur $[0,T]$. Pour $L=T/m$, $L \geq 4$ et $L \in \mathbb{N}$, nous avons*

$$\mathbb{P}(S(m,a) \leq k) \approx Q_2 \left(\frac{Q_3}{Q_2}\right)^{L-2}, \; k \geq 0. \tag{2.41}$$

où

$$Q_2 = \mathbb{P}(S(m,2m) \leq k)$$
$$Q_3 = \mathbb{P}(S(m,3m) \leq k)$$

sont évalués en utilisant le théorème 2.40.

Preuve. *Pour $1 \leq i \leq L-1$ et ν_t défini en (1.1), posons*

$$E_i = \left\{ \max_{(i-1)m \leq t \leq im} \nu_t \leq k \right\}, \; k \geq 0.$$

Partant de la remarque suivante

$$\mathbb{P}(S(m,a) \leq k) = \mathbb{P}\left(\bigcap_{i=1}^{L-1} E_i \right),$$

la preuve est identique à celle utilisée dans le cadre du théorème 2.1.

Dans [Glaz, 1992], les auteurs ont proposé une autre méthode d'approximation basée sur l'intersection des statistiques d'ordre.

Théorème 2.10. *Soient $X_1,\ldots,X_a$ une suite de variables aléatoires i.i.d. selon une loi uniforme sur $[0,T]$. Soit $X_{(1)} < X_{(2)} < \ldots < X_{(a)}$ les statistiques d'ordre associées. Pour $3 \leq k \leq a$, $1 \leq i \leq a-k$ et $0 < m < T/2$, définissons les évènements*

$$A_i = \{X_{(k+i)} - X_{(i)} \geq m\}.$$

Pour $2 \leq t \leq k \leq a/2$ posons

$$\alpha_1 = \mathbb{P}(A_1), \quad \alpha_t = \mathbb{P}\left(\bigcap_{j=1}^{k} A_j \right).$$

Alors nous avons

$$\mathbb{P}(S(m,a) \leq k) \approx \alpha_k \left(\frac{\alpha_k}{\alpha_{k-1}}\right)^{a-2k+1}. \tag{2.42}$$

Preuve (Esquisse)**.** *Par définition des A_i, $1 \leq i \leq a-k+1$, nous avons*

$$\mathbb{P}(S(m,a) \leq k) = \mathbb{P}\left(\bigcap_{i=1}^{a-k+1} A_i \right) = \alpha_{a-k+1}.$$

Or

$$\alpha_{a-k+1} = \mathbb{P}\left(A_1 \cap \ldots \cap A_k\right) \prod_{j=k+1}^{a-k+1} \mathbb{P}\left(A_j / A_1 \cap A_2 \cap \ldots \cap A_{j-1}\right).$$

Lorsque $j > k$, faisons une hypothèse de comportement markovien sur la dépendance entre les A_j, à savoir que le conditionnement sur le passé entier peut être approximé par le conditionnement sur les k derniers évènements :

$$\begin{aligned}\mathbb{P}\left(A_j/A_1 \cap A_2 \cap \ldots \cap A_{j-1}\right) &\approx \mathbb{P}\left(A_j/A_{j-1} \cap A_{j-2} \cap \ldots \cap A_{j-k}\right) \\ &= \mathbb{P}\left(A_k/A_1 \cap A_2 \cap \ldots \cap A_{k-1}\right) = \frac{\alpha_k}{\alpha_k - 1}.\end{aligned}$$

D'où

$$\alpha_{a-k+1} = \alpha_k \left(\frac{\alpha_k}{\alpha_{k-1}}\right)^{a-2k+1}.$$

Remarque 2.9. *L'approximation de (2.42) nécessite le calul de α_k et α_{k-1}. Pour $2 \leq t \leq k \leq a/2$ posons*

$$\alpha_1^* = \mathbb{P}(A_1^c), \quad \alpha_t^* = \mathbb{P}\left(A_1 \cap \left[\bigcap_{j=2}^{k} A_j^c\right]\right).$$

Nous admettrons, selon [Glaz, 1992], que pour $2 \leq i \leq a-k$

$$\alpha_i = \alpha_1 - \sum_{j=2}^{i} \alpha_j^*. \tag{2.43}$$

Pour $3 \leq t \leq a/2$ et $0 < m < T/2$ et $T = 1$, [Glaz, 1992, Théorème 1] montre que

$$\alpha_t^* = b(k-1, a, m) - b(k, a, m) + \sum_{j=t}^{a-k+1} (-1)^j \prod_{i=1}^{} k - 2\left[1 - \frac{j(j-1)}{i(i+1)}\right] b(k+j-1, a, m),$$

où $b(j; a; m) = \binom{a}{j} m^k (1-m)^{a-k}$ est la loi de probabilité de la loi binomiale $\mathcal{B}(a, m)$. Par ailleurs, le fait que ce théorème s'applique uniquement lorsque $T = 1$ ne constitue pas une limite car tout intervalle $[0, T]$, $T > 1$, peut être rapporté à un intervalle $[0, 1]$.

3.2 Statistique de scan bidimensionnelle

3.2.1 Statistique de scan bidimensionnelle discrète

Soit une région rectangulaire $[0, N_1] \times [0, N_2]$ avec $N_1, N_2 \in \mathbb{N}$. Considérons $\{X_{i,j}\}$, $1 \leq i \leq N_1$ et $1 \leq j \leq N_2$ un ensemble de variables aléatoires indépendantes, identiquement distribuées et à valeurs dans $\mathbb{N}$. Chaque variable aléatoire $X_{i,j}$ désigne le nombre d'évènements observés dans une région élémentaire de dimension $[i-1, i] \times [j-1, j]$. De surcroît, le nombre total d'évènements,

$$\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} X_{ij} = a, \tag{2.44}$$

est considéré connu. En considérant $S(m_1, m_2, N_1, N_2, a)$ définie en (1.6), nous cherchons à évaluer

$$\mathbb{P}(S(m_1, m_2, N_1, N_2, a) \leq k) = \mathbb{P}\left(S(m_1, m_2, N_1, N_2) \leq k \Big/ \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} X_{ij} = a\right), \; k \geq 0.$$

Dans [Chen and Glaz, 2002], les auteurs ont proposé une approximation de type "Product-type"

Théorème 2.11. *Pour $m_1 = m_2 = m$ et $N_1 = N_2 = N$, $m \in \mathbb{N}^*$, $2 \leq m\ N-1$, $N \in \mathbb{N}$ et $L = N/m$, $L \in \mathbb{N}$, $k \geq 0$, nous avons*

$$\mathbb{P}(S(m_1, m_2, N_1, N_2, a) \leq k) \approx q_{m,N}(a) \left(\frac{q_{m+1,N}(a)}{q_{m,N}(a)}\right)^{L-1}, \tag{2.45}$$

avec

$$q_{m,N}(a) = \mathbb{P}(S(m,m,m,N,a) \leq k),$$

et

$$q_{m+1,N}(a) = \mathbb{P}(S(m,m,m+1,N,a) \leq k).$$

Preuve. *Pour* $1 \leq i_1 \leq N-m+1$, $1 \leq j \leq N-m+1$ *et* ν_{ts} *définis en (1.5) , posons*

$$A_{i_1,i_2} = \{\nu_{i_1,i_2} \leq k\}\,,\ k \geq 0.$$

Posons

$$B = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij} = a \right\}.$$

Remarquons que

$$\mathbb{P}(S(m_1,m_2,N_1,N_2,a) \leq k) = \mathbb{P}\left(\bigcap_{i_1=1}^{N-m+1} \bigcap_{i_2=1}^{N-m+1} A_{i_1,i_2} \Big/ B \right).$$

Pour $1 \leq i_1 \leq N-m$, *posons*

$$E_{i_1} = \bigcap_{i_2=1}^{N-m+1} A_{i_1,i_2}.$$

Aussi

$$\mathbb{P}(S(m_1,m_2,N_1,N_2,a) \leq k) = \mathbb{P}\left(\bigcap_{i_1=1}^{N-m+1} E_{i_1} \Big/ B \right),$$

qui peut s'écrire sous la forme

$$\mathbb{P}\left(\bigcap_{i_1=1}^{N-m+1} E_{i_1} \Big/ B \right) = \mathbb{P}(E_1/B) \prod_{i_1=2}^{N-m+1} \left[\frac{\mathbb{P}\left(\bigcap_{j=1}^{i_1} E_j \Big/ B \right)}{\mathbb{P}\left(\bigcap_{j=1}^{i_1-1} E_j \Big/ B \right)} \right].$$

En utilisant des hypothèses simplificatrices de type markovien (absence de mémoire et échangeabilité) sur la dépendance entre les E_{i_1}

$$\frac{\mathbb{P}\left(\bigcap_{j=1}^{i_1} E_j \Big/ B \right)}{\mathbb{P}\left(\bigcap_{j=1}^{i_1-1} E_j \Big/ B \right)} \approx \frac{\mathbb{P}\left(\bigcap_{j=i_1-1}^{i_1} E_j \Big/ B \right)}{\mathbb{P}\left(\bigcap_{j=i_1-1}^{i_1-1} E_j \Big/ B \right)} \approx \frac{\mathbb{P}\left(E_1 \cap E_2/B\right)}{\mathbb{P}\left(E_1/B\right)},$$

nous arrivons à une approximation de type "Product-type"

$$\mathbb{P}(S(m_1,m_2,N_1,N_2,a) \leq k) \approx \mathbb{P}(E_1/B) \left(\frac{\mathbb{P}(E_1 \cap E_2/B)}{\mathbb{P}(E_1/B)} \right)^{L-1}.$$

Or $q_{m,N}(a) = \mathbb{P}(S(m,m,m,N,a) \leq k) = \mathbb{P}(E_1/B)$ *et* $q_{m+1,N}(a) = \mathbb{P}(S(m,m,m+1,N,a) \leq k) = \mathbb{P}(E_1 \cap E_2/B)$, *donc*

$$\mathbb{P}(S(m_1,m_2,N_1,N_2,a) \leq k) \approx q_{m,N}(a) \left(\frac{q_{m+1,N}(a)}{q_{m,N}(a)} \right)^{L-1}.$$

Il n'existe pas de formules exactes pour $q_{m,N}(a)$ et $q_{m+1,N}(a)$. Dans [Chen and Glaz, 2002], les auteurs ont proposé d'estimer ces quantités par simulation dont la méthodologie est fonction du modèle régissant les X_{ij}, $1 \leq i \leq N_1$, $1 \leq j \leq N_2$.

Modèle Binomial. Posons $\{X_{i,j}\}$, $1 \leq i,j \leq N$ une famille de variables aléatoires i.i.d. selon une loi binomiale $\mathcal{B}(n,p)$, $n \geq 1$, $0 < p < 1$. Considérons a, défini en (2.44), le nombre total d'évènements observés, connu. La distribution jointe des $X_{i,j}$, conditionnellement à a, est une loi hypergéométrique multivariée

$$\mathbb{P}(X_{i,j} = x_{i,j},\ 1 \leq i,j \leq N/B) = \frac{\binom{n}{x_{1,1}}\binom{n}{x_{1,2}}\cdots\binom{n}{x_{N,N}}}{\binom{N^2 n}{a}}. \tag{2.46}$$

où $x_{i,j} \geq 0$ et $x_{N,N} = a - \sum_{j=1}^{N-1}\sum_{i=1} N x_{i,j} \geq 0$. Soient $V_{1,1},\ldots,V_{N,N}$ une suite de variables aléatoires distribuées selon la loi hypergéométrique définie en (2.46). Pour $t \in \{1,2\}$, nous avons

$$q_{m+t-1,N}(a) = \mathbb{P}\left(\bigcap_{s=1}^{t} E_s \middle/ B\right) = \mathbb{P}\left[\bigcap_{s=1}^{t}\bigcap_{i_1=1}^{m+t-1}\sum_{j=i_2}^{i_2+m-1} V_{i,j} \leq k\right].$$

Aussi, il suffit de réaliser R simulations, en utilisant l'algorithme développé dans [Patefield, 1981], des variables $V_{1,1},\ldots,V_{m+1,N}$ dont la distribution jointe est

$$\mathbb{P}(V_{i,j} = v_{i,j},\ 1 \leq i \leq m+1,\ 1 \leq j \leq N) = \frac{\binom{n}{v_{1,1}}\cdots\binom{n}{v_{m+1,N}}\binom{N(N-m-2)n}{a-\sum_{i=1}^{m+2}\sum_{j=1}^{N} v_{i,j}}}{\binom{N^2 n}{a}},$$

pour estimer les valeurs de $q_{m+t-1,N}(a)$, $1 \leq t \leq 2$, données par

$$\widehat{q}_{m+t-1,N}(a) = \frac{\sum_{i=1}^{R} \mathbb{1}_{\{S(m,m,m+t-1,N,a) \leq k\}}}{R}. \tag{2.47}$$

Modèle de Poisson. Posons $\{X_{i,j}\}$, $1 \leq i,j \leq N$ une famille de variables aléatoires i.i.d. selon une loi de Poisson $\mathcal{P}(\lambda)$. Considérons a, défini en (2.44), le nombre total d'évènements observés, connu. La distribution jointe des $X_{i,j}$, conditionnellement à a, est une loi multinomiale

$$\mathbb{P}(X_{i,j} = x_{i,j},\ 1 \leq i,j \leq N/B) = \binom{a}{x_{1,1},\ldots,x_{N,N}}\left(\frac{1}{N^2}\right)^a, \tag{2.48}$$

où $x_{i,j} \geq 0$ et $x_{N,N} = a - \sum_{j=1}^{N-1}\sum_{i=1} N x_{i,j} \geq 0$. Posons $W_{1,1},\ldots,W_{N,N}$ une suite variables aléatoires distribuées selon la loi multinomiale définie en (2.48). Pour $t \in \{1,2\}$, nous avons

$$q_{m+t-1,N}(a) = \mathbb{P}\left(\bigcap_{s=1}^{t} E_s \middle/ B\right) = \mathbb{P}\left[\bigcap_{s=1}^{t}\bigcap_{i_1=1}^{m+t-1}\sum_{j=i_2}^{i_2+m-1} V_{i,j} \leq k\right].$$

Aussi, il suffit de réaliser R simulations des variables $W_{1,1},\ldots,W_{m+1,N}$ dont la distribution jointe est

$$\mathbb{P}(W_{i,j} = w_{i,j},\ 1 \leq i \leq m+1,\ 1 \leq j \leq N)$$

$$= \binom{a}{w_{1,1},\ldots,w_{m+1,N},w^*}\left(\frac{1}{N^2}\right)^{\sum_{i=1}^{m+2}\sum_{j=1}^{N} w_{i,j}}\left(1 - \frac{(m+2)N}{N^2}\right)^{a-w^*},$$

où $w^* = a - \sum_{i=1}^{m+2}\sum_{j=1}^{N} w_{i,j}$, pour estimer les valeurs de $q_{m+t-1,N}(a)$, $1 \leq t \leq 2$, définies en (2.47).

3.2.2 Statistique de scan bidimensionnelle continue

Modèle de Poisson. Soit une région rectangulaire $[0,L]\times[0,K]$, $\{L,K\} \in \mathbb{R}^2$ et $K,L < \infty$. Soit $N = \{N_{(t,s)}, t \in [0,L], s \in [0,K]\}$ un processus de Poisson homogène bidimensionnel d'intensité λ.

Supposons que a évènements aient été observés sur $[0, L] \times [0, K]$. Soient $\{u, v\} \in \mathbb{R}^{2*}$, $u < L$ et $v < K$. En considérant $S(u, v, K, L, a)$ définie en (1.8), nous cherchons à déterminer

$$\mathbb{P}(S(u, v, L, K, a) \leq k),\ k \geq 0,$$

ou encore

$$\mathbb{P}(S(u, v, L, K, \lambda, a) \leq k),\ k \geq 0.$$

Conditionnellement au nombre total d'évènements observés, a, nous pouvons reformuler le problème de la manière suivante : soient $X_{11}, \ldots, X_{1,a}$ et $X_{21}, \ldots, X_{2,a}$ des variables aléatoires i.i.d. selon une loi uniforme sur $[0, L]$ et $[0, K]$ respectivement. Pour $1 \leq i \leq a$, posons $\mathbf{X}_i = (X_{1i}, X_{2i})$. Aussi, $\mathbf{X}_1, \ldots, \mathbf{X}_a$ peut être vue comme a points distribués de manière aléatoire sur $[0, L] \times [0, K]$.

Une borne supérieure de $\mathbb{P}(S(u, v, L, K, a) > k) = 1 - \mathbb{P}(S(u, v, L, K, a) \leq k)$ découle immédiatement du fait s'il existe un rectangle de dimension $u \times v$ contenant au moins $k + 1$ évènements alors il y a au moins $k + 1$ évènements parmi $X_{11}, \ldots, X_{1,a}$ contenus dans un intervalle de taille u et $k + 1$ évènements parmi $X_{21}, \ldots, X_{2,a}$ contenus dans un intervalle de taille v. Etant donné que ces variables aléatoires sont i.i.d. selon les lois uniformes précitées, nous avons

$$\mathbb{P}(S(u, v, L, K, a) > k) \leq \mathbb{P}(S(u, L, a) > k)\mathbb{P}(S(v, K, a) > k), \tag{2.49}$$

$S(u, L, a)$ et $S(v, K, a)$ étant définies en (1.4). Par ailleurs, $k + 1$ évènements ou plus sont contenus dans un rectangle de taille $u \times v$ si au moins $k + 1$ des premières coordonnées de $\mathbf{X}_1, \ldots, \mathbf{X}_a$ sont contenues dans un intervalle de longueur u et si au moins $k + 1$ des secondes coordonnées de $\mathbf{X}_1, \ldots, \mathbf{X}_a$, correspondant à ces premières coordonnées, sont comprises dans un intervalle de longueur v. Aussi, cela implique que

$$\mathbb{P}(S(u, L, a) > k)\mathbb{P}(S(v, k, a) > k) \leq \mathbb{P}(S(u, v, L, K, a) > k). \tag{2.50}$$

Comme l'inégalité (2.50) est symétrique en u et v, nous obtenons

$$\begin{aligned} \max\{\mathbb{P}(S(u, L, a) > k)\mathbb{P}(S(v, k, a) > k), \mathbb{P}(S(u, k, a) > k)\mathbb{P}(S(v, K, a) > k)\} \\ \leq \mathbb{P}(S(u, v, L, K, a) > k). \end{aligned} \tag{2.51}$$

Dans [Naus, 1966, Théorème 3], les auteurs ont montré que la borne supérieure (2.49) converge vers la borne inférieure (2.51) lorsque $k > a/2$. Aussi, l'auteur conseille d'utiliser la borne inférieure comme approximation de $\mathbb{P}(S(u, v, L, K, a) > k)$.

TABLE 2.6 – Récapitulatif des différentes formules exactes et approximations de la distribution des différentes statistiques de scan. Il est important de préciser que les formules exactes citées dans ce tableau correspondent aux quantités Q_2 et Q_3 pour le cas unidimensionnel, qui sont utilisées par la suite dans les approximations. Par ailleurs, dans le cas bidimensionnel continu conditionnel, l'approximation proposée dans [Naus, 1966] est constituée du produit de deux distributions de la statistique de scan unidimensionnelle continue conditionnelle.

	Cas Non Conditionnel			
	Unidimensionnel		Bidimensionnel	
	Discret	Continu	Discret	Continu
Approximations	Naus (2.2) / Haiman (2.6)	Naus (2.2) / Haiman (2.6)	Haiman (2.16) / Glaz (2.12)	Alm (2.25) / Haiman (2.27)
Bernoulli	Formules exactes (2.3)	-	Simulations	-
Binomial	Simulations	-	Simulations	-
Poisson	Simulations	Formules exactes (2.4)	Simulations	Simulations

	Cas Conditionnel			
	Unidimensionnel		Bidimensionnel	
	Discret	Continu	Discret	Continu
Approximations	Glaz (2.35)	Naus (2.41) / Glaz (2.10)	Glaz (2.45)	Naus (2.51)
Bernoulli	Formules exactes (2.34)	-	Simulations	-
Binomial	Simulations	-	Simulations	-
Poisson	Simulations	Form. Exactes : (2.40) ou (2.43)	Simulations	Approx uni (2.41)

4 Influence de la forme de la fenêtre sur l'approximation de la statistique de scan bidimensionnelle discrète

4.1 Introduction

Dans le cadre des statistiques de scan bidimensionnelles, la forme de la fenêtre de scan a fait l'objet de recherches décrites en section 3.3. En résumé, il est possible de distinguer deux types de fenêtres : les formes paramétriques (rectangle, cercle, ellipse, ...) et les formes non-paramétriques permettant de détecter des clusters de formes très irrégulières. Dans une grande majorité des cas, ces deux types de formes de fenêtre ont été évaluées par le biais d'étude de puissance afin de tester leur capacité de détection de clusters d'évènements sous l'hypothèse alternative (étude de puissance).

Par ailleurs, les différentes techniques d'approximation de la distribution des statistiques de scan bidimensionnelles, décrites au sein des sections 2.2 et 3.2, considèrent des fenêtres de scan de forme uniquement rectangulaire. Cela amène à la question suivante : la forme de la fenêtre a-t-elle une influence sur l'approximation de la distribution de la statistique de scan ? Dans le cadre des statistiques de scan bidimensionnelles continues, les auteurs ont considéré différentes formes de fenêtre. En effet, [Naus, 1965a, Loader, 1991] ont utilisé des rectangles, [Alm, 1997, Alm, 1998, Anderson and Titterington, 1997] ont considéré des rectangles et des cercles alors que [Alm, 1997, Alm, 1998] ont pris en compte des triangles, ellipses et autres formes convexes. Dans le cadre des statistiques de scan bidimensionnelles discrètes, l'influence de la forme de la fenêtre n'a, à notre connaissance, jamais été étudiée.

Nous proposons d'étudier, par le biais d'une étude de simulation, l'influence de la forme de la fenêtre sur l'approximation de la statistique de scan bidimensionnelle discrète. Pour ce faire nous avons utilisé des formes géométriques discrètes (carré, rectangle et cercle discret) ainsi que la technique approximation proposée par [Haiman and Preda, 2006] et décrite en section 2.2.1, qui présente l'avantage de fournir une erreur d'approximation. Aussi, dans l'objectif de comparer les approximations de la distribution de la statistique de scan bidimensionnelle discrète en fonction de la forme de la fenêtre de scan, seule cette technique peut permettre de mettre en évidence une différence qui ne serait pas liée à une erreur d'approximation. En d'autres termes, il est possible de comparer, pour deux formes de fenêtre distinctes de même surface, l'approximation de la distribution de la statistique de scan associée, à une certaine décimale si leur erreurs d'approximation associées sont inférieures à cette décimale. Par ailleurs, nous proposons une étude de puissance du test basé sur la statistique de scan lorsque le cluster existant sous l'hypothèse alternative est de forme carrée, rectangulaire ou circulaire.

La Section 4.2 présente les différentes formes de fenêtres utilisées. La Section 4.3 explicite la méthode de comparaison des approximations de la distribution de la statistique de scan en fonction de la forme de la fenêtre de scan ainsi que l'estimation de la puissance du test basé sur la statistique de scan, par le biais d'une étude de simulations.

4.2 Forme de la fenêtre de scan

La statistique de scan bidimensionnelle discrète est initialement définie avec une fenêtre de scan de forme rectangulaire. Cependant, le processus de scan peut être réalisé avec n'importe quelle forme convexe. Dans cette section, nous nous sommes intéressés à l'observation des différences au sein de la distribution de la statistique de scan lorsque que le processus de scan est réalisé avec des fenêtres de même surface, mais de formes différentes. En particulier, nous détaillons les cas des fenêtres rectangulaires et circulaires (cercle discret).

4.2.1 Fenêtre de scan rectangulaire.

Considérons le cas rectangulaire. Pour une surface $A \in \mathbb{N}$, nous avons utilisé l'ensemble des formes possibles de rectangle en déterminant toutes les valeurs $m_1, m_2 \in \mathbb{N}^*$ satisfaisant $m_1 \times m_2 = A$. Par exemple, pour $A = 16$, toutes les configurations possibles de rectangles sont données au sein de la Figure 2.6 pour $(m_1, m_2) \in \{(1, 16), (2, 8), (4, 4), (8, 2), (16, 1)\}$.

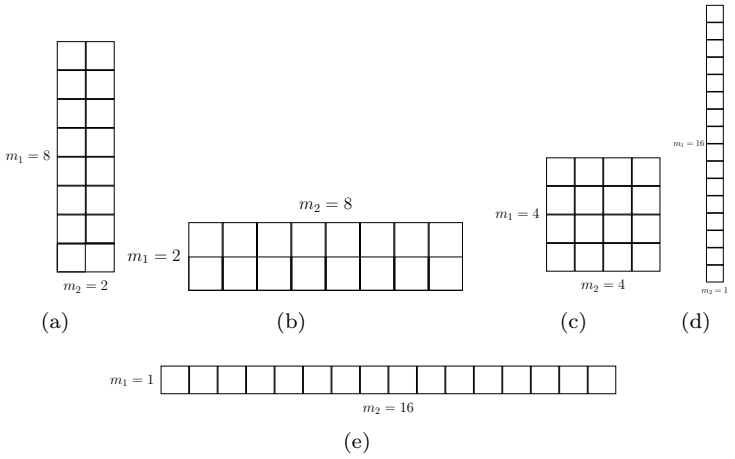


FIGURE 2.6 – Configurations possibles de formes rectangulaires pour $A = 16$.

4.2.2 Fenêtre de scan circulaire (cercle discret).

Les cercles discrets peuvent être vus comme la discrétisation dans $\mathbb{Z}^2$ des cercles euclidiens définis dans $\mathbb{R}^2$. Nous avons utilisé l'algorithme des "2 points" pour le tracé des cercles discrets, proposé par J. Brensenham [Bresenham, 1977]. Soit $\mathcal{B}(\mathbf{0}, R)$ le cercle discret de Bresenham de centre $\mathbf{O} = (0, 0)$, de rayon $R \in \mathbb{N}$. L'idée clé de cet algorithme repose sur le principe de symétrie du cercle. En effet, on constate qu'il existe quatre axes de symétrie au sein d'un cercle, deux suivant les axes du repère centré sur le cercle et deux bissectrices. Aussi, le tracé d'un cercle discret est uniquement réalisé sur le deuxième octant de $\mathbb{Z}^2$, c'est à dire, l'ensemble $\{(i, j) \in \mathbb{Z}^2 / 0 \leqslant i \leqslant j\}$ (Figure 2.7(a)). Les autres pixels sont tracés en utilisant leur images symétriques dans les sept autres octants de $\mathbb{Z}^2$. De surcroît, l'algorithme est dit incrémental car la position du prochain pixel tracé est déterminée par la position du pixel précédent et non en utilisant une formule générale pour l'ensemble des pixels.

Soit $\mathcal{C}(\mathbf{O}, R)$ le cercle euclidien de centre $\mathbf{O} = (0, 0)$ et de rayon R, décrit par l'équation suivante

$$f(i, j) = i^2 + j^2 - R^2 = 0.$$

L'algorithme de Bresenham est incrémental en i et début au pixel trivial $(0, R)$ d'une grille $\mathcal{G}$ définie dans $\mathbb{Z}^2$. Au sein du deuxième octant de $\mathbb{Z}^2$, j est une fonction décroisante de i. Aussi, pour un pixel tracé (i, j), J. Bresenham a montré qu'il n'existe que deux candidats possibles pour devenir le prochain pixel tracé : soit le pixel de coordonnées $(i + 1, j)$ soit le pixel de coordonnées $(i + 1, j - 1)$ (Figure 2.7(a)). Parmi ces deux candidats, nous devons choisir celui qui est le plus proche de $\mathcal{C}(\mathbf{0}, R)$ et cela est réalisé en utilisant la différence suivante

$$\delta(i, j) = |f(i + 1, j)| - |f(i + 1, j + 1)|.$$

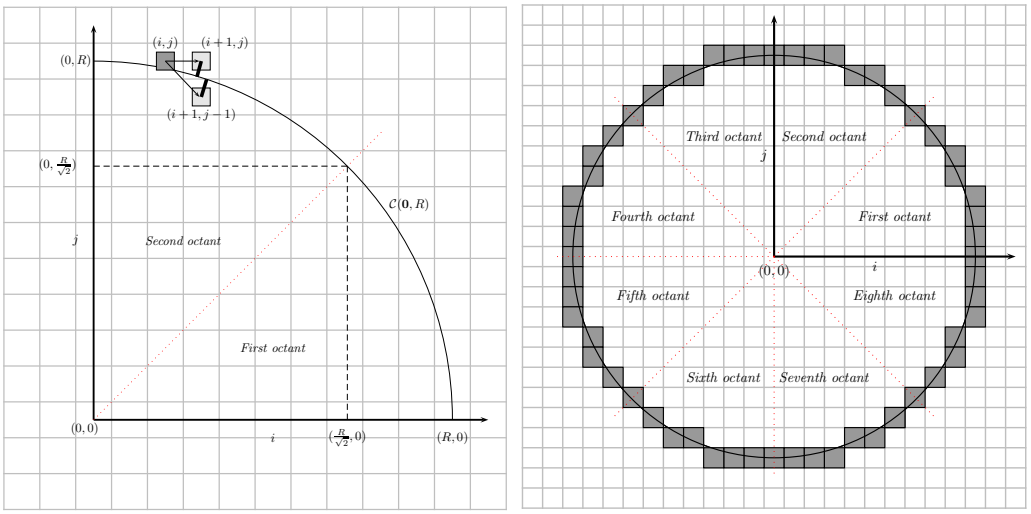
Si $\delta(i, j) \leqslant 0$ alors le pixel de coordonnées $(i + 1, j)$ est le plus proche de $\mathcal{C}(\mathbf{0}, R)$ et il est donc sélectionné pour être le prochain pixel tracé. Sinon, le pixel de coordonnées $(i + 1, j - 1)$ est le plus proche et il est tracé. Par ailleurs, pour un pixel tracé de coordonnées (i, j) dans le second octant de $\mathbb{Z}^2$, les pixels de coordonnées (j, i), $(-i, j)$, $(-j, i)$, $(-j, -i)$, $(-i, -j)$, $(i, -j)$, $(j, -i)$ sont également tracés dans les sept autres octants.

Afin de comparer une fenêtre de scan de forme rectangulaire de surface $A \in \mathbb{N}$ avec une fenêtre de type cercle discret, nous devons déterminer le rayon optimal R qui minimise $|A - D|$, D étant le nombre de pixels dans le disque discret de centre $\mathbf{0}$ et de rayon R. Dans [Kulpa, 1979], les auteurs ont montré que D peut être déterminé par l'expression suivante :

$$D = 8 \sum_{j=1}^{\lfloor \alpha \rfloor} \lfloor \frac{1}{2} + \sqrt{R^2 - j^2} \rfloor - 4 \lfloor \alpha \rfloor^2 + 4R + 1, \tag{2.52}$$

Algorithme 2 Tracé de $\mathcal{B}(\mathbf{0}, R)$ dans $\mathbb{Z}^2$ [Bresenham, 1977]

$i \leftarrow 0$
$j \leftarrow R$
Tantque $i \leqslant j$ **Faire**
  trace_pixel(j, i) // *Premier octant*
  trace_pixel(i, j) // *Second octant*
  trace_pixel$(-i, j)$ // *Troisième octant*
  trace_pixel$(-j, i)$ // *Quatrième octant*
  trace_pixel$(-j, -i)$ // *Cinquième octant*
  trace_pixel$(-i, -j)$ // *Sixième octant*
  trace_pixel$(i, -j)$ // *Septième octant*
  trace_pixel$(j, -i)$ // *Huitième octant*
  Si $\delta(i, j) > 0$ **Alors**
    $j \leftarrow j - 1$
  finSi
  $i \leftarrow i + 1$
fin Tantque



(a) Processus de tracé de cercle discret

(b) Cercle discret de rayon $R = 10$

FIGURE 2.7 – Cercle discret dans $\mathbb{Z}^2$.

où $\lfloor \alpha \rfloor = \lfloor \frac{1}{4}(1 + \sqrt{(8R^2 - 1)}) \rfloor$. Des exemples de comparaisons de surfaces entre des fenêtres de forme carrée et circulaire sont données dans la Table 2.7.

TABLE 2.7 – Comparaison de surface entre formes carrées et circulaires avec un rayon optimal R.

Rectangle	**Discrete Circle**		
A	R	D	$\lvert A - D \rvert$
25	2	21	4
36	3	37	1
64	4	61	3
100	5	97	3

4.3 Applications numériques : approximation et puissance du test basé la statistique de scan

Dans cette section, nous présentons, dans un premier temps, une étude de simulation ayant pour objectif d'évaluer le changement dans la distribution de la statistique de scan bidimensionnelle discrète lorsque le processus de scan utilise différentes formes de fenêtres. Dans un deuxième temps, nous évaluons la puissance du test basé sur la statistique de scan en fonction de la forme de la fenêtre de scan et celle du cluster simulé sous l'hypothèse alternative. Nous considérons ici les formes carrée, rectangulaire et circulaire (cercle discret), de même surface, pour la fenêtre de scan et les clusters simulés.

4.3.1 Approximation et forme de la fenêtre de scan

Nous présentons des approximations de la distribution de la statistique de scan, $\mathbb{P}(S \leq n)$, lorsque les X_{ij} sont distribués selon une loi binomiale $\mathcal{B}(\nu, p)$ et une loi de Poisson $\mathcal{P}(\lambda)$. La méthodologie utilisée pour obtenir ces approximations est celle proposée par [Haiman and Preda, 2006] et décrite en Section 2.2.1. Pour chaque modèle, nous comparons les distributions des statistiques de scan lorsque la fenêtre de scan est carrée, rectangulaire et circulaire. La Table 2.8 présente la distribution de la statistique de scan bidimensionnelle discrète lorsque le processus de scan est réalisé sur une une région de taille $N_1 \times N_2 = 42 \times 42$ avec une fenêtre de scan de taille fixe ($A = 36$) et de formes variées : carrée (6×6), rectangulaire (3×12) et circulaire ($R = 3$) [1]. Comme $N_1 = N_2$, la distribution des statistiques de scan avec une fenêtre de scan $m_1 \times m_2$ est identique à la distribution avec une fenêtre $m_2 \times m_1$. Nous considérons différents modèles pour les X_{ij} : $\mathcal{B}(1, 0.01)$, $\mathcal{B}(5, 0.05)$ et $\mathcal{P}(0.25)$. Concernant les fenêtres de scan rectangulaires, nous avons utilisé la technique d'*Importance Sampling* présenté dans [Amarioarei and Preda, 2013] qui fournit des intervalles de confiance des erreurs beaucoup plus restreints que la méthode de Monte Carlo que nous avons présentée pour le cercle discret.

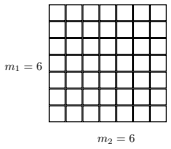
En comparant les valeurs d'approximation et en prenant en compte les erreurs d'approximation, on observe que les distributions des statistiques de scan associées à différentes formes de fenêtres de scan sont très proches les unes des autres. Remarquons, dans cet exemple, que le quantile d'ordre 0.95 de la statistique de scan ne change pas avec la forme de la fenêtre. Cependant, comme nous l'attendions, on peut remarquer des faibles différences significatives comme par exemple, pour le modèle de Poisson, lorsque que l'on scanne avec une fenêtre 6×6 ($\mathbb{P}(S(6, 6, 42, 42) \leq 22) < 0.96$) et avec une fenêtre 3×12 ($(\mathbb{P}(S(3, 12, 42, 42) \leq 22) > 0.96)$).

A partir de cette étude de simulation, mais sans généraliser, nous pouvons conclure que la forme de la fenêtre de scan ne change pas considérablement la distribution de la statistique de scan bidimensionnelle discrète.

1. Dans le cas d'une taille de fenêtre $A = 16$, des résultats numériques supplémentaires sont disponibles en Annexe C.

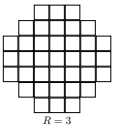
TABLE 2.8 – Approximations de $\mathbb{P}(S \leqslant n)$: Scan d'une région $N_1 \times N_2 = 42 \times 42$ avec différentes formes de fenêtres.

Scanning window shape — **Approximation and approximation error for $\mathbb{P}(S \leqslant n)$**



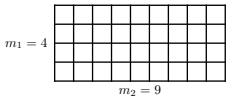
Square
$m_1 = m_2 = 6,\ K = L = 7,\ N = 10^6$

$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$		
k	App. (2.16)	Err. (2.23)	k	App. (2.16)	Err. (2.23)	k	App. (2.16)	Err. (2.23)
3	0.869707	0.042916	21	0.935139	0.007260	21	0.900014	0.017640
4	0.987184	0.000761	22	0.974401	0.001505	22	0.956164	0.003483
5	0.999136	0.000035	23	0.990531	0.000386	23	0.982368	0.000862
6	0.999955	0.000002	24	0.996714	0.000109	24	0.993254	0.000248
7	0.999998	$5.4E^{-08}$	25	0.998922	0.000031	25	0.997553	0.000076



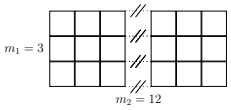
Discrete Circle
$R = 3,\ K = L = 6,\ N = 10^6$

$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$		
n	App. (2.16)	Err. (2.23)	n	App. (2.16)	Err. (2.23)	n	App. (2.16)	Err. (2.23)
3	0.861413	0.015934	21	0.908622	0.041011	21	0.868727	0.081296
4	0.986067	0.002879	22	0.962759	0.008972	22	0.940899	0.018439
5	0.998741	0.000610	23	0.985878	0.002814	23	0.974819	0.005116
6	0.999951	0.000120	24	0.994401	0.001400	24	0.990174	0.002073
7	0.999984	0.000049	25	0.998256	0.000764	25	0.996016	0.001184



Rectangle
$m_1 = 4,\ m_2 = 9,\ K = \{4, 5\},\ L \in \{10, 11\},\ N = 10^6$

$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$		
n	App. (2.16)	Err. (2.23)	n	App. (2.16)	Err. (2.23)	n	App. (2.16)	Err. (2.23)
3	0.877179	0.083146	21	0.939068	0.010084	21	0.905948	0.030105
4	0.988156	0.000700	22	0.975922	0.001605	22	0.959257	0.004212
5	0.999200	0.000029	23	0.991213	0.000356	23	0.983609	0.000863
6	0.999959	0.000001	24	0.996940	0.000095	24	0.993726	0.000224
7	0.999998	$4.6E^{-08}$	25	0.998999	0.000027	25	0.997715	0.000066



Rectangle
$m_1 = 3,\ m_2 = 12,\ K = \{3, 4\},\ L = 14,\ N = 10^6$

$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$		
n	App. (2.16)	Err. (2.23)	n	App. (2.16)	Err. (2.23)	n	App. (2.16)	Err. (2.23)
3	0.889072	0,028246	21	0.946773	0.007472	21	0.916710	0.037748
4	0.989557	0,000009	22	0.979108	0.001088	22	0.964482	0.003788
5	0.999321	$3.1E^{-07}$	23	0.992349	0.000208	23	0.985723	0.000587
6	0.999965	7.9^{E-10}	24	0.997357	0.000050	24	0.994554	0.000129
7	0.999999	$1.3E^{-12}$	25	0.999136	0.000013	25	0.998028	0.000034

4.3.2 Puissance du test basé sur la statistique de scan : formes de la fenêtre et du cluster

Nous évaluons la puissance du test basé sur la statistique de scan au regard de la forme de la fenêtre de scan et celle du cluster simulé sous l'hypothèse alternative. Nous considérons ici le modèle binomial et modèle de Poisson pour la distribution des X_{ij}.

Pour le modèle binomial, le test basé sur la statistique de scan vérifie l'hypothèse nulle $\mathcal{H}_0$: $X_{ij} \sim \mathcal{B}(\nu, p_0)$ contre un hypothèse alternative $\mathcal{H}_1$ supportant l'existence d'une sous-région (cluster) $\mathcal{C} \in \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$ telle que pour tout $(i, j) \in \mathcal{C}$, les X_{ij} sont distribuées selon une loi binomiale $\mathcal{B}(\nu, p_1)$, $p_1 > p_0$. A l'extérieur de $\mathcal{C}$, les X_{ij} sont distribuées selon une loi binomiale $\mathcal{B}(\nu, p_0)$. Pour le modèle de Poisson, $\mathcal{H}_0 : X_{ij} \sim \mathcal{P}(\lambda_0)$ et $\mathcal{H}_1$ supporte l'existence d'un cluster $\mathcal{C}$ tel que pour tout $\{i, j\} \in \mathcal{C}$, les X_{ij} sont distribuées selon une loi de Poisson $\mathcal{P}(\lambda_1)$, $\lambda_1 > \lambda_0$ et, à l'extérieur, $X_{ij} \sim \mathcal{P}(\lambda_0)$.

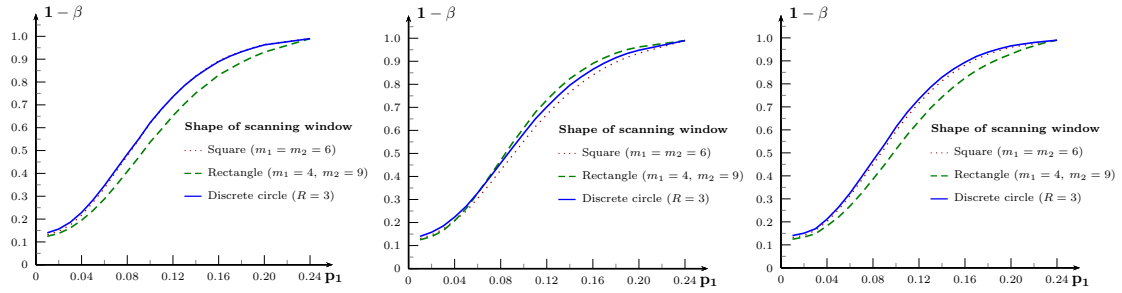
Dans ce qui suit, la forme du cluster $\mathcal{C}$ sera carrée, rectangulaire et circulaire. Pour une forme fixée du cluster, nous évaluons la puissance de la statistique de scan pour détecter le cluster lorsque la fenêtre de scan est de forme variée, *i.e.* carrée, rectangulaire et circulaire. Nous avons adopté la procédure suivante :

- Fixer le risque de première espèce à $\alpha = 0.05$;
- Pour les modèles binomial et Poisson, générer un champ aléatoire $\{X_{ij}\}$, $\{i, j\} \in \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$ tel qu'il y ait un changement dans les paramètres de distribution des X_{ij} dans le cluster $\mathcal{C}$. L'emplacement de $\mathcal{C}$ au sein de $\{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$ est généré de manière aléatoire ;
- Sous $\mathcal{H}_1$, pour le modèle binomial, nous considérons p_1 dans l'intervalle $[p_0, 1]$ prenant toutes les valeurs avec un pas de 0.02 De manière similaire, pour le modèle de Poisson, $\lambda_1 \in [\lambda_0, \lambda_{max}]$ prenant toutes les valeurs avec un pas de 0.05. La valeur λ_{max} est telle que $\lambda_{max} \geq \lambda_0$ et que la puissance de la statistique de scan ne change pas (*i.e.* égale à 1) pour toutes les formes de fenêtre de scan considérées ;
- Pour chaque forme particulière de la fenêtre de scan (carrée, rectangulaire et circulaire),
  - Posons $n_{1-\alpha}$ le plus petit entier tel que $\mathbb{P}(S \leqslant n_{1-\alpha})) \geqslant 1 - \alpha$. Il est obtenu comme quantile d'ordre $1 - \alpha$ de la distribution approximée de S donnée par (2.16).
  - Sous $\mathcal{H}_1$, pour chaque forme spécifique de cluster (carrée, rectangulaire et circulaire) et chaque paramètre (p_1 et λ_1), simuler $N = 10^6$ réalisation du champ aléatoire $\{X_{i,j}\}$. Pour chaque réalisation i, $1 \leqslant i \leqslant N$, la valeur observée de la statistique de scan S_i est comparée à $n_{1-\alpha}$. La puissance $1 - \beta$ est estimée par :

$$1 - \beta = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S_i \geqslant n_{1-\alpha}\}}.$$

Pour $N_1 \times N_2 = 42 \times 42$, l'estimation de la courbe de puissance du test basé sur les statistiques de scan est présentée au sein des Figures 2.8, 2.9 and 2.10.

Au sein de la Figure 2.8 nous présentons le modèle de Bernoulli et, pour chaque forme de cluster (carré (a), rectangulaire (b) et circulaire (c)), nous avons tracé la courbe de puissance estimée correspondant aux statistiques de scan avec différentes formes de fenêtre de scan. Nous pouvons observer que pour un cluster carré, la détection est plus efficace lorsque que l'on scanne avec une fenêtre de forme carrée et circulaire que lorsque qu'une fenêtre de scan rectangulaire est utilisée (Figure 2.8(a)). Pour un cluster de forme rectangulaire, la fenêtre de scan rectangulaire est la plus efficace (Figure 2.8(b)) parmi toutes les formes. Pour un cluster de forme circulaire, une fenêtre de scan de forme cercle discret est la plus efficace (Figure 2.8(c)). Les mêmes commentaires sont valables pour les Figures 2.9 et 2.10.

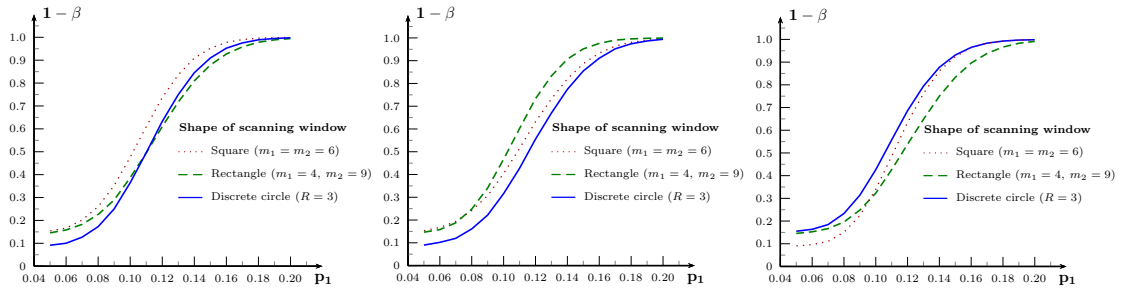


(a) Square simulated cluster ($m_1 = m_2 = 6$)

(b) Rectangular simulated cluster ($m_1 = 4,\ m_2 = 9$)

(c) Discrete circle shaped simulated cluster ($R = 3$)

FIGURE 2.8 – Modèle de Bernoulli $\mathcal{B}(1, 0.01)$: estimation de la puissance du test basé sur la statistique de scan.

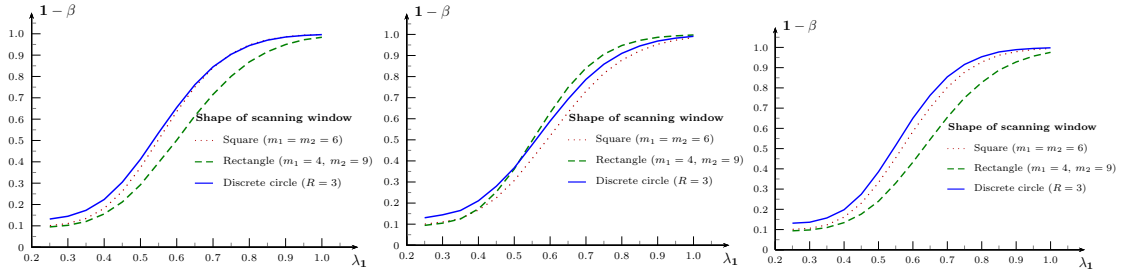


(a) Square simulated cluster ($m_1 = m_2 = 6$)

(b) Rectangular simulated cluster ($m_1 = 4,\ m_2 = 9$)

(c) Discrete circle shaped simulated cluster ($R = 3$)

FIGURE 2.9 – Modèle binomial $\mathcal{B}(5, 0.05)$: estimation de la puissance du test basé sur la statistique de scan.



(a) Square simulated cluster ($m_1 = m_2 = 6$)

(b) Rectangular simulated cluster ($m_1 = 4,\ m_2 = 9$)

(c) Discrete circle shaped simulated cluster ($R = 3$)

FIGURE 2.10 – Modèle de Poisson $\mathcal{P}(0.25)$: estimation de la puissance du test basé sur la statistique de scan.

5 Conclusion

Dans un premier temps, ce chapitre s'est attaché à décrire les différentes formules exactes et approximations de la distribution des statistiques de scan dans les cas non-conditionnel (Section 2) et conditionnel (Section 3), faisant, pour chaque cas, la distinction entre les cas unidimensionnel et bidimensionnel, discret et continu.

Dans un deuxième temps, nous avons évalué l'influence de la forme de la fenêtre de scan sur la distribution des statistiques de scan bidimensionnelles discrètes. Nous avons réalisé une étude de simulation prenant en compte des fenêtres de scan de forme carrée, rectangulaire et circulaire (cercle discret). Cette étude a mis en évidence le fait que les distributions des statistiques de scan associées à ces formes sont très proches les unes des autres mais significativement différentes. Par ailleurs, nous avons mis en évidence, par le biais d'une étude de simulation, que la puissance du test basé sur les statistiques de scan est liée à la forme de la fenêtre ainsi qu'à la forme du cluster existant sous l'hypothèse alternative.

Chapitre 3

Statistiques de scan spatiales

Sommaire

1 Introduction 63
2 Données spatiales 64
2.1 Modèle général 64
2.2 Données géostatistiques 64
2.3 Données latticielles 65
2.4 Données ponctuelles 66
3 Cadre général de la statistique de scan spatiale 69
3.1 Phase de détection 70
3.2 Phase d'inférence statistique 72
3.3 Notion de clusters secondaires 73
4 Principaux modèles 74
4.1 Modèle de Bernoulli 74
4.2 Modèle de Poisson 77
5 Limites des statistiques de scan spatiales 81
5.1 Temps de calcul 81
5.2 Forme de la fenêtre 86
6 Une alternative à la méthode de Monte Carlo pour le TRVG 86
6.1 Méthodologie 86
6.2 Etude de simulation 88
6.3 Résultats 90
7 Conclusion 92

1 Introduction

Les statistiques de scan spatiales constituent une extension des statistiques de scan bidimensionnelles, appliquées aux données spatiales. Ces méthodes, initialement proposées dans [Kulldorff and Nagarwalla, 1995], étendues dans [Kulldorff, 1997] puis dans [Kulldorff, 2006], présentent la particularité d'utiliser une fenêtre de scan de forme circulaire de taille variable et d'utiliser des simulations de Monte Carlo pour estimer la distribution de la statistique de test. Elles sont de loin les plus utilisées par la communauté scientifique [1] notamment par le fait qu'elles soient implémentées au sein du logiciel SaTScan© [Kulldorff, 2011] qui dispose d'une interface graphique et, par conséquent, propose une utilisation relativement aisée.

Dans un premier temps, ce chapitre va s'attacher à définir les différents types de données spatiales que sont les données géostatistiques, les données latticielles et les processus ponctuels spatiaux. Dans un second temps, le cadre général des statistiques de scan sera explicité, notamment

1. Elles comptabilisent, en 2013, plus de 2370 références sur Google Scholar.

les deux phases principales de ces méthodes que sont la phase de détection d'un cluster potentiel et la phase d'inférence statistique. Dans un troisième temps, nous décrivons les modèles de Bernoulli et de Poisson, en proposant une définition plus rigoureuse, d'un point de vue mathématique, que celle énoncée dans [Kulldorff, 1997]. Dans un quatrième temps, nous mettons en évidence les deux faiblesses majeures de la méthode que sont la forme de la fenêtre et son temps de calcul. Cette dernière fait l'objet d'une étude plus approfondie car nous proposons, dans un cinquième temps, une alternative à l'approximation de la distribution de la statistique de test par la méthode de Monte Carlo. En effet, nous proposons une modélisation basée sur les suites de variables aléatoires 1-dépendantes afin de donner une approximation de la probabilité critique associée au cluster détecté. Cette modélisation s'avère être plus précise tout en réduisant de manière drastique les temps de calculs inhérents aux méthodes de statistiques de scan spatiales.

2 Données spatiales

En statistique spatiale, on associe aux observations leurs coordonnées spatiales, et ce sont celles-ci qui vont intervenir dans la modélisation statistique. Il existe trois grandes familles de données spatiales : les données géostatistiques, les données latticielles et les données ponctuelles. Les modèles statistiques spatiaux varient en fonction de la nature des données spatiales. Cependant, il est important de déterminer un modèle général, flexible, qui sert de *framework* aux différents modèles liés à la nature des données.

2.1 Modèle général

Définition 3.1. Soit $X = \{X_d,\, d \in D\}$ un processus aléatoire indexé par un ensemble spatial D et à valeurs dans un espace d'états E.

La localisation d'un site $d \in D$ peut être soit fixée ou soit aléatoire en fonction du modèle de données. L'ensemble spatial D peut être bi-dimensionnel ($D \subseteq \mathbb{R}^2$), tri-dimensionnel ($D \subseteq \mathbb{R}^3$), spatio-temporel (indexation d'une observation par $(d,t) \in \mathbb{R}^2 \times \mathbb{R}^+$). Par ailleurs, l'espace d'états E peut être de nature :

- $E \subseteq \mathbb{R}^p$ (champ gaussien)
- $E \subseteq \mathbb{R}^+$ (champ exponentiel ou Gamma)
- $E \subseteq \mathbb{N}$ (champ poissonien)
- $E = \{a_0, a_1, \ldots, a_k\}$ (champ catégoriel)
- $E = \{0, 1\}$ (Champ binaire)

En pratique, on observe n données spatiales $\{x_{d_1}, x_{d_2}, \ldots, x_{d_n}\}$, qui constituent une réalisation de X et dont les coordonnées spatiales sont $\{d_1, d_2, \ldots, d_n\}$, appelées également index spatiaux. La nature des ensembles E et D détermine le type de données spatiales. On distingue trois grands types de données que sont les **données géostatistiques**, les **données latticielles** et les **données ponctuelles**. Les parties suivantes décrivent, en se basant sur ce modèle général, ces trois types de données.

2.2 Données géostatistiques

Définition 3.2. Lorsque qu'un processus spatial X à valeurs dans $E \subseteq \mathbb{R}^p$, $(p \geq 1)$ est observé au travers de n sites fixés $\{d_1, d_2, \ldots, ; d_n\} \in D$, D étant un sous-espace *continu* fixé de $\mathbb{R}^q$, $(D \subseteq \mathbb{R}^q,\, q \geq 2)$, on parle de **données géostatistiques**.

En d'autres termes, si les coordonnées spatiales de n sites fixés $\{d_1, d_2, \ldots, d_n\}$ sont comprises dans un espace continu de $\mathbb{R}^q$ et les variables aléatoires X_d associées sont à valeurs dans $\mathbb{R}^p$, nous sommes en présence de données géostatistiques. De surcroît, ces données rentrent dans la catégorie des champs aléatoires de second ordre ($E[X_d^2] < \infty$). Par ailleurs, la localisation des sites peut être régulière $D \subseteq \mathbb{Z}^q$ ou irrégulière ($D \subseteq \mathbb{R}^q$) avec $q \geq 2$ (Figure 3.1).

Par exemple, considérons les relevés de pluviométrie dans une région. Ces relevés peuvent être, en principe, réalisés à n'importe quel endroit de la région mais sont fixés par le météorologue.

A chaque site d de relevé on associe une mesure de la pluviométrie qui est une réalisation d'une variable aléatoire $X(d)$ à valeurs dans $\mathbb{R}^+$.

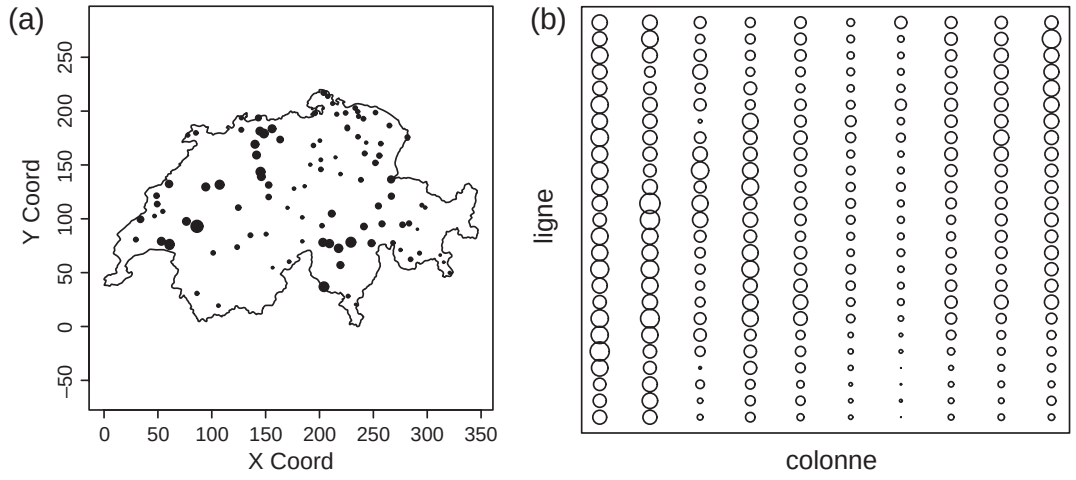


FIGURE 3.1 – (a) Cumuls pluviométriques sur le réseau météorologique suisse le 8 mai 1986 (passage du nuage de Chernobyl, données sic du package geoR du logiciel R, (b) Porosité d'un sol (données soil du package geoR)

2.3 Données latticielles

Définition 3.3. Lorsque qu'un processus aléatoire $X = \{X_d, d \in D\}$ est indexé par un ensemble discret fixé D défini comme un réseau discret structuré par un graphe de voisinage $\mathcal{G}$ d'ordre n, on parle de **données latticielles**.

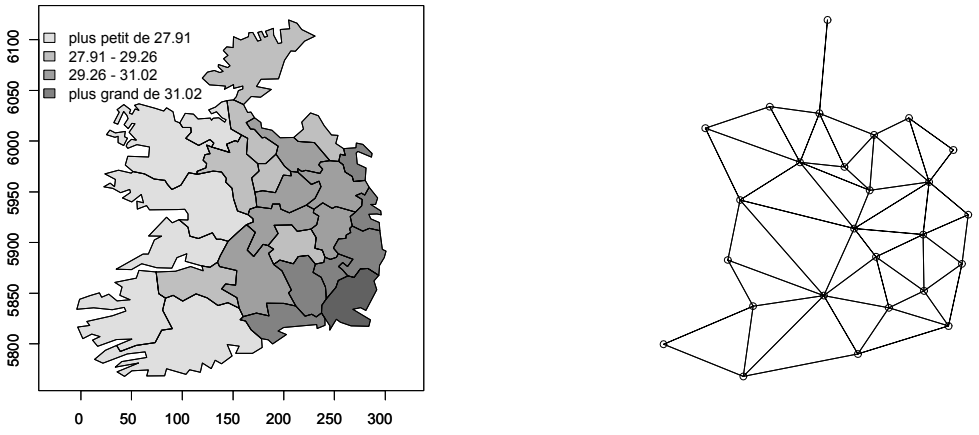
D peut être un réseau discret régulier ($D \subset \mathbb{Z}^p$) ou irrégulier $D \subset \mathbb{R}^p$ avec $p \geq 2$. On associe à D un graphe simple de voisinage $\mathcal{G}$ d'ordre n défini par $\mathcal{G} = \{D, A\}$, D étant l'ensemble des sommets du graphe et A l'ensemble des arêtes. On munit $\mathcal{G}$ d'un matrice d'adjacence $U_{n\times n}$ définie par :

$$U = (u_{ij}),\ i,j \in \{1,\dots,n\} \text{ avec } u_{ij} = \begin{cases} 1 & \text{si } (i,j) \in A \\ 0 & \text{sinon} \end{cases}$$

On munit également $\mathcal{G}$ d'une matrice W de poids positifs définie par :

$$W = (w_{ij}),\ i,j \in \{1,\dots,n\} \text{ avec } w_{ij} = \begin{cases} w_{ij} & \text{si } (i,j) \in A \text{ et } i \neq j \\ 0 & \text{sinon} \end{cases}$$

Ce type de données spatiales est très courant dans le milieu de la santé. En effet, les informations concernant des évènements de santé (cas de maladie, décès, ...) sont souvent agrégées à l'échelle d'une unité spatiale (iris, commune, canton, département, région, etc.), souvent représentée elle-même par son centre (barycentre de l'unité spatiale).
Un exemple de données latticielles (ou données agrégées) est donné par la Figure 3.2 dans laquelle est représenté le pourcentage de personnes du groupe sanguin A, donnée agrégée au comté d, en Irlande (D). Nous sommes donc en présence de données latticielles sur un réseau discret D. La variable aléatoire $X(d)$ correspond au nombre de personnes concernées dans un comté d de la région D, donc $E = \mathbb{N}$.



(a) (b)

FIGURE 3.2 – (a) Pourcentage de la population présentant le groupe sanguin A dans les 26 comtés d'Irlande, (b) Graphe de voisinage $\mathcal{G}$ des 26 comtés

2.4 Données ponctuelles

Les données ponctuelles sont modélisées par la notion de processus spatiaux ponctuels (PSP). Ceux-ci constituent un champ de la statistique qui étudie les ensembles d'évènements définis par leur coordonnées dans l'espace. La différence fondamentale entre les données géostatistiques, les données latticielles d'une part et les processus ponctuels spatiaux réside dans le fait que les coordonnées spatiales des sites $\{d_1, d_2, \ldots, d_n\}$, désormais déterminées par $\mathbf{x} = \{x_1, x_2, \ldots\}$ sont aléatoires. Par ailleurs, le nombre de sites observés $n(\mathbf{x}) = n$ est également aléatoire. Dans la littérature traitant des PSP, on peut distinguer les ouvrages de références de [Cressie, 1991], [Stoyan and Stoyan, 1994], [Moller and Waagepetersen, 2003] et [Diggle, 2003].

Considérons les notations suivantes : $\mathcal{B}_b(D)$ l'ensemble des boréliens fermés de D et μ la mesure de Lebesgue sur D.

2.4.1 Définitions

Définition 3.4. Un processus ponctuel spatial X sur $D \subseteq \mathbb{R}^p$, $p \geq 2$ est une application d'un espace de probabilité $\{\Omega, \mathcal{A}, P\}$ dans l'ensemble E des configurations localement finies telle que $\forall A \in \mathcal{B}_b(D)$, le nombre d'évènements $N_X(A)$ de X observés dans A est une variable aléatoire.

$$N_X(A) : E \to \mathbb{N}, \forall A \in \mathcal{B}_b(D)$$

Avec E la réunion de tous les espaces de configurations à n points sur D : $E = \cup_{n \geq 0} E_n$. E est appelé espace exponentiel des configurations.

Remarque 3.1. $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, $x_i \in D \subset \mathbb{R}^p$ *est la réalisation d'un processus spatial ponctuel* X *et est appelé* ***semis de points****.*

Propriétés 3.5. Si D est borné, alors $N_X(D)$ est presque sûrement fini. Le processus spatial ponctuel est dit **fini**.

Propriétés 3.6. S'il n'y a pas de répétitions de points au même d-uplet de coordonnées, le processus spatial ponctuel est dit **simple**.

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\} \subset D, \text{ si } n(\mathbf{x}) = n$$

Définition 3.7. La loi d'un processus spatial ponctuel X est définie comme la donnée pour tout $m \geq 0$ et tout m-uplet $(A_1, \ldots, A_m) \in \mathcal{B}_b(D)$ des distributions sur $\mathbb{N}^m$ de $(N(A_1), \ldots, N(A_m))$.

Définition 3.8. Un processus ponctuel spatial X est dit **marqué** si on associe une marque m appartenant à un espace métrique K à chaque point de X.

Par exemple, les marques peuvent être quantitatives (*ex. Mesure biologique,* $K = \mathbb{R}$), ou qualitatives (*ex. stades de cancers,* $K = \{m_1, m_2, \ldots, m_k\}$).

2.4.2 Processus binomial

Définition 3.9. Un **processus binomial** X sur $D \subset \mathbb{R}^p$ à n points est constitué de n points *i.i.d.* uniformes sur D si :

$$\begin{gathered} \{A_1, A_2, \ldots, A_k\} \text{ une partition borélienne de } D \\ (N(A_1), N(A_2), \ldots, N(A_k)) \sim \mathcal{M}(n; q_1, q_2, \ldots, q_k) \\ \text{avec } q_i = \frac{\mu(A_i)}{\mu(D)} \text{ et } \mu \text{ la mesure de Lebesgue sur } \mathbb{R}^p \end{gathered}$$

En d'autres termes, $\forall A \in \mathcal{B}_b(D)$, $N_X(A) \sim \mathcal{B}(n, \frac{\mu(A)}{\mu(D)})$

2.4.3 Processus de Poisson homogène

Définition 3.10. X est un **processus ponctuel de Poisson homogène** d'intensité $\lambda > 0$ sur $S \subset \mathbb{R}^p$ si :

- $N(D) \sim \mathcal{P}(\lambda\mu(D))$, $N(A) \sim \mathcal{P}(\lambda\mu(A))$, $\forall A \in \mathcal{B}_b(D)$
- Si A_i sont des boréliens disjoints, les $N(A_i)$ sont indépendantes.
- Si $N(D) = n$, les n points sont *i.i.d.* uniformes sur D

$$P(N(D) = n) = \frac{e^{-\lambda\mu(D)}(\lambda\mu(D))^n}{n!}$$

L'intensité λ correspond au nombre attendu d'évènements par unité de surface. Conditionnellement au nombre total d'évènements observés n, on peut estimer λ par $\hat{\lambda} = n/\mu(D)$.

L'hypothèse selon laquelle les données spatiales peuvent être modélisées par un processus de Poisson homogène est appelée *CSR* (*Completely Spatial Randomness*). Cela signifie qu'un évènement a la même probabilité de survenir à n'importe quel point de D, indépendamment de la localisation des autres évènements (les évènements sont indépendants les uns des autres). En d'autres termes, l'hypothèse *CSR* désigne l'absence de structure spatiale dans les données.

2.4.4 Processus de Poisson non-homogène

Le processus de Poisson homogène formule l'hypothèse que l'intensité λ est constante sur D. Cependant, cette hypothèse peut s'avérer parfois trop restrictive. En effet, dans le cadre des données de santé, la population à risque n'est pas distribuée de manière uniforme sur l'ensemble de la zone étudiée. Prenons l'exemple d'un découpage cantonal d'une aire géographique étudiée. On utilise comme données de population à risque les données d'un recensement. Aussi, certains cantons vont présenter une forte population à risque (cantons urbains) alors que d'autres vont être faiblement peuplés (cantons ruraux). On peut facilement imaginer que le nombre d'évènements observés va être plus important dans les zones fortement peuplées que dans les zones faiblement peuplées.
Dans la plupart des cas, on s'intéresse aux variations du nombre d'évènements qui ne seraient pas imputables aux variations de la population sous-jacente. Aussi, l'hypothèse d'une intensité λ constante sur toute la zone étudiée n'est pas adaptée. Pour pallier ce problème, on utilise un processus de Poisson non-homogène dans lequel l'intensité est définie comme une **mesure d'intensité** :

$$\forall A \in \mathcal{B}_b(D),\ \lambda(A) = \int_{x \in A} f(x).dx$$

Où x correspond à un point de A et $f(x)$ est une fonction d'intensité.

Définition 3.11. Soit $\lambda(\cdot)$ une mesure d'intensité sur $D \subset \mathbb{R}^p$. X est un **processus de Poisson non-homogène** d'intensité $\lambda(\cdot)$ sur D si :

— $\forall A \in \mathcal{B}_b(D)$, $N(A) \sim \mathcal{P}(\lambda(A))$
— si A_1, $A_2 \in \mathcal{B}_b(D)$ alors $N(A_1)$ et $N(A_2)$ sont indépendantes.

L'hypothèse selon laquelle les données peuvent être modélisées par un processus de Poisson hétérogène est notée *CRH* (*Constant risk hypothesis*). Cette hypothèse définit la présence d'une structure dans les données spatiales uniquement liée à la répartition de la population sous-jacente ($\lambda(\cdot)$). En d'autres termes, la réalisation d'un processus de Poisson non-homogène peut présenter des agrégats d'évènements dont la position est fixe et dépend de $\lambda(\cdot)$.

2.4.5 Autres processus spatiaux ponctuels

Poisson clusters processes. Ce type de processus est défini par un processus spatial ponctuel dans lequel chaque événement appartient à un cluster spécifique. Le concept est divisé en deux étapes :

1. On simule une réalisation d'un processus de Poisson (homogène ou non-homogène) qui constitue le processus père.
2. Chaque évènement simulé va donner naissance à un nombre aléatoire ou fixé d'évènements enfants qui vont se distribuer de manière aléatoire autour de l'événement du processus père.

Le processus de Neymann-Scott ([Neyman and Scott, 1958]) est un exemple de *Poisson clusters processes*.

Processus de contagion/inhibition. La principale caractéristique des **processus de contagion** réside dans le fait que l'apparition d'un événement à un point $x \in D$ augmente la probabilité de survenue d'autres évènements dans les environs de x. Ce type de processus s'avère pertinent dans le cadre des maladies contagieuses. Les **processus d'inhibition** fonctionnent de manière inverse. L'apparition d'un événement au point $x \in D$ diminue la probabilité de survenue d'autres évènements aux environs de x. Ce type de processus est utile lorsqu'on veut modéliser, par exemple, l'implantation d'un animal et la définition de son territoire. Les processus de Markov et les processus de Gibbs sont des exemples de processus de contagion/inhibition.

Processus de Cox. Le processus de Cox ([Cox, 1955]) est un processus de Poisson hétérogène qui présente la caractéristique de considérer la mesure d'intensité $\lambda(.)$ comme une variable aléatoire. Ce type de processus peut être intéressant dans le domaine de la santé car il possible de penser que la fonction d'intensité du processus peut être amenée à varier en fonction du temps (prise en compte de la démographie de la population à risque pendant une période de temps, évolution temporelle de facteurs environnementaux, etc.).

2.4.6 Exemples

La Figure 3.3 permet de visualiser deux réalisations de processus de Poisson : un processus homogène et un processus non-homogène. On voit très nettement, que des structures spatiales d'événements apparaissent au sein de la réalisation du processus de Poisson non-homogène, ceci étant dû à la mesure d'intensité. Ces structures spatiales peuvent être liées à des zones où la population à risque est plus élevée (valeurs de $\mu(.)$ élevées). L'agrégation d'événements à ces zones n'est donc pas si atypique.

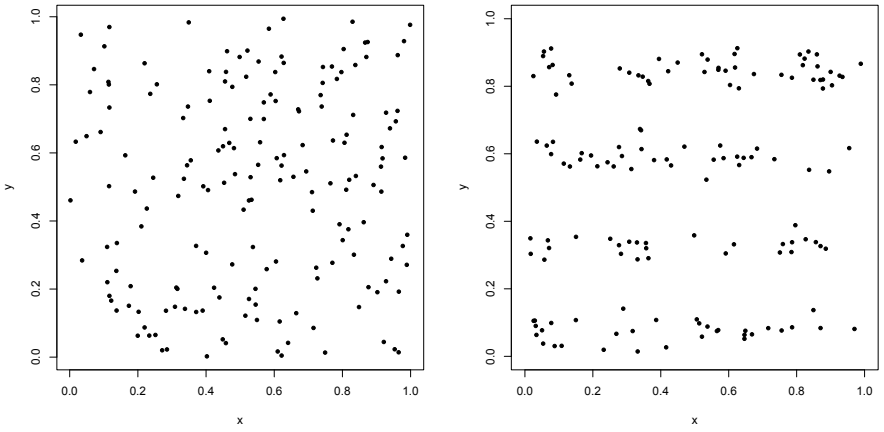


FIGURE 3.3 – (a) Réalisation d'un processus de Poisson homogène sur $D = [0,1]^2$, intensité $\lambda = 200$; (b) Réalisation d'un processus de Poisson hétérogène sur $D = [0,1]^2$ d'intensité $\lambda(x,y) = 400 * (\sin(24x) + \sin(24y))$.

3 Cadre général de la statistique de scan spatiale

Dans le domaine de la détection de clusters en statistique spatiale, il existe trois grandes classes de méthodes d'analyse. On distinguera les tests globaux, les tests de détection de clusters et les tests focalisés.

Tests globaux (ou généraux). Ils permettent de détecter une tendance globale à l'agrégation dans l'ensemble du domaine spatial étudié sans mettre en évidence la localisation des clusters éventuels. On distinguera les méthodes de [Alt and Vach, 1991], [Besag and Newell, 1991], [Cuzick and Edward, 1990], [Diggle and Chetwynd, 1991], [Grimson and Rose, 1991], [Moran, 1950],[Ranta et al., 1996], [Tango, 1995, Tango, 2000], [Walter, 1994].

Tests de détection de clusters (ou sans point source défini). Ils ont pour objectif la localisation sans connaissance *a priori* de clusters et le test de leur significativité. On distinguera les méthodes de [Turnbull et al., 1990], [Kulldorff and Nagarwalla, 1995, Kulldorff, 1997]. Aussi les statistiques de scan spatiales appartiennent à cette classe de méthodes.

Tests focalisés (ou avec point source défini). Ils permettent de définir un point source sur la base de connaissances *a priori* et de tester l'agrégation d'évènements autour de ce point. Par exemple, on pourrait imaginer que l'on veuille tester, à la suite d'une fuite de produits toxiques dans une usine, l'agrégation du nombre de malades autour du site. On distinguera le test de Stone [Stone, 2006], le test du score de Lawson-Waller [Lawson, 1993] et le test de Bithell [Bithell, 2007].

[Openshaw et al., 1987] ont développé une méthode nommée *Geographical Analysis Machine* (GAM) qui a pour objectif la détection de clusters potentiels sur une surface géographique en se basant sur l'utilisation d'un ensemble de cercles qui se superposent. Cependant, cette méthode souffre d'un problème de tests multiples car un test de significativité est réalisé pour chaque cercle. [Turnbull et al., 1990], ont proposé une technique afin de tester des clusters potentiels de taille différente, technique ajustée sur la multiplicité des tests par le biais de l'utilisation du rapport de vraisemblance. Les statistiques de scan spatiales proposées dans un premier temps par [Kulldorff and Nagarwalla, 1995] sont inspirées des deux méthodes précitées, à savoir la détection de clusters potentiels et le test de leur significativité sans problème de tests multiples. Cette méthode est actuellement la plus utilisée dans de nombreux champs d'applications en recherche médicale tels

que l'oncologie [Hjalmars et al., 1996, Kulldorff et al., 1997, Michelozzi et al., 2002, Thomas and Carlin, 2003, Klassen et al., 2005, DeChello and Sheehan, 2007, Amin et al., 2010], la cardiologie [Kuehl and Loffredo, 2006, Pedigo et al., 2011], les maladies infectieuses [Cousens et al., 2001, Fevre et al., 2001, D'Aignaux et al., 2002, Mostashari et al., 2003, Jennings et al., 2005, Wylie et al., 2005, Weisent et al., 2011, Wu et al., 2012], la gastro-entérologie [Ekbom et al., 1991, Green et al., 2006, Aamodt et al., 2008] ou encore la pédiatrie [George et al., 2001, Sankoh et al., 2001, Ozdenerol et al., 2005].

Les statistiques de scan spatiales présentent différentes caractéristiques spécifiques. Dans un premier temps, la forme de la fenêtre est uniquement circulaire, de taille variable. Cette forme a été choisie pour des raisons calculatoires car le nombre de paramètres liés à la fenêtre est limité, un cercle étant uniquement défini par son centre et son rayon. Dans un deuxième temps, la méthode permet de prendre en compte une intensité sous-jacente connue qui gouverne la distribution des évènements sous l'hypothèse nulle d'absence de cluster. Cette intensité μ sur un espace D peut prendre différentes formes en fonction de l'application. Par exemple, si $D \subset \mathbb{R}$, alors $\forall [a,b] \subseteq D$ $\mu([a,b])$, correspond à la mesure de Lebesgue sur $\mathbb{R}$, ou encore si $D \subset \mathbb{R}^2$ alors $\forall A \subseteq D$, $\mu(A)$ correspond à la mesure de Lebesgue sur $\mathbb{R}^2$. Cependant, la mesure μ peut prendre d'autres formes en fonction des applications. Par exemple, dans le cadre des données de santé, les épidémiologistes s'intéressent à la détection de clusters de cas de maladie. Or, il est important de prendre en compte l'hétérogénéité de la population sous-jacente afin de déterminer si la zone étudiée présente un nombre de cas effectivement élevé. En effet, si la méthode n'est pas ajustée sur la population sous-jacente, elle aura tendance à détecter des clusters dans les zones les plus peuplées, zones présentant, par conséquent, un nombre de cas de maladie plus élevé. Aussi, dans ce cas de figure, $\forall A \subseteq D$, $\mu(A)$ correspond à la population sous-jacente relative à A (ex : population d'un canton).

Par ailleurs, les statistiques de scan spatiales ont été étendues au cas spatio-temporel par [Kulldorff et al., 1998]. Elles permettent la détection de clusters d'évènements à la fois dans le temps et l'espace et ainsi, elle peuvent être qualifiées de statistiques de scan tridimensionnelles. Ces méthodes présentent les même caractéristiques que dans le cadre spatial, à l'exception de la forme de la fenêtre qui est cylindrique, de base et de hauteur variables.

Les méthodes statistiques de scan spatiales et spatio-temporelles se décomposent en deux étapes successives distinctes. La première représente la phase de détection de cluster d'évènements potentiellement atypique et la deuxième, consiste en l'inférence statistique de ces derniers. La principale différence entre le cas spatial et le cas spatio-temporel réside dans la phase détection.

3.1 Phase de détection

3.1.1 Cas spatial

Soit Z une fenêtre circulaire qui va se déplacer dans la totalité de l'ensemble spatial D. La taille de la fenêtre est variable et son rayon varie, en théorie, de 0 à $+\infty$. En pratique, il est considéré comme discret et sa valeur maximale est définie comme étant la longueur maximale séparant deux éléments de D. De surcroît, la méthode se limite aux zones Z contenant au maximum 50% des évènements. En effet, la détection d'un cluster comprenant plus de 50% des évènements reviendrait à mettre en évidence un nombre d'évènements beaucoup trop faible à l'extérieur et non à l'intérieur de ce dernier. Le balayage de l'ensemble spatial D est effectué par le centrage de la fenêtre circulaire en un élément de D et l'augmentation du rayon jusqu'aux limites définies *supra*. Cette procédure est réitérée pour chaque élément de D et conduit à l'ensemble discret fini $\mathcal{Z}$ défini par l'ensemble des zones Z de centres et tailles différents qui constitue l'ensemble des clusters potentiels (Figure 3.4).

3.1.2 Cas spatio-temporel

Considérons Z une fenêtre cylindrique dont la base représente l'espace et la hauteur la composante temporelle. Cette fenêtre de scan est de taille variable et se déplace sur l'ensemble de l'espace à trois dimensions, prenant comme centre un évènement ou encore un site (cas des données latticielles). A chaque position, la base de la fenêtre varie de 0 à un rayon tel que 50% du nombre d'évènements soit compris dans le cylindre. La hauteur de la fenêtre varie également de zéro à la moitié de la période temporelle étudiée. Cette procédure est réalisée pour chacun des évènements

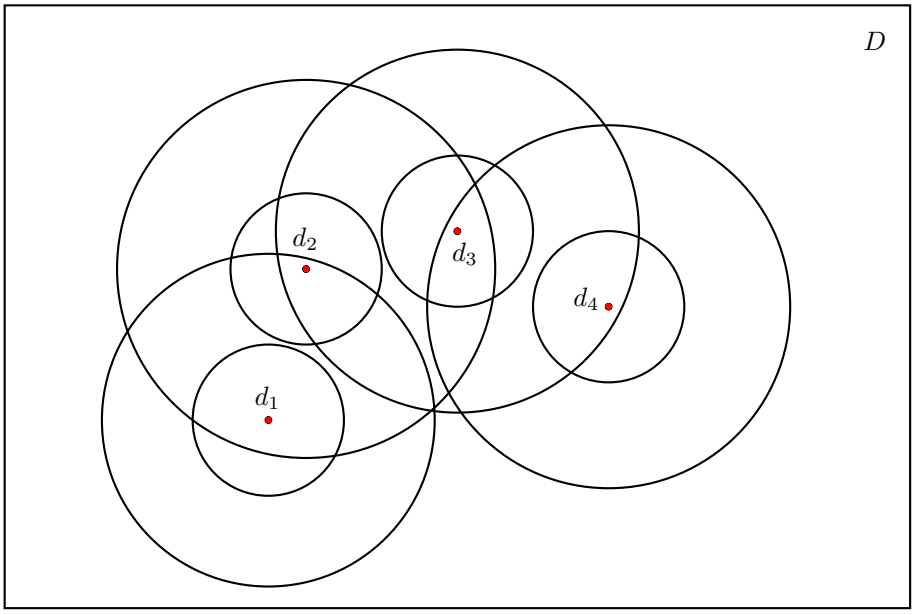


FIGURE 3.4 – Cas spatial - Ensemble $\mathcal{Z}$ de clusters potentiels

(ou sites) de l'espace tridimensionnel considéré et conduit ainsi à une collection $\mathcal{Z}$ de cylindres de base et de hauteur de tailles différentes, constituant l'ensemble des clusters spatio-temporels potentiels (Figure 3.5).

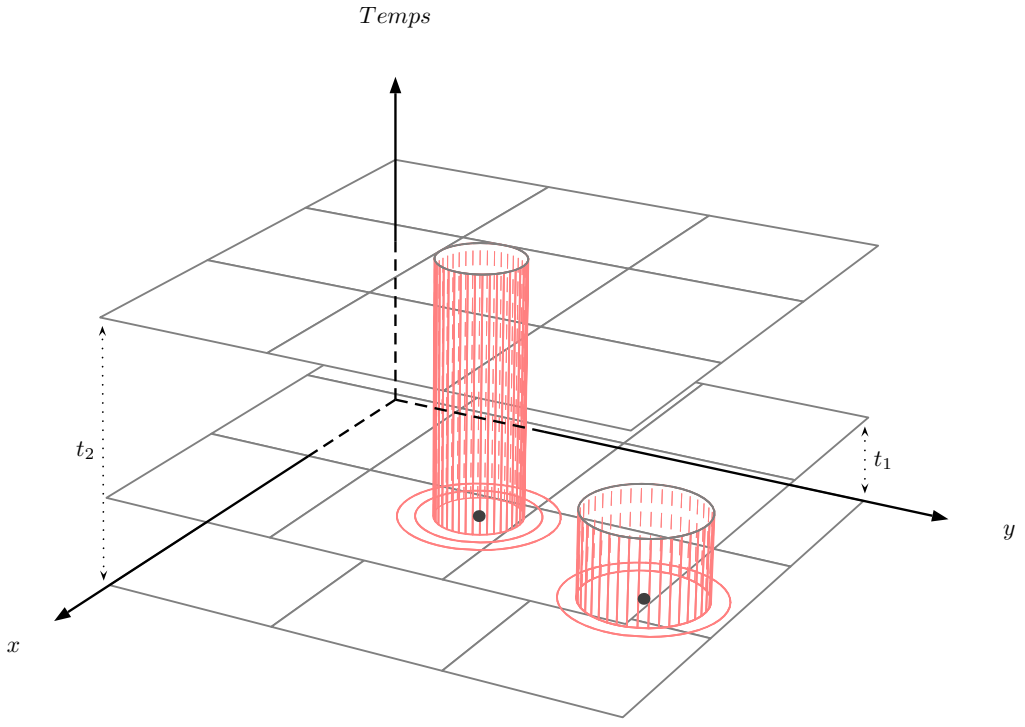


FIGURE 3.5 – Cas spatio-temporel - Ensemble $\mathcal{Z}$ de clusters potentiels.

3.1.3 Rapport de vraisemblance

Rappelons-nous que nous souhaitons tester l'hypothèse nulle $\mathcal{H}_0$ selon laquelle les évènements sont distribués de manière homogène sur l'espace étudié (*i.e.* les évènements observés sont distribués selon une loi de probabilité pré-spécifiée paramétrée par Θ_0) contre une hypothèse alternative $\mathcal{H}_1$ supportant la présence d'un cluster Z dans la répartition des évènements (*i.e.* les évènements observés à l'intérieur Z sont distribués selon la même loi de probabilité, mais présentant des paramètres différents Θ_1). Par ailleurs, la phase de détection conduit à un nombre élevé de clusters potentiels. Aussi, pour un cluster candidat Z de *taille fixe*, une fonction de vraisemblance sous $\mathcal{H}_0$, $L_{\mathcal{H}_0}$, et sous $\mathcal{H}_1$, $L_{\mathcal{H}_1}$ est calculée et la statistique de test de rapport de vraisemblance a pour expression

$$LR(Z) = \frac{\sup_{\Theta_1} L_{\mathcal{H}_1}}{\sup_{\Theta_0} L_{\mathcal{H}_0}}.$$

Pour Z de taille fixe, $LR(Z)$ est une fonction monotone croissante du nombre de cas à l'intérieur de Z. Or, comme la taille des clusters candidats varie, la statistique de test de rapport de vraisemblance a pour expression

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

Ainsi, parmi l'ensemble des clusters candidats $\mathcal{Z}$, celui qui maximise $LR(Z)$ est défini comme le cluster le plus probable (*Most Likely Cluster* (MLC)). C'est principalement ce cluster qui est testé lors de la phase d'inférence statistique. Cette technique permet d'éviter la problématique de multiplicité des tests car seul le MLC va être testé et non l'ensemble des clusters candidats issus de la phase de détection.

L'expression du rapport de vraisemblance λ est fonction du type de modèle utilisé pour modéliser la génération des évènements dans l'espace étudié. En effet, les premiers modèles développés par [Kulldorff and Nagarwalla, 1995, Kulldorff, 1997] sont les modèles de Bernoulli et modèles de Poisson, et sont respectivement explicités dans les sections 4.1 et 4.2. Depuis, plusieurs modèles ont été développés, notamment les modèles multinomiaux [Jung et al., 2010], ordinaux [Jung et al., 2007], normaux [Kulldorff et al., 2009] et exponentiels [Huang et al., 2007].

3.2 Phase d'inférence statistique

La phase de détection conduit à la mise en évidence du MLC et le test basé sur le rapport de vraisemblance permet d'évaluer la significativité de ce dernier. Cependant, la loi de probabilité, sous $\mathcal{H}_0$, de la statistique de test λ n'a pas de forme analytique. Afin de pallier ce problème, une procédure de test basée sur des simulations de Monte Carlo est utilisée. Les deux sections suivantes vont respectivement présenter le principe du test de Monte Carlo et l'application de ce dernier à la statistique de scan spatiale.

3.2.1 Principe du test de Monte Carlo

Un des problèmes importants dans la construction d'une statistique de test U repose sur le fait que dans certains cas il est impossible d'obtenir de manière analytique sa loi de probabilité sous l'hypothèse nulle. Pour palier ce problème, [Dwass, 1957] a proposé une procédure de test basée sur des simulations de Monte Carlo permettant de donner une approximation de la distribution de la statistique de test sous l'hypothèse nulle. Cette méthode a par la suite été étendue par [Barnard, 1963] et [Hope, 1968].

Soit U une statistique de test dont la distribution ne peut pas être déterminée de manière analytique. S'il est possible de générer R, $R \in \mathbb{N}$, échantillons de U sous l'hypothèse nulle $\mathcal{H}_0$, l'estimateur de maximum de vraisemblance de la probabilité critique, notée $\mathbb{P}_{mc}$, est donnée par le rapport entre le nombre de statistiques de test simulées, u_i $1 \leq i \leq R$, supérieures ou égales à la valeur observée de la statistique de test u_0 et le nombre total de simulations plus un $R+1$:

$$\mathbb{P}_{mc} = \frac{\#\{u_i > u_0\}}{R+1}. \tag{3.1}$$

Si $\mathbb{P}_{mc}$ est inférieure à un risque de première espèce α fixé, alors $\mathcal{H}_0$ est rejetée. Un des principaux avantages du test de Monte Carlo réside dans le fait que le test conserve correctement le risque α nominal. Ceci est simplement réalisable en choisissant un nombre de simulations R tel que $m = \alpha(R+1) \in \mathbb{N}$. Par exemple, si $\alpha = 0.05$ alors la probabilité de rejeter $\mathcal{H}_0$ est exactement 0.05 si R =19, 99, 999... Prenons $R = 19$ et $\alpha = 0.05$. Sous $\mathcal{H}_0$, la statistique de test observée u_0 et celles simulées u_i possèdent la même loi de probabilité. Comme les u_i sont indépendantes, la probabilité que la statistique de test observée ait une valeur supérieure à toutes les autres statistiques de test simulées est $1/20 = 0.05$. Or, si $\mathbb{P}_{mc} \leq \alpha = 0.05$, $\mathcal{H}_0$ est rejetée. Pour $R = 19$, le test conserve correctement le risque de première espèce $\alpha = 0.05$.

3.2.2 Application à la statistique de scan spatiale

Dans le cadre de la statistique de scan spatiale, et quel que soit le modèle utilisé, des simulations de Monte Carlo sont utilisées afin d'approximer la distribution du rapport de vraisemblance sous l'hypothèse nulle d'absence de cluster [Kulldorff, 1999]. Ainsi, la probabilité critique associée au MLC, $\mathbb{P}_{MLC}$ détecté peut être déterminée. La méthode est divisée en quatre étapes distinctes :

Algorithme 3 - Calcul de $\mathbb{P}_{MLC}$

Etape 1 : Calcul de λ sur données observées (MLC)
Etape 2 : simulation de R jeux de données sous $\mathcal{H}_0$ (fonction du modèle)
Etape 3 : A partir des simulations,
  Pour chaque simulation i, $1 \leq i \leq R$ **Faire**
    Calcul de la fonction de vraisemblance sous $\mathcal{H}_0$
    Phase de détection du MLC $\rightarrow \lambda$
  Fin Pour
Etape 4 :
  Ordonner de manière croissante les $R+1$ valeurs de λ (simulations + données réelles)
  Calcul de la probabilité critique $\mathbb{P}_{MLC} = \frac{r_0}{R+1}$, r_0 étant le rang du λ observé.

Par exemple, si l'on réalise 9999 simulations, alors, le cluster est significatif au risque de première espèce $\alpha = 0.05$ si la statistique de test λ calculée sur les données réelles se trouve parmi les 500 plus hautes valeurs des λ :

$$\mathbb{P}_{MLC} = \frac{r_0}{R+1} = \frac{500}{9999+1} = 0.05$$

La simulation des évènements sous $\mathcal{H}_0$ est bien entendu fonction du modèle choisi.

3.3 Notion de clusters secondaires

Jusqu'à présent, la méthode permet de mettre en évidence le MLC et de tester sa significativité. Cependant, il peut être également intéressant d'analyser les clusters potentiels présentant des valeurs élevées de rapports de vraisemblance sans pour autant avoir été déterminés comme MLC. Ces derniers sont définis comme clusters secondaires et peuvent être distingués selon deux types. Le premier correspond aux clusters secondaires qui ne se superposent avec le MLC et, par conséquent, partagent une ensemble d'évènements. Le deuxième dénote les clusters qui ne superposent pas au MLC. Le premier type s'avère être, dans la plupart des cas, de faible intérêt bien qu'il nous rappelle le fait que le MLC obtenu n'est simplement qu'une estimation du "vrai" cluster dans la réalité. Le deuxième type de cluster est nettement plus intéressant car l'analyste s'intéresse à des clusters qui sont situés dans une autre partie de la zone géographique étudiée.

L'inférence statistique des clusters secondaires se base sur les simulations de Monte Carlo ayant été réalisées pour tester la significativité du MLC. Les valeurs des rangs des rapports de vraisemblance associés sont comparés aux simulations et une probabilité critique est calculée. Cependant, cette procédure d'inférence présente un caractère conservatif et conduit à une perte de puissance statistique [Kulldorff, 1997, Kulldorff, 1999]. En effet, le test de significativité d'un cluster secondaire est réalisé sans prendre en compte l'existence préalable du MLC. Aussi, ce cluster secondaire

s'avèrera significatif s'il est capable, par lui-même, de rejeter l'hypothèse nulle, quelle que soit la significativité des autres clusters. Par analogie, ce problème peut être considéré comme une analyse de régression dans laquelle chaque variable explicative est entrée dans un modèle simple et évaluée sans ajustement sur les autres variables explicatives.

Une alternative à ce caractère conservatif a été proposée par [Zhang and Kulldorff, 2010]. Ces derniers proposent deux méthodes pour ajuster les probabilités critiques des clusters secondaires à la présence du MLC. D'une part, les auteurs proposent de retirer les données relatives au MLC et ainsi relancer l'analyse sur un jeu de données réduit. Cette façon de procéder est réitérée jusqu'à ce qu'aucun MLC ne soit significatif. Cette méthode permet de pallier le conservatisme mais, en contrepartie, diminue le nombre d'évènements observés à chaque étape. D'autre part, les auteurs proposent de remplacer les effectifs observés au sein du MLC par les effectifs théoriques sous l'hypothèse nulle d'absence de cluster.

4 Principaux modèles

4.1 Modèle de Bernoulli

4.1.1 Conditions d'applications

Dans ce modèle, chaque évènement est modélisé par une variable de Bernoulli de paramètre p. Ces variables peuvent représenter des individus présentant une pathologie ou non (ex : malade / non malade), ou encore des individus présentant différents type de pathologie, comme par exemple le premier ou dernier stade d'un cancer (ex : stade 1 / stade 4). Il est important de préciser que les données utilisées peuvent refléter les cas et contrôles d'une population sous-jacente (échantillon), ou alors la population dans son ensemble (étude exhaustive). De surcroît, les données peuvent être représentées sous forme de processus ponctuel marqué, la localisation de chaque observation étant nécessaire, ou alors, sous forme de données latticielles, les observations, supposées indépendantes entre elles, étant ainsi agrégées à une unité spatiale (ex : canton). Dans ce dernier cas, l'unité spatiale fait office d'unité statistique et par conséquent le nombre de cas observés dans une unité spatiale est distribué selon une loi binomiale dont les paramètres sont explicités *infra*.

4.1.2 Modèle

Données ponctuelles. Considérons $\mathbf{X}$ un processus spatial ponctuel sur $D \subset \mathbb{R}^2$ de marque $m \in K = \{0, 1\}$ à valeurs dans l'espace d'états $E = \mathbb{N}$. Posons X_A, $A \in \mathcal{B}_b(D)$, la variable aléatoire qui associe à A le nombre de cas ($m = 1$) observés et $\mu(A)$ le nombre de points ($m = 0$ ou $m = 1$) observés dans A. La variable aléatoire X_A est distribuée selon une loi binomiale $\mathcal{B}(\mu(A), p)$.

Données latticielles. Considérons D un ensemble discret fixé, $D \subset \mathbb{R}^2$, structuré par un graphe de voisinage $\mathcal{G} = \{D, A\}$ et muni d'une matrice d'adjacence $U = (u_{ij})$ et d'une matrice de poids $W = (w_{ij})$. Soit $X = \{X_d, d \in D\}$ un processus aléatoire indexé par $D = \{d_1, d_2, ..., d_n\}$ et à valeurs dans l'espace d'états $E = \mathbb{N}$. Dans ce modèle, w_{ij} est défini ici comme la distance euclidienne entre le site d_i et le site d_j, $1 \leq i, j, \leq n$.

$\forall d \in D$, $X_d \sim \mathcal{B}(\mu(d), p)$ ou $\mu(d)$ correspond au nombre d'évènements observés au sein de la localisation d. Par construction,

$$\forall A \subset D \ , \ \mu(A) = \sum_{d \in A} \mu(d),$$

et par stabilité par la somme de la loi binomiale

$$X_A = \sum_{d \in A} X_d \text{ et } X_A \sim \mathcal{B}(\mu(A), p).$$

Hypothèses de test. L'objectif est de tester l'hypothèse nulle, $\mathcal{H}_0$, d'absence de cluster stipulant que dans l'ensemble de l'aire étudiée, chaque observation a la même probabilité p d'être un cas contre une hypothèse alternative, $\mathcal{H}_1$, supportant l'existence d'au moins une zone $Z \in \mathcal{Z}$ dans

laquelle chaque observation a une probabilité p d'être un cas et que les observations se trouvant à l'extérieur de Z ont une probabilité q, $q < p$, d'être un cas. De manière plus formelle, les hypothèses de test associées au modèle de Bernoulli sont les suivantes :

$$\begin{cases} \mathcal{H}_0 \ : \ p = q, \ X_A \sim \mathcal{B}(\mu(A), p) \, \forall A \subset D \\ \mathcal{H}_1 \ : \ p > q, \exists \, Z \in \mathcal{Z} / X_A \sim \mathcal{B}(\mu(A), p) \, \forall A \subset Z \text{ et } X_A \sim \mathcal{B}(\mu(A), q) \, \forall A \subset Z^c \end{cases}$$

Autrement dit, sous $\mathcal{H}_1$, le nombre de cas à l'intérieur de la zone suit une loi binomiale $\mathcal{B}(\mu(A), p)$ et à l'extérieur de la zone, une loi binomiale de $\mathcal{B}(\mu(A), q)$, avec $p > q$.

4.1.3 Fonction de vraisemblance

Deux approches équivalentes vont être explicitées afin de définir l'expression de la fonction de vraisemblance. La première, énoncée par [Kulldorff, 1997] se base sur les observations individuelles (cas/témoins). La deuxième, que nous explicitons, est appliquée à des données latticielles.

Approche de M. Kulldorff. Sous $\mathcal{H}_0$, posons o_i une observation, $1 \leq i \leq \mu(D)$, où $\mu(D)$ correspond au nombre total d'observations sur la région étudiée D. Chaque observation o_i peut soit prendre la valeur 1 avec une probabilité p soit la valeur 0 avec une probabilité $1-p$. Considérons les variables aléatoires O_i, $1 \leq i \leq \mu(D)$, distribuées selon une loi de Bernoulli $\mathcal{B}(1, p)$. La fonction de vraisemblance a donc pour expression :

$$L(p) = \prod_{i=1}^{\mu(D)} \mathbb{P}(O_i = o_i) = \prod_{i=1}^{\mu(D)} p^{o_i} (1-p)^{1-o_i}.$$

Posons n_D le nombre de cas observés dans la région étudiée D. $L(p)$ atteint son maximum lorsque $p = n_D / \mu(D)$:

$$L_0 = \max_p L(p) = \prod_{i=1}^{\mu(D)} \left(\frac{n_D}{\mu(D)} \right)^{o_i} \left(1 - \frac{n_D}{\mu(D)} \right)^{1-o_i} = \left(\frac{n_D}{\mu(D)} \right)^{n_D} \left(\frac{\mu(D) - n_D}{\mu(D)} \right)^{\mu(D)-n_D}.$$

Sous $\mathcal{H}_1$, chaque observation o_i inclue dans Z, $1 \leq i \leq \mu(Z)$, peut soit prendre la valeur 1 avec une probabilité p soit la valeur 0 avec une probabilité $1 - p$. Par ailleurs, chaque observation o_i non inclue dans Z peut soit prendre la valeur 1 avec une probabilité q soit la valeur 0 avec une probabilité $1 - q$. Aussi, considérons les variables aléatoires O_i, $1 \leq i \leq \mu(D)$, dont la distribution est fonction de l'appartenance à Z :

$$\begin{cases} O_i \sim \mathcal{B}(1, p) \text{ si } O_i \in Z \\ O_i \sim \mathcal{B}(1, q) \text{ sinon.} \end{cases}$$

La fonction de vraisemblance sous $\mathcal{H}_1$ a donc pour expression :

$$L(Z, p, q) = \prod_{i \in Z} \mathbb{P}(O_i = o_i) \prod_{i \notin Z} \mathbb{P}(O_i = o_i) = \prod_{i \in Z} p^{o_i} (1-p)^{1-o_i} \prod_{i \notin Z} q^{o_i} (1-q)^{1-o_i}.$$

Posons n_Z le nombre de cas observés dans la zone Z et $\mu(Z)$ le nombre total d'observations dans la zone Z. La fonction de vraisemblance s'écrit alors :

$$L(Z, p, q) = p^{n_Z} (1-p)^{\mu(Z) - n_Z} q^{n_D - n_Z} (1-q)^{[(\mu(D) - \mu(Z)) - (n_D - n_Z)]}$$

L'objectif est de déterminer le MLC. Aussi, doit-on déterminer la zone $\widehat{Z}$ qui maximise $L(Z, p, q)$ [2]. Ceci est réalisé en deux étapes successives. La première consiste à maximiser la fonction de vraisemblance conditionnée en Z (*i.e.* déterminer les estimateur de maximum de vraisemblance de p et q). Posons $L(Z)$ telle que

$$L(Z) = \sup_{p > q} L(Z, p, q).$$

2. Il est important de noter que $\widehat{Z}$ est considéré comme l'estimateur de maximum de vraisemblance du paramètre Z.

$L(Z,p,q)$ atteint son maximum lorsque $p = n_Z/\mu(Z)$ et $q = (n_D - n_Z)/(\mu(D) - \mu(Z))$. Aussi $L(Z)$ peut s'écrire de la manière suivante :

$$L(Z) = \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(1 - \frac{n_Z}{\mu(Z)}\right)^{\mu(Z)-n_Z} \left(\frac{n_D - n_Z}{\mu(D) - \mu(Z)}\right)^{n_D-n_Z} \times \left(1 - \frac{n_D - n_Z}{\mu(D) - \mu(Z)}\right)^{[(\mu(D)-\mu(Z))-(n_D-n_Z)]}.$$

La deuxième étape consiste à déterminer la zone $\widehat{Z} = \{Z : L(Z) \geq (L(Z')\, \forall Z' \in \mathcal{Z}\}$.

Approche sur données latticielles. En considérant les hypothèses du modèle nous pouvons écrire, sous $\mathcal{H}_0$, que

$$\forall d_i \in D,\ X_{d_i} \sim \mathcal{B}(\mu(d_i), p) \text{ et } \mathbb{P}(X_{d_i} = n_{d_i}) = \binom{\mu(d_i)}{n_{d_i}} p^{n_{d_i}} (1-p)^{\mu(d_i)-n_{d_i}},$$

les X_{d_i} étant considérées indépendantes. Par conséquent la fonction de vraisemblance a pour expression

$$L(p) = \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_{d_i} = n_{d_i}\}\right) = \prod_{i=1}^{n} \mathbb{P}(X_{d_i} = n_{d_i}) = \prod_{i=1}^{n} \binom{\mu(d_i)}{n_{d_i}} p^{n_{d_i}} (1-p)^{\mu(d_i)-n_{d_i}},$$

se résumant, après simplification, à

$$L(p) = p^{n_D} (1-p)^{\mu(D)-n_D} \prod_{i=1}^{n} \binom{\mu(d_i)}{n_{d_i}}.$$

$L(p)$ atteint son maximum lorsque $p = n_D/\mu(D)$:

$$L_0 = \sup_p L(p) = \left(\frac{n_D}{\mu(D)}\right)^{n_D} \left(1 - \frac{n_D}{\mu(D)}\right)^{\mu(D)-n_D} \prod_{i=1}^{n} \binom{\mu(d_i)}{n_{d_i}}.$$

En considérant le modèle sous $\mathcal{H}_1$, nous obtenons

$$\forall d_i \in Z,\ X_{d_i} \sim \mathcal{B}(\mu(d_i), p) \text{ et } \mathbb{P}(X_{d_i} = n_{d_i}) = \binom{\mu(d_i)}{n_{d_i}} p^{n_{d_i}} (1-p)^{\mu(d_i)-n_{d_i}},$$
$$\forall d_i \in Z^c,\ X_{d_i} \sim \mathcal{B}(\mu(d_i), q) \text{ et } \mathbb{P}(X_{d_i} = n_{d_i}) = \binom{\mu(d_i)}{n_{d_i}} q^{n_{d_i}} (1-q)^{\mu(d_i)-n_{d_i}}.$$

En conséquence la fonction de vraisemblance a pour expression

$$L(Z,p,q) = \mathbb{P}\left(\bigcap_{i\in Z} \{X_{d_i} = n_{d_i}\}\right) \mathbb{P}\left(\bigcap_{i\in Z^c} \{X_{d_i} = n_{d_i}\}\right)$$

$$L(Z,p,q) = \prod_{i\in Z} \mathbb{P}(X_{d_i} = n_{d_i}) \prod_{i\in Z^c} \mathbb{P}(X_{d_i} = n_{d_i})$$

$$L(Z,p,q) = \prod_{i\in Z} \binom{\mu(d_i)}{n_{d_i}} p^{n_{d_i}} (1-p)^{\mu(d_i)-n_{d_i}} \prod_{i\in Z^c} \binom{\mu(d_i)}{n_{d_i}} q^{n_{d_i}} (1-q)^{\mu(d_i)-n_{d_i}},$$

se résumant, après simplification, à

$$L(Z,p,q) = p^{n_Z} (1-p)^{\mu(Z)-n_Z} q^{n_D-n_Z} (1-q)^{[\mu(D)-\mu(Z)-(n_D-n_Z)]} \prod_{i=1}^{n} \binom{\mu(d_i)}{n_{d_i}}$$

En conditionnant par rapport à Z, $L(Z,p,q)$ atteint son maximum lorsque $p = n_Z/\mu(Z)$ et $q = (n_D - n_Z)/(\mu(D) - \mu(Z))$:

$$L(Z) = \sup_{p>q} L(Z,p,q) = \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(1 - \frac{n_Z}{\mu(Z)}\right)^{\mu(Z)-n_Z} \left(\frac{n_D - n_Z}{\mu(D)-\mu(Z)}\right)^{n_D-n_Z} \times \left(1 - \frac{n_D - n_Z}{\mu(D)-\mu(Z)}\right)^{[(\mu(D)-\mu(Z))-(n_D-n_Z)]} \prod_{i=1}^{n} \binom{\mu(d_i)}{n_{d_i}}.$$

La deuxième étape consiste à déterminer la zone $\widehat{Z} = \{Z : L(Z) \geq (L(Z')\, \forall Z' \in \mathcal{Z}\}$.

Remarque 3.2. *Les expressions de L_0 et $L(Z)$ issues de la méthode présentée par M. Kulldorff (données ponctuelles) et celles basées sur les données latticielles sont, à une constante près, égales.*

4.1.4 Test du rapport de vraisemblance

A la suite de la détermination du MLC, $\widehat{Z}$, sa significativité est testée au moyen d'un test de rapport de vraisemblance qui pour expression :

$$\lambda = \frac{\sup_{Z\in\mathcal{Z}, p>q} L(Z,p,q)}{\sup_p L(p)} = \frac{L(\widehat{Z})}{L_0}$$

Remarque 3.3. *L_0 ne dépend pas de Z. Aussi, le cluster le plus probable maximise à la fois $L(Z)$ et la statistique de test λ.*

Remarque 3.4. *L'expression de λ est identique quel que soit le type de données : ponctuelles ou latticielles.*

Remarque 3.5. *Dans le cadre du modèle de Bernoulli, la distribution de λ ne possède pas de forme analytique. Elle est approximée au moyen de simulations de Monte Carlo sous $\mathcal{H}_0$.*

4.2 Modèle de Poisson

4.2.1 Conditions d'application

Le modèle de Poisson s'applique uniquement aux données latticielles. Les données sont ainsi agrégées à une unité spatiale (ex : iris, canton, département, etc.) et le nombre de cas au sein de chaque unité spatiale est distribué selon une loi de Poisson dont l'espérance est proportionnelle à une mesure d'intensité μ caractérisant la population sous-jacente de l'unité spatiale (ex : population du canton). Par ailleurs, ce type de modèle est conseillé lorsque le nombre de cas est négligeable face à la taille de la population sous-jacente.

4.2.2 Modèle

Considérons D un ensemble discret fixé, $D \subset \mathbb{R}^2$, structuré par un graphe de voisinage $\mathcal{G} = \{D, A\}$ et muni d'une matrice d'adjacence $U = (u_{ij})$ et d'une matrice de poids $W = (w_{ij})$. Soit $X = \{X_d, d \in D\}$ un processus aléatoire indexé par $D = \{d_1, d_2, ..., d_n\}$ et à valeurs dans l'espace d'états $E = \mathbb{N}$.

$\forall d \in D$, $X_d \sim \mathcal{P}(p\mu(d))$ ou $\mu(d)$ correspond à la population sous-jacente observée à la localisation d. Par construction,

$$\forall A \subset D\ ,\ \mu(A) = \sum_{d\in A} \mu(d),$$

et par stabilité par la somme de la loi de Poisson

$$X_A = \sum_{d\in A} X_d \text{ et } X_A \sim \mathcal{P}(p\mu(A)).$$

L'objectif est de tester l'hypothèse nulle $\mathcal{H}_0$ d'absence de cluster stipulant que dans l'ensemble de l'aire étudiée, le nombre attendu de cas dans chaque unité spatiale d_i, $1 \leq i \leq n$ est égal à

$p\mu(d_i)$, contre une hypothèse alternative, $\mathcal{H}_1$, supportant l'existence d'au moins une zone $Z \in \mathcal{Z}$ dans laquelle le nombre de cas attendus pour chaque unité spatiale est égal à $p\mu(d_i)$ et égal à $q\mu(d_i)$ à l'extérieur de Z, $p > q$. De manière plus formelle, les hypothèses de test du modèle de Poisson sont les suivantes :

$$\begin{cases} \mathcal{H}_0 \ : \ p = q, \ X_A \sim \mathcal{P}(p\mu(A)), \ \forall A \subset D \\ \mathcal{H}_1 \ : \ p > q, \ X_A \sim \mathcal{P}(p\mu(A \cap Z) + q\mu(A \cap Z^c)), \ \forall A \subset D \end{cases}$$

Autrement dit, sous $\mathcal{H}_1$, le nombre de cas à l'intérieur de la zone Z est distribué selon une loi de Poisson $\mathcal{P}(p\mu(Z))$ et, à l'extérieur de la zone, selon une loi de Poisson $\mathcal{P}(q\mu(Z^c))$, avec $p > q$.

4.2.3 Fonction de vraisemblance

Deux approches équivalentes vont être explicitées afin de définir l'expression de la fonction de vraisemblance. La première, énoncée par [Kulldorff, 1997] n'est pas classique dans son explicitation car l'expression de départ qui permet de construire la fonction de vraisemblance n'est pas celle d'une loi de Poisson ordinaire. Nous proposons une deuxième approche qui se base sur l'expression classique d'une fonction de vraisemblance de Poisson.

Approche de M. Kulldorff. Dans un premier temps, définissons la probabilité d'observer n_D cas sur D de population sous-jacente $\mu(D)$:

$$\mathbb{P}\left(X_D = n_D\right) = \frac{e^{-p\mu(Z)-q(\mu(D)-\mu(Z))}(p\mu(Z) + q[\mu(D) - \mu(Z)])^{n_D}}{n_D!}$$

Dans un deuxième temps, définissons la fonction de densité de probabilité, $f(d_i)$, qu'un cas soit observé au sein du site d_i, $1 \leq i \leq n$:

$$f(d_i) = \begin{cases} \dfrac{p\mu(d_i)}{p\mu(Z) + q(\mu(D) - \mu(Z))} \text{ si } d_i \in Z \\ \\ \dfrac{q\mu(d_i)}{p\mu(Z) + q(\mu(D) - \mu(Z))} \text{ si } d_i \notin Z \end{cases}$$

Sous $\mathcal{H}_0$ (*i.e.* $p = q$), la fonction de vraisemblance a pour expression :

$$L(p) = \mathbb{P}(X_D = n_D) \prod_{i=1}^{n} f(d_i) = \frac{e^{-p\mu(D)}(p\mu(D))^{n_D}}{n_D!} \prod_{i=1}^{n} \frac{\mu(d_i)}{\mu(D)} = \frac{e^{-p\mu(D)}}{n_D!}(p)^{n_D} \prod_{i=1}^{n} \mu(d_i),$$

et atteint son maximum lorsque $p = n_D/\mu(D)$:

$$L_0 = \max_p L(p) = \frac{e^{-n_D}}{n_D!} \left(\frac{n_D}{\mu(D)}\right)^{n_D} \prod_{i=1}^{n} \mu(d_i).$$

Sous $\mathcal{H}_1$ (*i.e.* $p > q$)), la fonction de vraisemblance s'écrit :

$$\begin{aligned} L(Z,p,q) = & \frac{e^{-p\mu(Z)+q(\mu(D)-\mu(Z))}(p\mu(Z) + q[\mu(D) - \mu(Z)])^{n_D}}{n_D!} \\ & \times \prod_{d_i \in Z} \frac{p\mu(d_i)}{p\mu(Z) + q(\mu(D) - \mu(Z))} \prod_{d_i \notin Z} \frac{q\mu(d_i)}{p\mu(Z) + q(\mu(D) - \mu(Z))}, \end{aligned}$$

et après simplification

$$L(Z,p,q) = \frac{e^{-p\mu(Z)-q(\mu(D)-\mu(Z))}}{n_D!} p^{n_Z} q^{n_D - n_Z} \prod_{i=1}^{n} \mu(d_i).$$

L'objectif est de déterminer le MLC. Aussi, doit-on déterminer la zone $\widehat{Z}$ qui maximise $L(Z,p,q)$. A l'instar du modèle de Bernoulli, ceci est réalisé en deux étapes successives. La première consiste à maximiser la fonction de vraisemblance conditionnée en Z (*i.e.* déterminer les estimateurs de maximum de vraisemblance de p et q). Posons $L(Z)$ telle que

$$L(Z) = \sup_{p>q} L(Z,p,q). \tag{3.2}$$

$L(Z)$ atteint son maximum lorsque $p = n_Z/\mu(Z)$ et $q = (n_D - n_Z)/(\mu(D) - \mu(Z))$. Aussi $L(Z)$ peut s'écrire de la manière suivante :

$$L(Z) = \frac{e^{-n_D}}{n_D!}\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_D - n_Z}{\mu(D)-\mu(Z)}\right)^{n_D-n_Z}\prod_{i=1}^{n}\mu(d_i).$$

La deuxième étape consiste à déterminer la zone $\widehat{Z} = \{Z : L(Z) \geq (L(Z')\,\forall Z' \in \mathcal{Z}\}$.

Approche traditionnelle. En considérant les hypothèses du modèle nous pouvons écrire, sous $\mathcal{H}_0$, que

$$\forall d_i \in D,\ X_{d_i} \sim \mathcal{P}(p\mu(d_i)) \text{ et } \mathbb{P}(X_{d_i} = n_{d_i}) = \frac{e^{-p\mu(d_i)}[p\mu(d_i)]^{n_{d_i}}}{n_{d_i}!},$$

les X_{d_i} étant considérées indépendantes. En conséquence, la fonction de vraisemblance a pour expression

$$L(p) = \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_{d_i} = n_{d_i}\}\right) = \prod_{i=1}^{n}\mathbb{P}(X_{d_i} = n_{d_i}) = \prod_{i=1}^{n}\frac{e^{-p\mu(d_i)}[p\mu(d_i)]^{n_{d_i}}}{n_{d_i}!},$$

se résumant, après simplification, à

$$L(p) = e^{-p\mu(D)}p^{n_D}\prod_{i=1}^{n}\frac{\mu(d_i)^{n_{d_i}}}{n_{d_i}!}.$$

$L(p)$ atteint son maximum lorsque $p = n_D/\mu(D)$:

$$L_0 = \max_p L(p) = e^{-n_D}\left(\frac{n_D}{\mu(D)}\right)^{n_D}\prod_{i=1}^{n}\frac{\mu(d_i)^{n_{d_i}}}{n_{d_i}!}.$$

En considérant le modèle sous $\mathcal{H}_1$, nous obtenons

$$\forall d_i \in Z,\ X_{d_i} \sim \mathcal{P}(p\mu(d_i)) \text{ et } \mathbb{P}(X_{d_i} = n_{d_i}) = \frac{e^{-p\mu(d_i)}[p\mu(d_i)]^{n_{d_i}}}{n_{d_i}!},$$
$$\forall d_i \in Z^c,\ X_{d_i} \sim \mathcal{P}(q\mu(d_i)) \text{ et } \mathbb{P}(X_{d_i} = n_{d_i}) = \frac{e^{-q\mu(d_i)}[q\mu(d_i)]^{n_{d_i}}}{n_{d_i}!}.$$

En conséquence la fonction de vraisemblance a pour expression

$$L(Z,p,q) = \mathbb{P}\left(\bigcap_{i\in Z}\{X_{d_i} = n_{d_i}\}\right)\mathbb{P}\left(\bigcap_{i\in Z^c}\{X_{d_i} = n_{d_i}\}\right),$$

$$L(Z,p,q) = \prod_{i\in Z}\mathbb{P}(X_{d_i} = n_{d_i})\prod_{i\in Z^c}\mathbb{P}(X_{d_i} = n_{d_i}),$$

$$L(Z,p,q) = \prod_{i\in Z}\frac{e^{-p\mu(d_i)}[p\mu(d_i)]^{n_{d_i}}}{n_{d_i}!}\prod_{i\in Z^c}\frac{e^{-q\mu(d_i)}[q\mu(d_i)]^{n_{d_i}}}{n_{d_i}!},$$

se résumant, après simplification, à

$$L(Z,p,q) = e^{-p\mu(Z)}p^{n_Z}e^{-q(\mu(D)-\mu(Z))}q^{n_D-n_Z}\prod_{i=1}^{n}\frac{\mu(d_i)^{n_{d_i}}}{n_{d_i}!}.$$

$L(Z,p,q)$ atteint son maximum lorsque $p = n_Z/\mu(Z)$ et $q = (n_D - n_Z)/(\mu(D) - \mu(Z))$. Aussi $L(Z)$, définie en (3.2) peut s'écrire de la manière suivante :

$$L(Z) = \max_{p>q} L(Z,p,q) = e^{-n_D}\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_D-n_Z}{\mu(D)-\mu(Z)}\right)^{n_D-n_Z}.$$

La dernière étape est identique à l'approche de M. Kulldorff, à savoir déterminer la zone $\widehat{Z} = \{Z : L(Z) \geq (L(Z')\,\forall Z' \in \mathcal{Z}\}$.

4.2.4 Test du rapport de vraisemblance

Après détermination du MLC, $\widehat{Z}$, sa significativité est testée au moyen d'un test de rapport de vraisemblance d'expression

$$\lambda = \frac{\sup_{Z\in\mathcal{Z},p>q} L(Z,p,q)}{\sup_p L(p)} = \frac{L(\widehat{Z})}{L_0}.$$

Les expressions des fonctions de vraisemblance par le biais des deux approches précitées mènent à la même formule de la statistique de test. En effet, l'approche de M. Kulldorff conduit à

$$\begin{aligned}\lambda &= \frac{\sup_{Z\in\mathcal{Z}} \frac{e^{-n_D}}{n_D!}\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_D-n_Z}{\mu(D)-\mu(Z)}\right)^{n_D-n_Z}\prod_{i=1}^{n}\mu(d_i)}{\frac{e^{-n_D}}{n_D!}\left(\frac{n_D}{\mu(D)}\right)^{n_D}\prod_{i=1}^{n}\mu(d_i)}\\ &= \sup_{Z\in\mathcal{Z}}\frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_D-n_Z}{\mu(D)-\mu(Z)}\right)^{n_D-n_Z}}{\left(\frac{n_D}{\mu(D)}\right)^{n_D}},\end{aligned}$$

et l'approche traditionnelle mène à l'expression

$$\begin{aligned}\lambda &= \frac{\sup_{Z\in\mathcal{Z}}\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_D-n_Z}{\mu(D)-\mu(Z)}\right)^{(n_D-n_Z)}}{\left(\frac{n_D}{\mu(D)}\right)^{n_D}}\\ &= \sup_{Z\in\mathcal{Z}}\frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_D-n_Z}{\mu(D)-\mu(Z)}\right)^{n_D-n_Z}}{\left(\frac{n_D}{\mu(D)}\right)^{n_D}}.\end{aligned}$$

4.2.5 Ajustement des analyses sur des facteurs de confusion

Dans le cadre de la détection de clusters atypiques, il est intéressant d'ajuster sur des facteurs de confusion pour trois raisons principales :

— Le facteur est lié à la maladie étudiée. Par exemple, dans le cas de la Maladie de Crohn, il existe une différence avérée d'incidence entre les hommes et les femmes, d'où l'intérêt d'ajuster sur le sexe ;
— Le facteur n'est pas distribué de manière aléatoire dans la région étudiée. Nous pouvons imaginer l'exemple de certaines zones dans lesquelles la moyenne d'âge est nettement plus élevée que dans le reste de la population étudiée, et, par conséquent, joue un rôle dans l'incidence de la pathologie étudiée ;

— La volonté de détecter des clusters que ne puissent pas être expliqués par un facteur de confusion.

Le modèle de Poisson permet l'ajustement sur des facteurs de confusion de nature catégorielle [Kulldorff, 1997, Kulldorff et al., 1997, Kulldorff et al., 1998]. Sous l'hypothèse nulle, $\mathcal{H}_0$ d'absence de cluster, la variable aléatoire X_{d_i}, qui associe à un site d_i, $1 \leq i \leq n$, son nombre de cas observés, est distribuée selon une loi de Poisson $\mathcal{P}(p\mu(d_i))$, le nombre de cas attendus au sein du site d_i est donc

$$\mathbb{E}(X_{d_i}) = p\mu(d_i).$$

Considérons désormais un facteur de confusion catégoriel à k modalités. Le nombre de cas attendus est calculé au moyen d'une standardisation indirecte et a pour expression

$$\mathbb{E}(X_{d_i}) = \sum_{j=1}^{k} \mathbb{E}(X_{d_i}^j) = \sum_{j=1}^{k} p^j \mu^j(d_i),$$

où, pour chaque modalité j, $X_{d_i}^j$ est le nombre de cas observés sur le site d_i, $\mu^j(d_i)$ est la taille de la population du site d_i, p^j étant la prévalence de la modalité j et enfin $\mu^j(D)$ est la taille de la population de D présentant la modalité j.

5 Limites des statistiques de scan spatiales

5.1 Temps de calcul

L'algorithme utilisé dans la phase de détection ainsi que l'inférence statistique basée sur des simulations de Monte Carlo laisse à penser que le temps de calcul relatif à la méthode est directement lié au nombre d'évènements et/ou sites ainsi qu'au nombre de simulations de Monte Carlo. Aussi, dans le cas de grandes bases de données (ex : registres nationaux) les temps de calculs peuvent devenir très longs. De surcroît, lorsqu'est considéré un risque de première espèce très faible, la précision de la probabilité critique se doit d'être élevée afin de statuer sur le rejet (ou non) de l'hypothèse nulle, précision directement liée au nombre de simulations. Les sections suivantes vont s'attacher à étayer les arguments précités.

5.1.1 Complexité algorithmique

Le logiciel de référence SaTScan© [Kulldorff, 2011] est celui dans lequel la plupart des différents modèles de statistiques de scan spatiales et spatio-temporelles sont implémentées. L'algorithme 4 décrit celui implémenté dans le cadre des analyses spatiales.

Algorithme 4 - Statistiques de scan spatiales : Détection et Inférence.

Etape 1 : Considérer un point. Calculer la distance entre ce point et les autres. Trier les distances par ordre croissant. Sauvegarder les points triés selon leur distance par rapport au point considéré.
Etape 2 : Répéter l'étape 1 pour chaque point.
Etape 3 : Considérer un point.
Etape 4 : Créer un cercle centré en ce point et augmenter de manière continue le rayon. Pour chaque point qui rentre dans le cercle, mettre à jour le nombre de cas et la population dans la fenêtre de scan.
Etape 5 : Répéter les étapes 3 et 4 pour chaque point. Reporter la fenêtre qui maximise la fonction de vraisemblance $L(Z)$.
Etape 6 : Répéter les étapes 3 à 5 pour chaque réplication de Monte Carlo.

Notons N le nombre d'évènements dans l'aire géographique étudiée et R le nombre de simulations de Monte Carlo sous l'hypothèse nulle d'absence de clusters. Une étude de complexité

algorithmique montre que les étapes 1 et 2 présentent une complexité de l'ordre de $\mathcal{O}(N^2 \log N)$ et que les étapes 3 à 5 présentent une complexité de l'ordre de $\mathcal{O}(RN^2)$. Aussi, le temps de calcul devient très long lorsque N est important et/ou R est important. Or, dans le cas de grandes bases de données, comme par exemple des registres nationaux, le nombre d'évènements (ou encore le nombre de centres dans le cas des données agrégées) peut être important. Ceci a une influence importante sur le temps de calcul, surtout quand le nombre de simulations de Monte Carlo est important.

A titre d'exemple, nous avons étudié le temps de calcul du logiciel SatScan© en se basant sur des données simulées. Le modèle étudié est le modèle de Poisson spatial discret, appliqué à des données agrégées (caractérisées par leur centres et leur population sous-jacente). L'étude du temps de calcul s'est faite en modulant à la fois le nombre de centres et le nombre de simulations de Monte Carlo.

TABLE 3.1 – Temps de calcul du logiciel SaTScan© en fonction du nombre de centres et du nombre de simulations de Monte Carlo - *CPU : 2.2Ghz Intel Core I7 - RAM : 8 Go 1333 Mhz DDR3*

	R			
Nombre de centres	999	9999	99999	999999
600	4s	4s	1min23s	19min20s
4055	16s	2min48s	30min17s	5h31min23s
6000	43s	6min52s	1h9min3s	9h47min36s
36700	1h21min47s	12h57min22s	$> 1j$	$> 1j$

A la suite de cette étude, il apparaît que le temps de calcul devient extrêmement important lorsque le nombre de simulations de Monte Carlo est grand et/ou que le le nombre de centres est conséquent (Tableau 3.1). En effet, 5h31min23s sont nécessaires pour réaliser une inférence statistique si la base de données contient l'ensemble des cantons français (4055) et le nombre de réplications de MC est égal à 999999. Le temps de calcul dépasse grandement la journée si la base de données contient l'ensemble des communes françaises (36700) et que le nombre de simulations est supérieur ou égal à 99999.

5.1.2 Probabilité critique associée au test de Monte Carlo, $\mathbb{P}_{mc}$

Valeur minimale de $\mathbb{P}_{mc}$. Par construction, la probabilité critique associée au test basé sur les simulations de Monte Carlo, $\mathbb{P}_{mc}$ définie en (3.1), présente une valeur minimale qui est directement liée au nombre de simulations :

$$\mathbb{P}_{mc} \geq \frac{1}{1+R}.$$

A titre d'exemple, si $R = 999$ alors $\mathbb{P}_{mc} \geq 0.001$ et si $R = 9999$ alors $\mathbb{P}_{mc} \geq 0.0001$. Aussi, afin d'obtenir une décimale supplémentaire, il faut multiplier par 10 le nombre de simulations.

Précision de $\mathbb{P}_{mc}$. C'est une notion importante qui intervient dans la capacité à rejeter l'hypothèse nulle. En effet, pour des risques de première espèce faibles, une précision importante de la probabilité critique est requise afin de statuer sur le rejet de $\mathcal{H}_0$. En se basant sur la formule d'un intervalle de confiance d'une proportion [3] π au niveau de confiance 95% sur un échantillon de taille R :

$$\mathrm{IC}_{\pi}^{95\%} = \left[\mathbb{P}_{mc} \pm z_{\alpha/2} \sqrt{\frac{\mathbb{P}_{mc}(1-\mathbb{P}_{mc})}{R}} \right],$$

3. Probabilité critique réelle dans la situation où la distribution, sous $\mathcal{H}_0$, de la statistique de test est connue.

on peut déduire la valeur minimum de R pour une précision δ :

$$R \geq \frac{\mathbb{P}_{mc}(1-\mathbb{P}_{mc})}{\left(\delta/z_{\alpha/2}\right)^2}.$$

La valeur du produit $\mathbb{P}_{mc}(1-\mathbb{P}_{mc})$ est maximale lorsque $\mathbb{P}_{mc}=0.5$. Aussi, pour une précision à la deuxième décimale ($\delta=0.01$), le nombre minimum de simulations est égal à 10^4. Or, en considérant un risque de première espèce $\alpha=0.05$ associé au test de Monte Carlo, une précision de $\mathbb{P}_{mc}$ à la deuxième décimale n'est pas suffisante. Dans ce cas de figure, une précision à au moins 3 décimales est requise. Suivant le même calcul, le nombre de simulations minimum pour obtenir $\delta=0.001$ est égal à 10^6.

5.1.3 Nombre de simulations et puissance du test

Soit U une statistique de test. Dans le cas d'un test classique, la loi de U est connue et par conséquent, pour un risque de première espèce α fixé, la région critique (rejet de $\mathcal{H}_0$) est bien définie. En revanche, dans le cadre d'un test de Monte Carlo, la loi de U n'est pas définie. Aussi, pour un risque α fixé, la région critique devient variable et est fonction des simulations sous $\mathcal{H}_0$. Cette variabilité de la région critique peut induire une perte de puissance statistique. Cependant, l'augmentation du nombre de simulations a pour effet de réduire la variabilité associée à la région critique et par conséquent rend négligeable l'effet de cette variabilité sur la perte de puissance [Hope, 1968]. La question fondamentale réside donc dans la détermination du nombre de simulations R nécessaires afin de rendre négligeable la variabilité de la région critique. Dans [Marriott, 1979], les auteurs se sont intéressés à l'effet du nombre de simulations sur la probabilité de rejeter l'hypothèse nulle en utilisant un test de Monte Carlo.

Posons u_0 la statistique de test observée sur les données et $U=\{u_1,u_2,...,u_R\}$ la distribution de U sous $\mathcal{H}_0$ basée sur les simulations de Monte Carlo. Posons $m=\alpha R$, le rang minimal que u_0 doit avoir pour rejeter $\mathcal{H}_0$. En d'autres termes, si u_0 se trouve dans les m plus grandes valeurs de U, alors $\mathcal{H}_0$ est rejetée. Posons $\pi=P(U>u_0)$. Dans le cadre d'un test classique au risque α, $P(U>u_0)<\alpha$ conduit au rejet de $\mathcal{H}_0$. Nous appellerons π, la probabilité critique théorique. Dans le cadre d'un test de Monte Carlo, chaque statistique de test simulée a une probabilité π d'être supérieure à u_0. Or, comme les u_i sont *i.i.d.*, nous pouvons poser $Y\sim\mathcal{B}(R,\pi)$, la variable aléatoire qui associe à un ensemble de R simulations le nombre de fois où les statistiques de test résultant des simulations ont été supérieures à u_0. La probabilité de rejeter $\mathcal{H}_0$, notée $\mathbb{P}_{rejet\ \mathcal{H}_0}$ est définie par :

$$\mathbb{P}_{rejet\ \mathcal{H}_0}=\mathbb{P}(Y\leq m-1)=\sum_{i=0}^{m-1}\binom{R}{i}\pi^i(1-\pi)^{R-i}.$$

A partir de cette expression, il est aisé d'évaluer $\mathbb{P}_{rejet\ \mathcal{H}_0}$ en fonction des valeurs de π et de α. Les tables 3.2 et 3.3 montrent les différentes valeurs de $\mathbb{P}_{rejet\ \mathcal{H}_0}$ en fonction de π pour des risques α respectivement égaux à 0.05 et 0.01. On remarque que lorsque la probabilité critique théorique π est inférieure à α, $\mathbb{P}_{rejet\ \mathcal{H}_0}$ tend vers 1 lorsque le nombre de simulations augmente. *A contrario*, lorsque π est inférieure à α, $\mathbb{P}_{rejet\ \mathcal{H}_0}$ tend vers 0 lorsque le nombre de simulations augmente. Dans le cas où $\pi=\alpha$, $\mathbb{P}_{rejet\ \mathcal{H}_0}$ tend vers $1/2$ lorsque le nombre de simulations augmente.

On peut en conclure que des faibles valeurs de α nécessitent un nombre de simulations plus important. Dans [Besag and Diggle, 1977], les auteurs suggèrent que $m=5$ peut être une valeur appropriée pour la plupart des utilisations du test de Monte Carlo. Cependant cette étude ne considère que la probabilité de rejeter $\mathcal{H}_0$ en utilisant un test de Monte Carlo. Dans [Jockel, 1986], les auteurs sont allés plus loin en investiguant la perte de de puissance d'un test de Monte Carlo par rapport à un test uniformément le plus puissant (UPP). Posons $\beta(\alpha)$ la puissance d'un test UPP en fonction d'un risque de première espèce α. Posons $\beta_R(\alpha)$ la puissance d'un test de Monte Carlo basé sur R simulations. Dans [Dwass, 1957], les auteurs ont défini l'efficacité, en termes de puissance, d'un test de Monte Carlo par rapport à un test UPP par le rapport suivant :

$$\frac{\beta_R(\alpha)}{\beta(\alpha)}.$$

TABLE 3.2 – Valeur de $\mathbb{P}_{rejet\ \mathcal{H}_0}$ en fonction de π pour $\alpha = 0.05$

		$\alpha = 0.05$				
		π				
m	R	0.1	0.075	0.05	0.025	0.01
1	19	0.135	0.227	0.377	0.618	0.826
2	39	0.088	0.199	0.413	0.745	0.942
5	99	0.025	0.128	0.445	0.897	0.997
25	499	0.000	0.011	0.475	0.999	1
50	999	0.000	0.001	0.483	0.999	1

TABLE 3.3 – Valeur de $\mathbb{P}_{rejet\ \mathcal{H}_0}$ en fonction de π pour $\alpha = 0.01$

		$\alpha = 0.01$				
		p				
m	R	0.02	0.015	0.01	0.005	0.001
1	99	0.135	0.224	0.370	0.609	0.906
2	199	0.091	0.199	0.407	0.738	0.983
5	499	0.028	0.131	0.441	0.892	1
25	2499	0.000	0.012	0.474	0.999	1
50	4999	0.000	0.001	0.481	0.999	1

Dans [Jockel, 1986], les auteurs ont défini une borne inférieure de $\beta_R(\alpha)/\beta(\alpha)$:

$$e^D_{R,\alpha} = 1 - \frac{E|Z_{R,\alpha} - \alpha|}{2\alpha} \leq \frac{\beta_R(\alpha)}{\beta(\alpha)},$$

où

$$\frac{E|Z_{R,\alpha} - \alpha|}{2\alpha} = \frac{\left(\alpha^\alpha (1-\alpha)^{1-\alpha}\right)^{R+1}}{(R+1)\alpha B((R+1)\alpha, (R+1)(1-\alpha))}.$$

B désigne la densité de probabilité de la loi Beta et $e^D_{R,\alpha}$ est *l'efficacité de Dwass* pour un échantillon simulé de taille R et un risque α fixé.

A partir de cette formule, il est possible d'évaluer l'efficacité d'un test de Monte Carlo par rapport à un test UPP en fonction du risque α et du nombre R de simulations. La table 3.4 montre clairement que plus les valeurs du risque α sont faibles, plus le nombre de simulations doit être important pour assurer une efficacité correcte du test de Monte Carlo.

Dans [Jockel, 1986], les auteurs ont également fourni une approximation du nombre de simulations minimum afin d'obtenir une efficacité fixée [4] :

$$R \approx \frac{1-\alpha}{2\pi\alpha(1-e^D)^2} \tag{3.3}$$

A titre d'exemple, considérons un système de surveillance prospective qui réalise une détection journalière de cluster de cas de maladie. Un risque $\alpha = 0.05$ conduirait en moyenne à un faux positif tous les 20 jours. Dans le cadre de déclenchement d'alarme de santé, ce nombre moyen annuel de faux positifs est tout à fait inacceptable. Aussi, il est nécessaire de fixer un risque de première espèce beaucoup plus faible afin de se prémunir des faux positifs. Considérons que nous

4. Dans cette formule, π désigne le nombre pi.

TABLE 3.4 – Valeurs de $e^D_{R,\alpha}$ pour différentes valeurs de α et de R

	α				
R	0.0001	0.001	0.01	0.05	0.1
19	–	–	–	0.64	0.74
99	–	–	0.63	0.83	0.88
999	–	0.63	0.76	0.95	0.96
9999	0.63	0.88	0.96	1	1
99999	0.88	1	1	1	1

voulons minimiser le nombre moyen de faux positifs à un tous les dix ans. Le risque α associé est donc :

$$\alpha = 1/3650 \approx 3.10^{-4}$$

Dans ce cas de figure, le nombre de simulations nécessaires pour assurer une efficacité de 90% à ce risque α est donnée par 3.3 :

$$\frac{1 - 3.10^{-4}}{2\pi 3.10^{-4}(1-0.9)^2} \approx 60\,000 \text{ simulations.}$$

5.1.4 Conclusion

Les sections précédentes ont démontré que le temps de calcul inhérent aux statistiques de scan spatiales est directement lié au nombre d'évènements et/ou sites, N, ainsi qu'au nombre de simulations de Monte Carlo, R. De surcroît, la méthode d'inférence statistique fait intervenir l'algorithme de détection du MLC à chaque simulation, présentant une complexité en $\mathcal{O}(RN^2)$. Aussi, l'inférence statistique se positionne comme le principal facteur de l'augmentation du temps de calcul.

Afin de pallier ce problème, deux champs de réflexion s'offrent à nous. Le premier consiste à travailler sur l'optimisation de l'algorithme de détection du MLC. Cependant, celui implémenté dans le logiciel SaTScan© a été modifié plusieurs fois suite à des travaux de [Kulldorff, 1999, Section 14.3] et [Walther, 2010], la version actuelle étant relativement optimisée. Le deuxième champ de réflexion se base sur le fait trouver une alternative à la méthode basée sur les simulations de Monte Carlo. Dans [Besag and Clifford, 1991], les auteurs ont proposé une méthode de Monte Carlo séquentielle qui, appliquée aux statistiques de scan spatiales, consiste à stopper la procédure de simulation lorsqu'un nombre fixé h de simulations présentent un rapport de vraisemblance supérieur à celui calculé sur les données observées. Dans [Silva et al., 2009], les auteurs ont démontré que si $h = \alpha R - 1$ alors il n'y a pas de perte de puissance par rapport à une procédure de Monte Carlo classique. Cette méthode alternative présente un intérêt uniquement si le cluster considéré est très loin d'être significatif, à l'inverse le nombre de simulations sera identique à la procédure classique. Par ailleurs, compte tenu du fait que la précision de la probabilité critique est directement liée au nombre de simulations de Monte Carlo, [Abrams et al., 2010] ont proposé une méthode basée sur l'approximation de la distribution du rapport de vraisemblance par une loi de Gumbel. Le principe consiste à réaliser un nombre limité de simulations de Monte Carlo (minimum 999 selon les auteurs) et d'ajuster une loi de Gumbel, par la méthode des moments, à cette distribution empirique. Par conséquent, la probabilité critique associée au cluster peut être estimée *via* la loi de Gumbel. Cette méthode présente l'intérêt de donner une bonne précision à la probabilité critique mais nécessite tout de même l'utilisation de simulations de Monte Carlo, qui, dans le cas de grandes bases de données, peuvent s'avérer très consommatrices en termes de temps de calcul.

5.2 Forme de la fenêtre

Une des principales caractéristiques des statistiques de scan spatiales réside dans la forme circulaire de la fenêtre de scan. Cette forme a été choisie pour des raisons calculatoires car le fait qu'un cercle ne soit défini par par son centre et son rayon a pour finalité de limiter le cardinal de $\mathcal{Z}$, l'ensemble des clusters potentiels. [Kulldorff et al., 2003] ont montré que les statistiques de scan spatiales s'avèrent être très puissantes lorsque le "vrai" cluster est de forme circulaire. En revanche, la méthode démontre une faible puissance pour détecter des clusters de formes arbitraires. A titre d'exemple, la méthode sera peu capable de détecter un cluster de cas de maladie le long d'une rivière.

Fort de ce constat, la méthode a été généralisée pour la détection de clusters de formes elliptiques [Kulldorff et al., 2006]. Au sein de cette version, l'ensemble des clusters potentiels $\mathcal{Z}$ n'est désormais plus défini uniquement par un ensemble de centres et de rayons, mais aussi par l'excentricité [5] (la forme) et l'angle de l'ellipse par rapport à l'axe des abscisses. Cette nouvelle méthode présente une puissance légèrement supérieure à la méthode classique lorsque le "vrai" cluster a une forme d'ellipse allongée. Les deux méthodes présentent des puissances équivalentes lorsque l'excentricité de l'ellipse tend vers 1. La prise en compte de paramètres supplémentaires pour la définition de $\mathcal{Z}$ a pour conséquence une importante augmentation de son cardinal et, par conséquent, l'augmentation du temps de calcul en phase détection mais également en phase d'inférence statistique car cette méthode utilise également une procédure basée sur des simulations de Monte Carlo. Bien que cette méthode propose une meilleure flexibilité que la méthode classique, elle impose une forme paramétrique au clusters potentiels. Aussi, les auteurs préconisent d'utiliser des versions non paramétriques pour détecter des clusters de formes très irrégulières, décrites en section 3.3. En outre, ces méthodes utilisent, toutes sans exception, une procédure d'inférence basée sur des simulations de Monte Carlo.

6 Une alternative à la méthode de Monte Carlo pour le TRVG

Dans les sections précédentes, nous avons démontré que la précision de la probabilité critique associée au MLC est fonction du nombre de simulations de Monte Carlo. Aussi, lorsque la situation exige une précision importante, un grand nombre de simulation est nécessaire, augmentant de manière significative les temps de calcul. Nous proposons, dans cette section, une méthodologie appliquée aux statistiques de scan spatiales qui fait office d'alternative à la procédure de test basée sur les simulations de Monte Carlo pour le TRVG. Tout d'abord nous expliciterons le principe de la méthode, puis nous l'appliquerons sur des données simulées et enfin nous discuterons les résultats.

6.1 Méthodologie

Rappelons brièvement que la méthode de statistiques de scan spatiales se décompose en deux phases. La première consiste à détecter le MLC qui maximise le rapport de vraisemblance et, la deuxième permet de tester la significativité de ce dernier au moyen de simulations de Monte Carlo.

Notre méthodologie se base sur la conservation de la phase de détection qui conduit à la mise en évidence du MLC et d'un certain nombre d'informations associées telles que n_Z, le nombre de cas dans le cluster, $\mu(Z)$ la population du cluster et $\mu(D)$ la population totale de la surface étudiée D. A partir de ces informations, nous proposons deux modélisations permettant de tester la significativité du MLC.

Modélisation 1. Il est possible de modéliser les observations par $\{X_1, \ldots, X_{\mu(D)}\}$, une suite de variables aléatoires i.i.d. selon une loi de Bernoulli $\mathcal{B}(1, p)$. Aussi chaque observation de la population complète $\mu(D)$ est modélisée par une v.a. de Bernoulli ayant une probabilité p d'être un cas. Cette première modélisation, nous permet de calculer la probabilité qu'une statistique de

5. L'excentricité est définie par le rapport de la longueur du demi grand axe sur celle du demi petit axe. Aussi, le cercle est un cas particulier d'ellipse, d'excentricité égale à 1.

scan unidimensionnelle de fenêtre de taille $\mu(Z)$ sur la suite $\{X_1, \ldots, X_{\mu(D)}\}$ soit supérieure à n_Z cas :

$$\mathbb{P}(S_{\mu(Z)} > n_Z) = \mathbb{P}\left(S(\mu(Z), \mu(D)) > n_Z\right). \tag{3.4}$$

Modélisation 2. Soit une région carrée $[0, \sqrt{\mu(D)}] \times [0, \sqrt{\mu(D)}]$. Considérons $\{X_{ij}\}$, $1 \leq i, j \leq \sqrt{\mu(D)}$ un ensemble de variables aléatoires i.i.d. selon une loi de Bernoulli $\mathcal{B}(1, p)$. Chaque variable aléatoire X_{ij} désigne une observation de $\mu(D)$ qui a une probabilité p d'être un cas. Cette deuxième modélisation, évoquée dans [Glaz et al., 2001, Chapitre 5], nous permet de calculer la probabilité qu'une statistique de scan bidimensionnelle de fenêtre de taille $\sqrt{\mu(D)} \times \sqrt{\mu(D)}$ sur la région $[0, \sqrt{\mu(D)}] \times [0, \sqrt{\mu(D)}]$ soit strictement supérieure à n_Z :

$$\mathbb{P}(S_{\sqrt{\mu(Z)}, \sqrt{\mu(Z)}} > n_Z) = \mathbb{P}\left(S\left(\sqrt{\mu(Z)}, \sqrt{\mu(Z)}, \sqrt{\mu(D)}, \sqrt{\mu(D)}\right) \leq n_Z\right). \tag{3.5}$$

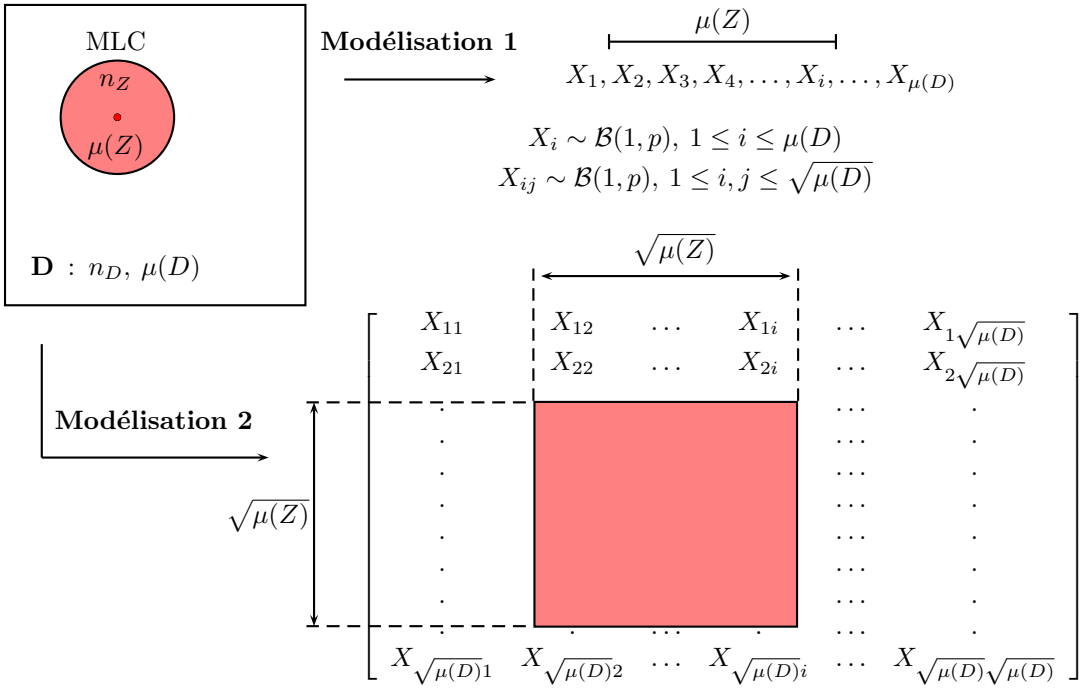


FIGURE 3.6 – Deux modélisations possibles.

Dans le cadre de la modélisation 1, le calcul de $\mathbb{P}(S_{\mu(Z)} > n_Z)$ peut être réalisé en utilisant l'approximation d'Haiman définie en (2.6) :

$$\mathbb{P}(S_{\mu(Z)} \leq n_z) \approx \frac{2Q_2 - Q_3}{(1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2)^{\frac{\mu(D)}{\mu(Z)} - 1}}$$

Cette approximation fait intervenir les quantités Q_2 et Q_3 définies en (2.1). Le principal avantage de cette modélisation réside dans le fait que des formules exactes, dans le cadre d'un modèle de Bernoulli unidimensionnel, existent pour les quantités Q_2 et Q_3, réduisant de manière drastique le temps de calcul.

En utilisant la modélisation 2, le calcul de (3.4) peut être réalisé en utilisant l'approximation d'Haiman définie en (2.16) faisant intervenir les quantités Q_2 et Q_3 définies en (2.15) et respectivement approximées par (2.19) et (2.20). Ces deux dernières approximations font intervenir les quantités Q_{ij}, $i, j \in \{2, 3\}$ définies en (2.18) et évaluées par le biais de simulations de Monte Carlo. Or, le principal objectif étant de réduire le temps de calcul, nous n'avons pas conservé cette

modélisation, trop gourmande en termes de temps de calcul.

L'objectif de cette étude est, au sein de la méthode de statistiques de scan spatiales, de conserver la phase de détection du MLC et de fournir une alternative à la méthode de Monte Carlo pour le TRVG en utilisant la modélisation 1. Celle-ci présente, d'une part, l'avantage de réduire de manière importante le temps de calcul et, d'autre part, fournit une erreur d'approximation de la distribution, ce qui n'est pas le cas dans la procédure de test basée sur les simulations de Monte Carlo. Cette alternative sera étudiée par le biais d'une étude de simulation décrite dans la Section 6.2, les résultats étant présentés en Section 6.3.

6.2 Etude de simulation

6.2.1 Données de simulation

Les données simulées sont issues d'un ensemble de données de *benchmark* développé par [Kulldorff et al., 2003] qui fait office de référence dans l'évaluation de méthodes de détection de clusters spatiaux [Kulldorff, 2006, Duczmal et al., 2006, Lawson, 2006, Duczmal et al., 2008]. Ces données de *benchmark* représentent des évènements (ex : cas de maladie) répartis sur 245 comtés du Nord-Est des Etats-Unis (Figure 3.7). La proportion d'évènements, p_0, est égale à 2.10^{-4}. Les données sont agrégées à l'échelle du comté, caractérisé par son centre, noté d_i, $1 \leq i \leq 245$, en termes de latitude et de longitude. A chaque comté est associée sa population sous-jacente, notée $\mu(d_i)$, issue du recensement américain de 1990. Posons D l'aire géographique étudiée. La population totale, $\mu(D)$, est égale à 29 535 210 habitants. La simulation des évènements a été réalisée à la fois



FIGURE 3.7 – Données simulées sur 245 comtés du Nord-Est des Etats-Unis.

sous l'hypothèse nulle, $\mathcal{H}_0$, d'absence de cluster et sous une hypothèse alternative, $\mathcal{H}_1$, supportant l'existence d'un cluster d'évènements.

Simulations sous $\mathcal{H}_0$. Posons X_{d_i} la variable aléatoire qui associe au canton d_i, $1 \leq i \leq 245$ le nombre d'évènements observés. Sous $\mathcal{H}_0$, X_{d_i} est distribuée selon une loi de Poisson $\mathcal{P}(p_0\mu(d_i))$. Aussi, une simulation des données sous l'hypothèse nulle consiste à simuler, dans chaque canton i, une loi de Poisson $\mathcal{P}(p_0\mu(d_i))$. Le *benchmark* est constitué de 999 999 simulations des données sous $\mathcal{H}_0$.

Simulations sous $\mathcal{H}_1$. L'objectif est de simuler la présence d'un cluster, noté $\mathcal{C}$, dans les données. Trois types de clusters peuvent être distingués : un cluster en zone rurale de faible population (Grand Isle) constitué de 3 comtés, un cluster en zone mixte (zone urbaine entourée de zones rurales) de population moyenne (Pittsburgh), constitué de 2 comtés, et un cluster en zone urbaine (Manhattan) de population élevée constitué de 2 comtés. Posons $X_{\mathcal{C}}$ la variable aléatoire qui associe au cluster $\mathcal{C}$ son nombre d'évènements et $\mu(\mathcal{C}) = \sum_{d_i \in \mathcal{C}} \mu(d_i)$ la population sous-jacente de $\mathcal{C}$.

La simulation de $X_{\mathcal{C}}$ se base sur un test de Poisson unilatéral dont les hypothèses sont :

$$\begin{cases} \mathcal{H}_0 \ : \ X_{\mathcal{C}} \sim \mathcal{P}(p_0\mu(\mathcal{C})) \\ \mathcal{H}_1 \ : \ X_{\mathcal{C}} \sim \mathcal{P}(p_1\mu(\mathcal{C})), p_1 > p_0. \end{cases}$$

La valeur de p_1 est déterminée telle qu'un test binomial unilatéral rejette l'hypothèse nulle au profit de l'hypothèse alternative avec une probabilité égale à 0.999, en considérant un risque de première espèce $\alpha = 0.05$. La détermination de p_1 est réalisée selon les deux étapes suivantes :

1. Sous $\mathcal{H}_0$, $X_{\mathcal{C}} \sim \mathcal{P}(p_0\mu(\mathcal{C}))$. Déterminer $k \in \mathbb{N}$ tel que $\mathbb{P}(X_{\mathcal{C}} \leq k) = 0.95$.
2. k déterminé, on se place sous $\mathcal{H}_1$ ($X_{\mathcal{C}} \sim \mathcal{P}(p_1\mu(\mathcal{C}))$) et on détermine p_1 telle que $\mathbb{P}(X_{\mathcal{C}} > k) = 0.999$ en utilisant de manière incrémentale la fonction de répartition de la loi de Poisson.

Nous pouvons remarquer que 0.999 correspond à la puissance maximale atteignable par le test de Poisson unilatéral précité. En dehors de $\mathcal{C}$, la simulation des X_{d_i} consiste à simuler une loi de Poisson $\mathcal{P}(p_0\mu(d_i)$. Par ailleurs, le *benchmark* comprend 10 000 simulations des données sous $\mathcal{H}_1$, pour chaque type de cluster (rural, mixte, urbain).

6.2.2 Méthodologie de calcul des probabilités critiques

Compte tenu du type de données simulées, nous avons utilisé des statistiques de scan spatiales basées sur un modèle de Poisson, avec une fenêtre circulaire de taille variable. L'objectif de l'étude est de comparer, pour un MLC détecté, sa probabilité critique basée sur les simulations de Monte Carlo à la probabilité basée sur l'approximation d'Haiman définie en (3.4). La méthodologie de calcul des probabilités critiques se base à la fois sur les données simulées sous $\mathcal{H}_0$ et sous $\mathcal{H}_1$.

Données simulées sous $\mathcal{H}_0$. L'objectif est de donner une approximation de la distribution du rapport de vraisemblance λ en utilisant les 999 999 jeux de données simulés sous $\mathcal{H}_0$. Aussi, pour chaque jeu de données, nous avons utilisé la phase de détection afin de mettre en évidence un MLC et calculer le rapport de vraisemblance associé. En définitive, nous avons obtenu 999 999 valeurs de λ ce qui nous permet d'obtenir une distribution approchée de λ sous $\mathcal{H}_0$.

Données simulées sous $\mathcal{H}_1$. Pour chaque jeu de données, la phase détection fournit un MLC, sa population $\mu(Z)$, son nombre d'évènements observés n_Z ainsi que son rapport de vraisemblance associé, λ_{obs}. Dans un premier temps, la probabilité critique associée au MLC, $\mathbb{P}(\lambda > \lambda_{obs})$, a été calculée en utilisant la distribution de λ précédemment calculée *via* les simulations de Monte Carlo. Dans un deuxième temps, grâce au informations relatives au clusters (population et nombre d'évènements) la probabilité (3.4) a été calculée en utilisant l'approximation d'Haiman définie en (2.6). Cette dernière impose la condition d'application suivante : $1 - Q_2 \leq 0.025$ qui est vérifiée. Si cette dernière n'est pas respectée alors la probabilité n'est pas calculée et la simulation n'est pas comptabilisée.

6.2.3 Evaluation des deux méthodes

Nous avons évalué la concordance entre les probabilités issues des simulations de Monte Carlo et celle issues de la modélisation 1 par le biais du calcul d'un coefficient de corrélation intra-classe (CCI). Par ailleurs, nous avons évalué la puissance, $1-\beta$, des deux méthodes en prenant en compte un risque de première espèce $\alpha = 0.05, 0.01$ et 0.001. Le calcul de la puissance a été réalisé de la manière suivante

$$1-\beta = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{\mathbb{P}<\alpha\}},$$

avec $\mathbb{P} = \mathbb{P}(\lambda > \lambda_{obs})$ pour la méthode basée sur les simulations de Monte Carlo et $\mathbb{P} = \mathbb{P}(S_{\mu(Z)} > n_Z)$ pour la méthode basée sur la modélisation 1. N correspond au nombre de simulations sous $\mathcal{H}_1$ qui ont été comptabilisées, c'est-à-dire, les simulations pour lesquelles l'approximation d'Haiman a pu être appliquée.

6.3 Résultats

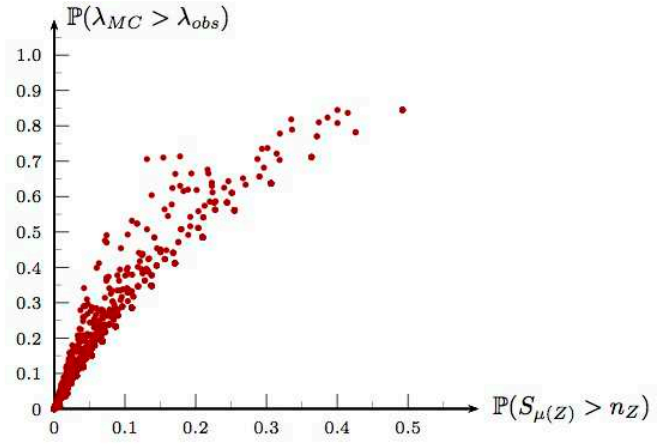
Le nombre de simulations sous $\mathcal{H}_1$ comptabilisées est de 9816 pour celles comprenant un cluster urbain, 9921 un cluster mixte et 9998 un cluster rural.

La concordance entre les deux méthodes est acceptable pour les simulations comprenant un cluster urbain (CCI = 0.638, IC 95% = [0.626 - 0.649], $p < 10^{-4}$) et un cluster mixte (CCI = 0.635, IC 95% = [0.623 - 0.647], $p < 10^{-4}$). En ce qui concerne les simulations avec cluster rural, la concordance entre les deux méthodes est excellente (CCI = 0.895, IC 95% = [0.893 - 0.901], $p < 10^{-4}$).

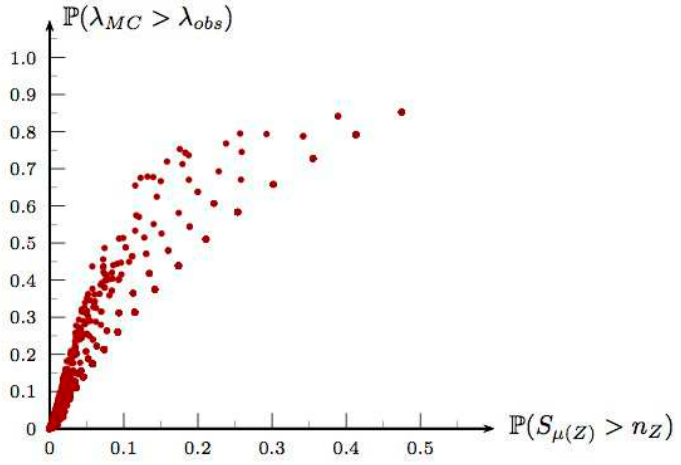
La Figure 3.8 présente le croisement des probabilités issues de simulations de Monte Carlo avec les probabilités issues de la modélisation 1, pour chaque type de cluster (urbain (a), mixte (b), rural(c)). Dans le cas des clusters urbains et mixte, nous pouvons remarquer que les probabilités issues de la modélisation 1 ont tendance à être inférieures à celles issues des simulations de Monte Carlo, ce qui justifie les valeurs des CCI associés. En ce qui concerne les simulations comprenant un cluster rural, les probabilités issues des deux méthodes sont similaires et très proches de 0, justifiant également la valeur du CCI associé.

La Table 3.5 présentent l'estimation de la puissance des deux méthodes avec différentes valeurs du risque de première espèce ($\alpha = 0.05$, 0.01 et 0.001). En ce qui concerne les simulations comprenant un cluster urbain ou mixte, la méthode basée sur la modélisation 1 présente une puissance systématique supérieure, quelle que soit la valeur de α. Pour les simulations avec cluster rural, les puissances des deux méthodes sont quasiment identiques.

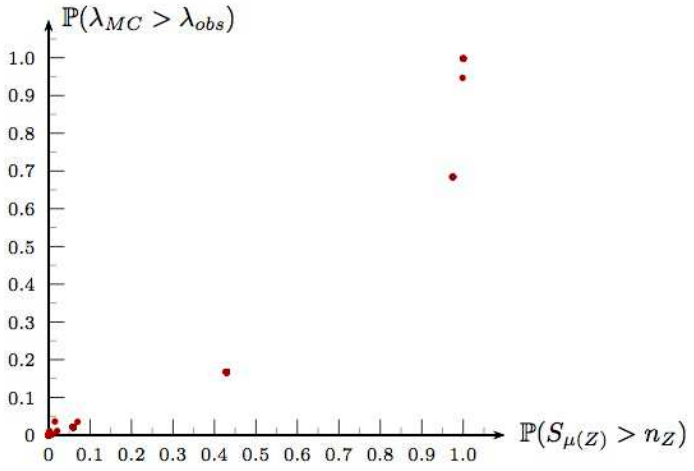
En utilisant un ordinateur de puissance normale, le temps de calcul moyen de $\mathbb{P}(S_{\mu(Z)} > n_Z)$ par simulation est de 7.3 ms. Par le biais du logiciel SaTScan© développé par Martin Kulldorff, le temps de calcul nécessaire pour réaliser 999 999 simulations de Monte Carlo sous l'hypothèse nulle est d'environ 36 min.



(a) Cluster urbain



(b) Cluster mixte



(c) Cluster rural

FIGURE 3.8 – Croisement des probabilités issues de la méthode de Monte Carlo et celle issues de la modélisation 1.

TABLE 3.5 – Comparaison de puissance entre l'approximation basée sur les simulations de Monte Carlo et l'approximation d'Haiman.

	α					
	0.05		0.01		0.001	
Type de cluster	MC	App. (3.4)	MC	App. (3.4)	MC	App. (3.4)
Urbain	0.8527883	0.9376083	0.7449281	0.8366806	0.5305332	0.6425731
Mixte	0.8893261	0.9516747	0.7884383	0.8692494	0.6023002	0.7041969
Rural	0.9898981	0.9737948	0.9734947	0.9735947	0.9432887	0.9432887

7 Conclusion

Dans le domaine des statistiques de scan, les statistiques de scan spatiales développées par Martin Kulldorff sont aujourd'hui, de loin, les méthodes les plus utilisées dans la détection de clusters spatiaux.

Dans un premier temps, ce chapitre s'est attaché à définir les différents types de données spatiales, les statistiques de scan spatiales ainsi que les principaux modèles (Bernoulli et Poisson).

Dans un deuxième temps, ce chapitre a décrit les deux principales limites de ces méthodes : la forme de la fenêtre de scan et le temps de calcul. En ce qui concerne la forme de la fenêtre, les statistiques de scan spatiales considèrent uniquement une fenêtre de forme circulaire ce qui implique une faible puissance de détection de clusters de forme irrégulière. Le temps de calcul de ces méthodes est fonction à la fois du nombre de simulations de Monte Carlo utilisées pour estimer la distribution de la statistique de test sous l'hypothèse nulle et du nombre d'évènements / sites observés. De surcroît, la précision de la probabilité critique et la puissance du test basé sur des simulations de Monte Carlo est fonction du nombre de simulations. Dans le cas de grandes bases de données et/ou lorsque que l'analyste désire une précision élevée de la probabilité critique, les temps de calculs inhérents aux simulations de Monte Carlo deviennent très longs, voire excessifs.

Dans un troisième temps, nous avons proposé, dans le cadre des statistiques de scan spatiales, une alternative à la méthode de Monte Carlo pour le TRVG. Le principe de cette alternative réside dans le fait de conserver la phase de détection du MLC et, en utilisant une modélisation basée sur une suite variables aléatoires de Bernoulli, d'estimer la probabilité associée au MLC. L'approximation de cette probabilité a été réalisée par la formule proposée dans [Haiman, 2007] basée sur les suites de variables aléatoires 1-dépendantes. Cette approximation, qui donne également une erreur d'approximation, présente l'avantage d'être constituée de quantités (Q_2 et Q_3) qui sont calculables par le biais de formules exactes, diminuant ainsi les temps de calculs. Nous avons réalisé une étude de simulation visant à comparer la méthode basée sur les simulations de Monte Carlo et celle basée sur notre modélisation. A partir de simulations sous l'hypothèse alternative, nous avons mis en évidence que les deux méthodes fournissent des probabilités concordantes, notre méthode présentant une puissance supérieure à celle basée sur les simulations de Monte Carlo et un temps de calcul nettement plus faible.

Chapitre 4

Application à l'étude de la maladie de Crohn

Sommaire

1 Introduction 93
2 Matériels et méthodes 94
2.1 Données 94
2.2 Méthodes statistiques 95
3 Résultats 96
3.1 Variations spatiales et spatio-temporelles de l'incidence de MC 96
3.2 Analyse de l'offre de soin en gastro-entérologie 99
4 Discussion 99
5 Conclusion 100

1 Introduction

Les Maladies Inflammatoires Chroniques de l'Intestin (MICI) comprennent la maladie de Crohn (MC) et la rectocolite hémorragique (RCH). Ce sont des inflammations chroniques du tube digestif atteignant exclusivement le rectum et le colon pour la RCH et l'ensemble du tube digestif avec une préférence pour la région iléo-caecale pour la MC. Bien que ces maladies soient non létales, mais en raison de leur survenue précoce dans la vie et de leur chronicité, elles induisent une morbidité élevée qui altère la qualité de vie des malades. En dépit de récentes avancées, notamment dans le domaine génétique, dans la compréhension de la physiopathologie de ces maladies, leur étiologie reste à ce jour inconnue. Cependant, les causes de nature environnementale ont toujours été suspectées de jouer un rôle critique dans l'expression des MICI [Baron et al., 2005, Cho, 2008, Cosnes, 2010, Duerr et al., 2006, Ng et al., 2012]. En effet, l'évolution de l'épidémiologie des MICI d'un point de vue temporel et spatial laisse à penser que les facteurs environnementaux jouent un rôle majeur en modifiant l'expression de la maladie. L'émergence de cette dernière dans des pays en développement indique que l'augmentation de l'incidence est liée à l'occidentalisation du mode de vie et l'industrialisation [Thia et al., 2008]. Cette constatation est retrouvée chez les populations migrantes de pays en développement vers les pays développés qui présentent un risque accru de MICI et, de ce fait, soutiennent l'hypothèse de l'importance de l'influence de l'environnement [Barreiro-de Acosta et al., 2011, Gearry et al., 2010]. L'urbanisation des sociétés ainsi que les changements socio-économiques peuvent se produire différemment dans différentes régions géographiques et différentes populations. Par conséquent, il est important de prendre en compte l'hétérogénéité des facteurs de risques applicables à un patient. Aussi, l'étude des variations géographiques de l'incidence des MICI fournit des indices aux chercheurs afin de mener des investigations sur les possibles facteurs étiologiques environnementaux.

Le nord de la France est caractérisé par une incidence élevée de MC qui a connu une recrudescence depuis les vingt dernières années, notamment chez les adolescents et les jeunes adultes

[Chouraki et al., 2011, Colombel et al., 1996]. [Declercq et al., 2010] ont montré, par le biais de modèles hiérarchiques bayésiens (modèle de Besag, York et Mollié), une hétérogénéité spatiale de l'incidence de MC avec une prédominance de la maladie dans les zones agricoles. Cependant, ces méthodes bayésiennes ne permettent pas de mettre en évidence des clusters atypiques et statistiquement significatifs de cas de maladie ainsi que leur évolution temporelle. Afin de palier les limites des méthodes précitées, les statistiques de scan spatio-temporelles ont été utilisées. Aussi, l'objectif de cette étude est d'analyser l'hétérogénéité spatiale de l'incidence de MC dans le nord de la France en mettant en évidence des clusters atypiques spatiaux et spatio-temporels grâce aux statistiques de scan.

2 Matériels et méthodes

2.1 Données

Notre étude a été réalisée dans le nord de la France qui présente une population totale de 5 790 526 habitants et est constitué de quatre départements : Nord, Pas-De-Calais, Somme et Seine Maritime (Figure 4.1). Au sein de cette zone géographique, les données ont été agrégées à l'échelle cantonale, menant à un nombre total de 273 cantons dont la population varie de 1 500 à 212 000 habitants. Les centres des cantons, définis par leurs latitude et longitude, ont été utilisés dans les analyses statistiques.



FIGURE 4.1 – Départements couverts par le registre EPIMAD.

Pour la période 1990-2006, l'estimation de la population annuelle moyenne des cantons a été réalisée sur la base des recensements de la population française de 1990 et de 1999. Pour la période de 1990 à 1999, les données de population ont été estimées par le biais de méthodes d'interpolation exponentielle utilisant les taux d'accroissement estimés par les recensements de 1990 et 1999. Les données de 2000 à 2006 ont été estimées par extrapolation exponentielle. Aussi, pour chaque canton, nous avons obtenu la population stratifiée par sexe et par classe d'âge (moins d'un an, 1 - 4 ans, 18 classes d'âge quinquennales de 5 à 89 ans et 90 ans et plus).

Les données épidémiologiques ont été extraites du registre EPIMAD, répertoriant de manière exhaustive l'ensemble des cas incidents de MC et de rectocolite hémorragique dans le nord de la France (Figure 4.1) depuis 1988. La méthodologie du registre EPIMAD a été décrite en détail dans [Gower-Rousseau et al., 1994]. Pour résumer, les enquêteurs du registre ont collecté des données sur tous les patients diagnostiqués par l'ensemble des gastro-entérologues du secteur privé ou public ($n = 262$). Seuls les patients résidant dans la zone couverte par le registre au moment de leur diagnostic ont été inclus et la localisation de chaque cas a été définie par le lieu de résidence du patient.

Chaque gastro-entérologue collecte l'ensemble des patients consultant pour la première fois et présentant des symptômes cliniques compatibles avec une maladie inflammatoire chronique de l'intestin. Parallèlement, ce gastro-entérologue est contacté au minimum trois fois par an par un enquêteur du registre et, à la suite, ce dernier se déplace sur le lieu de consultation afin de recueillir les données issues des dossiers patients par le bais d'un questionnaire standardisé. Le diagnostic final est réalisé par deux experts gastro-entérologues et est enregistré comme définitif, probable ou possible selon un critère défini dans [Gower-Rousseau et al., 1994]. Dans le cadre de cette étude, seuls les patients présentant un diagnostic définitif ou probable durant la période de 1990 à 2006 ont été inclus, soit au au total 6472 cas de MC.

2.2 Méthodes statistiques

Les rapports standardisés d'incidence (*Standardized Incidence Ratio* (SIR)) ont été calculés par le biais de la méthode utilisée dans [Declercq et al., 2010]. Cette méthode est basée sur un modèle hiérarchique bayésien développé par [Besag et al., 1991] qui fournit un lissage des SIR prenant en compte leur instabilité vis-à-vis des unités spatiales de faible population, l'hétérogénéité spatiale globale et l'autocorrélation spatiale. Cependant, bien que cette méthode permette de mettre en évidence une hétérogénéité spatiale globale, elle ne permet pas de détecter des zones géographiques atypiques en termes d'incidence de MC (clusters d'unités spatiales) et de tester leur significativité. Par ailleurs, le principal risque inhérent à cette méthode est de sélectionner une zone, qui semble atypique, et de l'utiliser pour d'autres comparaisons statistiques, conduisant à un biais de pré-sélection. De surcroît, ce modèle bayésien ne permet pas de prendre en compte la dimension temporelle et ainsi évaluer l'évolution dans le temps des zones géographiques précédemment détectées.

Les statistiques de scan spatiales et spatio-temporelles [Kulldorff, 1997, Kulldorff et al., 1998] ont été utilisées pour tester la présence de clusters de MC et identifier leur localisation sans biais de pré-sélection. Ces méthodes ont été ajustées, au moyen de standardisation indirecte, sur l'âge au moment du diagnostic et le sexe, qui sont des facteurs de confusion connus de MC [Cosnes et al., 2011]. La méthode de statistique de scan spatio-temporelle peut être décomposée en deux étapes : la détection et l'inférence.

Lors de la phase de détection, la méthode utilise une fenêtre de scan cylindrique dont la base représente la composante spatiale et la hauteur la composante temporelle. La fenêtre est de taille variable et se déplace sur l'ensemble de la zone géographique étudiée, utilisant comme centres les centres des cantons. A chaque position, la base de la fenêtre cylindrique varie de 0 à un rayon tel que 50 % du nombre de cas de MC soit compris dans cette base. La hauteur de la fenêtre varie également de zéro à la moitié de la période d'étude. De surcroît, la méthode permet la détection de clusters constants sur l'ensemble de la période étudiée, appelés clusters spatiaux constants, et de clusters évoluant lors de la période étudiée, appelés clusters spatiaux non-constants. La phase de détection conduit à un nombre élevé de clusters candidats, chacun contenant un ou plusieurs cantons. Pour chaque cluster candidat, et faisant l'hypothèse que le nombre de cas dans chaque canton est distribué selon une loi de Poisson, une fonction de vraisemblance est calculée. Le cluster candidat maximisant cette fonction est défini comme le cluster le plus probable (*Most Likely Cluster*(MLC)). La fonction de vraisemblance dépend du nombre de cas à l'intérieur du cluster candidat et à l'extérieur de celui-ci.

Lors de la phase d'inférence, l'hypothèse nulle correspond à l'absence de cluster qui est : "le risque de développer la MC est constant parmi tous les cantons et toute la période d'étude". Afin de tester cette hypothèse nulle, un test basé sur un rapport de vraisemblance (LLR) est utilisé : le numérateur correspond à la valeur de la fonction de vraisemblance associée au MLC et le dénominateur désigne la fonction de vraisemblance sous l'hypothèse nulle. La distribution du LLR n'ayant pas de forme analytique, la méthode utilise 9 999 réplications de Monte Carlo sous l'hypothèse nulle afin d'obtenir une probabilité critique [Dwass, 1957]. Dans cette procédure, une réplication consiste à générer les données distribuées selon une loi de Poisson sous l'hypothèse nulle. Les calculs ont été réalisés par le biais du logiciel SaTScan [Kulldorff, 2011] et un cluster a été considéré comme significatif si sa probabilité critique unilatérale associée était inférieure à 0.05. Par ailleurs, les clusters secondaires, clusters qui ne se superposent pas avec le MLC et disposent d'une valeur élevée de LLR, ont été pris en compte. Leur significativité a été évaluée en calculant

leur probabilité critique selon la méthode décrite dans [Zhang and Kulldorff, 2010].

Comme la présence de cluster spatiaux et spatio-temporels de MICI peut être imputable à une hétérogénéité de l'offre de soins en gastro-entérologie, nous avons utilisé les statistiques de scan afin de détecter la présence de clusters de gastro-entérologues. Des analyses temporelles, spatiales et spatio-temporelles ont été réalisées entre 1990 et 2006 en utilisant le code postal du lieu d'exercice des praticiens.

3 Résultats

3.1 Variations spatiales et spatio-temporelles de l'incidence de MC

De 1990 à 2006, le taux moyen d'incidence brute annuel de MC était de 6.5 pour 100 000 habitants dans l'ensemble de la région étudiée. La répartition spatiale des SIR lissés, ajustés sur le sexe et l'âge au diagnostic, déterminée par la méthode hiérarchique bayésienne, suggère fortement une hétérogénéité de l'incidence de MC au sein de la zone géographique étudiée (Figure 4.2), et notamment une zone de sous-incidence dans le département de la Seine-Maritime.

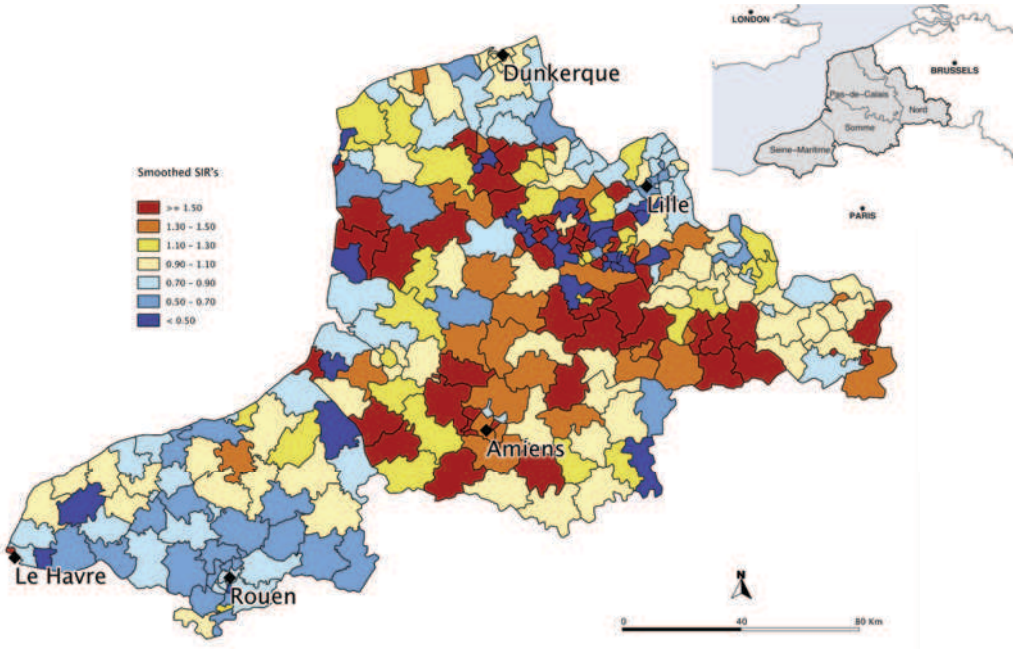


FIGURE 4.2 – Rapport standardisés d'incidence lissés de MC, ajustés sur le sexe et l'âge au diagnostic, de 1990 à 2006, au sein de la zone géographique couverte par le registre EPIMAD dans le nord de la France.

Les méthodes de statistiques de scan spatio-temporelles ont fourni deux types de clusters : les clusters spatiaux constants et les clusters spatiaux non-constants. La description des caractéristiques des clusters détectés significatifs (nombres de cas observé et attendu, probabilité critique, période temporelle et localisation géographique) sont disponibles au sein de la Table 4.1. La répartition des clusters spatiaux constants est montrée en Figure 4.3. Les analyses ont mis en évidence quatorze clusters significatifs dont cinq clusters de sur-incidence (total : 726 patients) et neuf clusters de sous-incidence (total : 521 patients). Le risque relatif varie de 1.88 à 9.80 pour les clusters de sur-incidence. Dans six clusters de sous-incidence, aucun cas de MC n'a été observé. Dans les quatre clusters restants, le risque relatif varie de 0.56 à 0.66. Par ailleurs, les analyses ont mis en évidence quatre clusters spatiaux non-constants significatifs (Figure 4.4), dont trois clusters de sur-incidence (total : 779 patients) au cours d'une période de 9 à 12 ans, et un cluster de sous-incidence (total : 4 patients) sur une période de 7 ans. Le risque relatif varie, pour les cluster de sur-incidence, de 1.66

à 2.03, et a pour valeur 0.14 pour l'unique cluster de sous-incidence. Parmi ces clusters spatiaux non constants, trois d'entre eux n'étaient plus significatifs à la fin de la période d'étude (2006), dont deux clusters de sur-incidence et un cluster de sous-incidence. Le cluster restant significatif à la fin de la période d'étude est apparu en 1996 et est un cluster de sur-incidence. Il est situé en zone côtière, au bord de la mer du Nord (Boulogne : cluster 16 ; Figure 4.4).

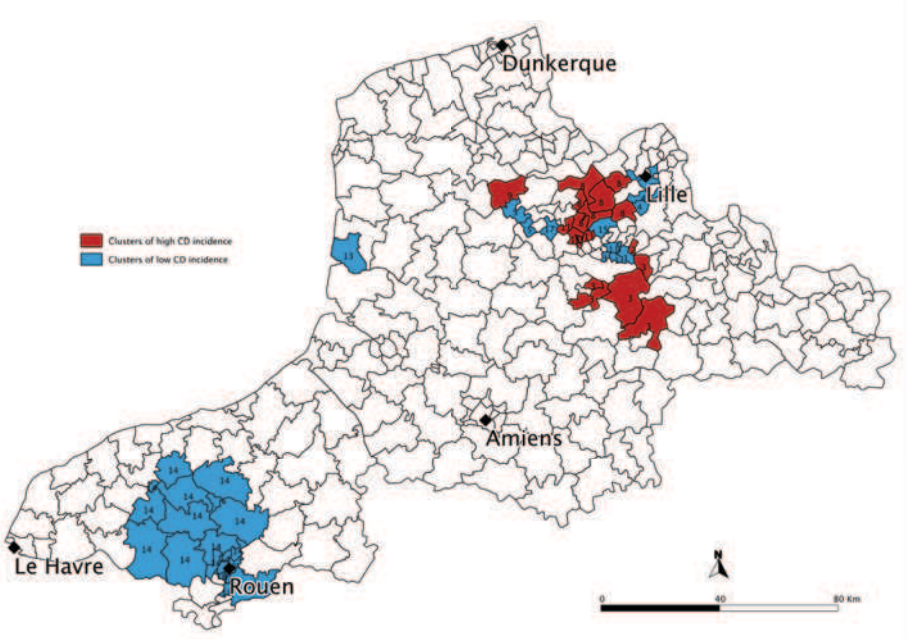


FIGURE 4.3 – Risques relatifs de MC dans le nord de la France pendant la période 1990 - 2006, clusters spatiaux constants pendant l'ensemble de la période étudiée.

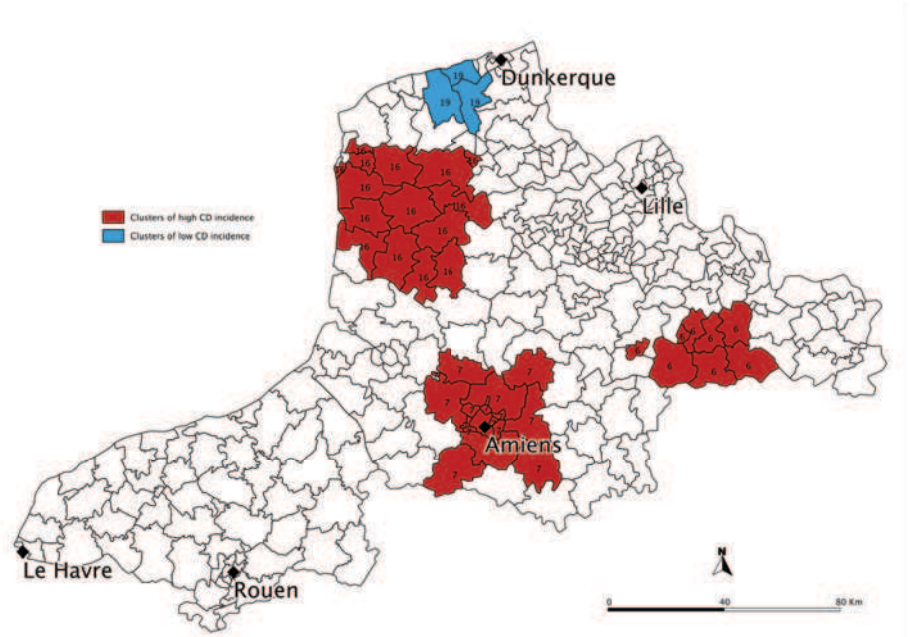


FIGURE 4.4 – Risques relatifs de MC dans le nord de la France pendant la période 1990 - 2006, clusters spatiaux non-constants pendant l'ensemble de la période étudiée.

TABLE 4.1 – Description des clusters de MC issus des analyses spatio-temporelles, nord de la France, 1990-2006.

ID cluster[a]	Type[b]	Date de départ	Date de fin	Population	Rayon (Km)[c]	Cas observés	Cas attendus	RR[a]	LLR[a]	p
Clusters spatiaux constants										
1	M[a]	01/01/90	31/12/06	83996	3.79	0	90.33	0	90.967521	0.0001
4	S[a]	01/01/90	31/12/06	42523	3.64	0	46.78	0	46.954198	0.0001
5	S	01/01/90	31/12/06	41190	7.06	0	42.28	0	42.420722	0.0001
13	S	01/01/90	31/12/06	23652	7.32	0	25.45	0	25.500852	0.0001
15	S	01/01/90	31/12/06	22803	4.82	0	24.13	0	24.176649	0.0001
17	S	01/01/90	31/12/06	18762	3.18	0	19.96	0	19.992701	0.0310
10	S	01/01/90	31/12/06	215032	8.58	178	309.23	0.56	34.306021	0.0001
14	S	01/01/90	31/12/06	247561	19.06	163	269.42	0.59	25.417637	0.0001
18	S	01/01/90	31/12/06	225942	10.39	180	269.43	0.66	17.466933	0.0270
2	S	01/01/90	31/12/06	5703	2.50	61	6.28	9.8	84.215722	0.0001
9	S	01/01/90	31/12/06	19768	12.05	69	20.58	3.38	35.234284	0.0001
3	S	01/01/90	31/12/06	88445	12.56	234	94.22	2.54	74.634570	0.0001
11	S	01/01/90	31/12/06	49753	4.62	117	52.66	2.24	29.385633	0.0001
8	S	01/01/90	31/12/06	123905	8.89	245	132.9	1.88	38.754113	0.0001
Clusters spatiaux non-constants										
19	S	01/01/90	31/07/96	67713	10.78	4	29.17	0.14	17.268793	0.0320
6	S	01/06/92	31/05/04	144912	17.35	214	107.4	2.03	41.835657	0.0010
7	S	01/09/94	30/11/05	268953	21.78	352	212.15	1.7	39.955464	0.0010
16	S	01/08/96	30/06/06	213764	27.65	213	129.86	1.66	22.805932	0.0010

[a] M most likely ; S secondaire ; RR risque relatif dans le cluster comparé au reste de la zone étudiée ; LLR log likelihood ratio

[b] ID, localisation géographique des clusters au sein des Figures 4.3 and 4.4

[c] Rayon, distance en Km entre le centre du cluster et ses frontières

3.2 Analyse de l'offre de soin en gastro-entérologie

En 1990, il y avait 225 gastro-entérologues consultant dans 98 secteurs médicaux (public ou privé) et en 2006, il y avait 262 gastro-entérologues consultant dans 116 secteurs médicaux. De 1990 à 2006, au sein de la zone géographique couverte par le registre EPIMAD, aucun cluster temporel, spatial ou spatio-temporel de gastro-entérologues n'a été mis en évidence.

4 Discussion

Cette étude a permis de montrer une forte hétérogénéité spatiale de l'incidence de MC dans le nord de la France pendant la période de 1990 à 2006, ce qui permet de confirmer et d'étendre les résultats publiés dans [Declercq et al., 2010]. En utilisant les statistiques de scan spatio-temporelles, 18 clusters ont été identifiés dont 14 clusters spatiaux constants et 4 clusters spatiaux non-constants. Parmi ces 14 clusters, 5 clusters de sur-incidence (total : 726 patients) et 9 clusters de sous-incidence (total : 521 patients) ont été mis en évidence. Au sein des 4 clusters spatiaux non-constants, 3 clusters de sur-incidence (total : 779 patients) et 1 cluster de sous-incidence (total : 4 patients) ont été détectés.

Les statistiques de scan spatiales ont été précédemment utilisées dans des études portant sur la détection de clusters de cas de MC. Dans la province canadienne de Manitoba, [Green et al., 2006] ont mis en évidence des clusters sur-incidence et de sous-incidence de MC et ont démontré une relation avec les caractéristiques de la population. En effet, ils ont montré une corrélation négative entre l'incidence de MC et les taux d'infection entérique à déclaration, ce qui tend à confirmer l'hypothèse hygiénique. Dans une étude réalisé en Norvège [Aamodt et al., 2008], un cluster de cas de MICI constitué de 4 municipalités a été identifié. De surcroît, une analyse écologique a montré que l'incidence de MC était plus élevée de 33% dans les municipalités ayant le plus fort niveau d'éducation en comparaison avec celles ayant le plus bas niveau d'éducation, et l'incidence de MC était plus élevée de 35% dans les municipalités urbaines par rapport à celles rurales. Cependant, aucune étude n'a pris en compte la dimension temporelle dans l'utilisation des méthodes de statistiques de scan, limitant ainsi le poids de facteurs de risque évoluant au cours du temps.

Par ailleurs, le niveau de vie, le stress et les facteurs environnementaux sont considérés comme des facteurs déclencheurs d'apparition de la MC chez les personnes présentant des pré-dispositions génétiques. Aussi, l'existence de clusters spatio-temporels de cas de MC pourrait refléter des variations dans la distribution spatiale de ces possibles facteurs étiologiques . En ce qui concerne le niveau de vie, [Declercq et al., 2010] n'ont pas montré de relation significative entre l'incidence de MC et la défaveur sociale évaluée par l'indice de Towsend. Concernant des évènements de vie stressants susceptibles de déclencher l'apparition de la MC, [Lerebours et al., 2007] ont réalisé, à partir des données du registre EPIMAD, une étude cas-témoins qui avait pour objectif de déterminer si le stress, évoluant au travers d'évènements de vie, était associé à l'apparition de la MC. Après ajustements sur des scores de dépression et d'anxiété ainsi que sur d'autres caractéristiques telles que le statut tabagique et des caractéristiques socio-économiques, aucune association significative n'a été mise en évidence. En ce qui concerne les facteurs environnementaux, les études rétrospectives sont difficiles à mettre en oeuvre du fait du biais de mémorisation. En outre, la détection de clusters géographiques sans biais de pré-sélection au moyen des statistiques de scan, va permettre de réaliser des études épidémiologiques interventionnelles focalisés sans *a priori*.

Au sein de cette étude, deux types de clusters ont été identifiés : certains sont purement spatiaux et par conséquent constants tout au long de la période de temps étudiée, et d'autres étaient non-constants dans le temps. La présence de clusters purement spatiaux peut refléter soit une différence constante d'exposition aux facteurs environnementaux tout au long de la période étudiée soit la présence de contextes de susceptibilité génétique à la MC différents. Il est intéressant de remarquer le fait que le cluster présentant le risque le plus élevé (ID = 2) se situe à proximité de deux fonderies de métaux non-ferreux. Or, une précédente étude [Leroyer et al., 2001] a mis en évidence le fait que les sols de ces municipalités contiennent entre 100 et 1 700 ppm de plomb et que 30% des hommes et 12% des femmes habitant dans cette zone montrent des taux de plombémie excédant 100microg/L. Par ailleurs, le plomb et l'aluminium ont précédemment été cités comme des facteurs environnementaux probable de la MC [Lerner, 2007]. Autre fait intéressant, les principaux clusters de sous-incidence ont été détectés dans le département de la Seine-Maritime. Ce dernier présente

la particularité d'avoir subi une importante invasion Viking en 911 ce qui a conduit à un flux substantiel de gènes scandinaves. Or, il a été mis en évidence par [Riis et al., 2007] que le taux de mutation du gène CARD15/NOD2, mutation responsable d'une prédisposition à la MC, était moins courant dans les populations scandinaves par rapport aux autres pays européens.

La principale innovation de cette étude réside dans l'identification, par le biais des analyses spatio-temporelles, de 3 clusters de sur-incidence et 1 cluster de sous-incidence non constants dans le temps, ce qui suggère fortement l'influence de facteurs environnementaux. Deux clusters de sur-incidence (ID = 6,7) se situent dans des zones rurales qui ont connu, au cours des trente dernières années, un passage de l'agriculture traditionnelle vers la culture céréalière intensive caractérisée par une mécanisation complète des installations et l'utilisation d'engrais chimiques. D'autres études sont nécessaires afin de confirmer le lien entre les cluster de cas de MC et la présence de polluants.

Les principales forces de cette étude résident dans le fait que le registre EPIMAD constitue le plus grand registre mondial (9.3% de la population française) et a permis de disposer d'un grand nombre de cas. Par ailleurs, la méthodologie de collecte des cas est bien adaptée au système de soins français et son exhaustivité a été évaluée à 96% [Gower-Rousseau et al., 1994]. Aussi, la combinaison de l'exhaustivité du registre et l'utilisation de statistiques de scan spatio-temporelles est particulièrement bien adaptée à l'identification de clusters géographiques d'origines environnementales. Par ailleurs, l'homogénéité de l'offre de soin en gastro-entérologie dans la zone géographique couverte par le registre confirme que les clusters de cas de MC mis en évidence ne sont pas imputables à des modifications temporelles ou spatiales de l'offre de soin.

Néanmoins, cette étude présente également des faiblesses. En effet, n'ont pas été pris en compte le manque d'informations sur les conditions de vie des patients, leur statut socio-professionnel et sur d'autres facteurs de risques potentiels de MC qui ne permettent pas de mettre en corrélation directe les clusters géographiques et les caractéristiques environnementales. De surcroît, le registre renseigne le lieu de résidence des patients à la date du diagnostic et non avant l'apparition de la maladie.

En outre, la principale force de cette étude réside dans la taille important du registre qui inclut des patients vivant dans différentes zones en termes d'environnement (urbain, péri-urbain et rural) et, en termes de pollution, ce qui permet la réalisation de futures études épidémiologiques analytiques telles que des études cas/témoins.

5 Conclusion

Dans le cadre de l'épidémiologie, et notamment au sein des études écologiques, les statistiques de scan spatiales et spatio-temporelles se révèlent être des outils pertinents et robustes afin de détecter des clusters atypiques (*i.e.* statistiquement significatifs) en termes de sur-incidence ou sous-incidence de cas d'une pathologie. La prise en compte des dimensions spatiale et temporelle rend ces méthodes singulières dans le domaine des techniques de détection de clusters.

Au travers de cette étude, nous avons mis en évidence le fait que le nord de la France est caractérisé par une forte hétérogénéité spatiale et spatio-temporelle de l'incidence de la maladie de Crohn. L'existence de clusters spatiaux constants et non constants tout au long de la période d'étude suggère que les facteurs de risque de la MC sont encore à l'oeuvre dans cette région. Cette spécificité ouvre la voie à de futures études portant sur les différences de distribution des variants génétiques et des facteurs environnementaux entre les clusters de sur-incidence et ceux de sous-incidence en utilisant une approche objective.

Conclusion générale et perspectives

1 Conclusion générale

Dans le cadre de cette thèse, nous nous sommes intéressés au domaine de la détection de clusters au moyen des statistiques de scan. Ce travail s'est tout d'abord axé sur l'influence de la forme de la fenêtre de scan sur l'approximation de la distribution de la statistique de scan bidimensionnelle discrète (Chapitre 2). Ensuite, nous nous sommes intéressés au cas des statistiques de scan spatiales et notamment à l'approximation de la distribution de la statistique de test (Chapitre 3). Enfin, nous avons appliqué ces méthodes à l'étude de la Maladie de Crohn dans le Nord-Pas-De-Calais par le biais des données collectées par le registre EPIMAD (Chapitre 4).

Au sein du chapitre 2, nous avons évalué l'influence de la forme de la fenêtre de scan sur la distribution des statistiques de scan bidimensionnelles discrètes. Nous avons réalisé une étude de simulation prenant en compte des fenêtres de scan de forme carrée, rectangulaire et circulaire (cercle discret). Cette étude a mis en évidence le fait que les distributions des statistiques de scan associées à ces formes sont très proches les unes des autres mais significativement différentes. Par ailleurs, nous avons mis en évidence, par le biais d'une étude de simulation, que la puissance du test basé sur les statistiques de scan est liée à la forme de la fenêtre ainsi qu'à la forme du cluster existant sous l'hypothèse alternative.

Dans le chapitre 3, nous avons proposé, dans le cadre des statistiques de scan spatiales, une alternative à la méthode de Monte Carlo pour le test de rapport de vraisemblance généralisé. Le principe de cette alternative réside dans le fait de conserver la phase de détection du MLC et, en utilisant une modélisation basée sur une suite variables aléatoires de Bernoulli, d'estimer la probabilité associée au MLC. L'approximation de cette probabilité a été réalisée par la formule proposée dans [Haiman, 2007] basée sur les suites de variables aléatoires 1-dépendantes. Cette approximation, qui donne également une erreur d'approximation, présente l'avantage d'être constituée de quantités qui sont calculables par le biais de formules exactes, diminuant ainsi les temps de calculs. Nous avons réalisé une étude de simulation visant à comparer la méthode basée sur les simulations de Monte Carlo et celle basée sur notre modélisation. A partir de simulations sous l'hypothèse alternative, nous avons mis en évidence que les deux méthodes fournissent des probabilités concordantes, notre méthode présentant une puissance supérieure à celle basée sur les simulations de Monte Carlo et un temps de calcul nettement plus faible.

Au sein du chapitre 4, nous avons appliqué les statistiques de scan spatio-temporelles à l'étude de la répartition spatiale des cas de maladie de Crohn (MC) dans le nord de la France, de 1990 à 2006. Cette pathologie chronique, dont l'étiologie est inconnue, n'est pas mortelle mais altère significativement la vie des patients. Nous avons isolés 18 clusters significatifs de deux types : 14 clusters spatiaux constants dans le temps et 4 clusters spatiaux non-constants dans le temps. Parmi les 14 clusters du premier type, 5 clusters de sur-incidence et 4 clusters de sous-incidence ont été détectés. Parmi les 4 clusters du deuxième type, 3 clusters de sur-incidence et 1 cluster de sous-incidence ont été détectés. Ces clusters ouvrent la voie à des recherches de facteurs causaux de nature génétique et/ou environnementale.

2 Perspectives

Les statistiques de scan se révèlent être un domaine de recherche offrant un nombre conséquent de perspectives, autant sur le plan théorique qu'appliqué. Cette section présente, de manière succincte, les perspectives de recherche que nous voudrions développer.

Statistiques de scan à fenêtre variable. Le chapitre 2 s'est attaché à décrire la plupart des approximations de la distribution des statistiques de scan à fenêtre fixe. Or, dans [Glaz and Zhang, 2004], les auteurs ont mis en évidence le fait que les statistiques de scan à fenêtre variable présentent une meilleure sensibilité concernant la détection de clusters atypiques. Depuis, [Glaz and Zhang, 2006, Zhang and Glaz, 2008] ont proposé des méthodes de type *Maximum scan score-type statistics* et *Minimum p-value scan statistics* afin de donner une approximation des statistiques de scan à fenêtre variable, en se basant sur des approximations de type "Product-Type". Nous proposons de développer une méthodologie, basée sur l'approximation d'Haiman, afin de donner une approximation des distributions des statistiques de scan à fenêtre variable, dans les cas uni et bidimensionnels.

Statistiques de scan spatiales : prise en compte de variables explicatives. Les statistiques de scan spatiales ne permettent pas, à l'heure actuelle, de prendre en compte des variables explicatives dans le calcul du nombre attendu de cas dans une unité spatiale. Or, en épidémiologie, l'hétérogénéité spatiale du nombre de cas de maladie peut être imputable à de nombreux facteurs tels que les caractéristiques socio-démographiques propres à chaque unité spatiale. La prise en compte de ces facteurs est réalisée, pour le moment, par le biais de standardisation indirecte et en considérant uniquement des variables qualitatives. Une piste de solution réside dans l'utilisation de modèles spatiaux log-linéaires afin de prendre en compte tout type de variables explicatives et, par conséquent, donner une estimation du nombre attendu de cas par unité spatiale.

Prise en compte de données manquantes. Dans le cadre temporel, une donnée manquante peut être définie par un évènement intervenant dans un intervalle de temps mais dont la date précise est inconnue. Dans le cadre spatial, une donnée manquante est définie par un évènement apparaissant dans le domaine géographique étudié mais dont les coordonnées spatiales sont inconnues. A l'heure actuelle, les méthodes de statistiques spatiales prennent en compte que des données complètes. Dans le cas de présence de données manquantes, deux possibilités s'offrent à nous : la suppression de ces données ou leur imputation. La première possibilité implique *de facto* une perte de puissance statistique et nécessite l'hypothèse d'une répartition uniforme des données manquantes (*missing at random*). Or, si cette hypothèse n'est pas vérifiée, les clusters détectés pourraient être uniquement considérés comme des artefacts liés à disparité des fréquences de données manquantes selon les unités temporelles ou spatiales. Aussi, la prise en compte de données manquantes par le biais de méthodes d'imputation de données temporelles ou spatiales nous semble une perspective de recherche pertinente.

Applications à l'épidémiologie. Le chapitre 4 a mis en évidence l'intérêt des méthodes de statistiques de scan dans la détection de clusters atypiques de clusters de cas de maladies. Dans le cadre de pathologies dont l'étiologie est peu connue voire inconnue, la mise en évidence de zones géographiques atypiques permet aux cliniciens d'émettre des hypothèses sur les causes de ces maladies. A la vue du nombre important de registres de cas de maladies dans le domaine médical, il apparaît évident que les méthodes de statistiques de scan vont être amenées à être utilisées de manière récurrente. A l'heure actuelle, nous appliquons ces méthodes dans l'analyse de la répartition spatiale des cas de lymphomes non-hodgkiniens dans le Nord-Pas-De-Calais (Registre Lymphonor), les cas d'insuffisance rénale, dans la même région (registre REIN) ainsi que les cas de tentative de suicide (F2RSM).

Les perspectives d'application de ces méthodes dans le domaine de la Santé Publique sont nombreuses et notamment dans le cadre de données génétiques et environnementales. Par ailleurs, les statistiques de scan peuvent être également utilisées dans le cadre de systèmes de surveillance prospectifs, particulièrement dans la détection d'épidémies.

Annexe A

Test basé sur la statistique de scan : cas bidimensionnel discret

1 Modèle binomial

Soit $[0, N_1] \times [0, N_2]$ une région rectangulaire, $N_1, N_2 \in \mathbb{N}$. Soit $\{X_{ij}\}$ une famille de variables aléatoires binomiales $\mathcal{B}(n, p_k)$, indépendantes :

$$\mathbb{P}(X_{ij} = x_{ij}) = \binom{n}{x_{ij}} p_k^{x_{ij}} (1 - p_k)^{n - x_{ij}} \qquad 1 \leq i \leq N_1,\ 1 \leq j \leq N_2,$$

où

$$\forall t, s, m_1, m_2 \in \mathbb{N},\ 1 \leq m_1 < N_1,\ 1 \leq m_2 < N_2,\ 1 \leq t \leq N_1 - m_1 + 1,\ 1 \leq s \leq N_2 - m_2 + 1$$

$$k = \begin{cases} 0 & \text{si} \quad \{i, j\} \in [0, N_1] \times [0, N_2] \backslash [t, t + m_1 - 1] \times [s, s + m_2 - 1] \\ 1 & \text{si} \quad \{i, j\} \in [t, t + m_1 - 1] \times [s, s + m_2 - 1] \end{cases}$$

Les X_{ij} correspondent au nombre d'évènements observés dans la sous-région élémentaire $[i-1, i] \times [j-1, j]$. Supposons m_1 et m_2 fixés et les paramètres p_k connus. On souhaite tester l'hypothèse nulle $\mathcal{H}_0$ selon laquelle les X_{ij} sont *i.i.d.* $\mathcal{B}(n, p_0)$:

$$\mathcal{H}_0 : p_0 = p_1 \tag{A.1}$$

Contre une hypothèse alternative $\mathcal{H}_1$ supportant un cluster d'évènements dans une fenêtre de taille $m_1 \times m_2$:

$$\mathcal{H}_1 : p_1 > p_0 \tag{A.2}$$

Proposition A.1. *Le test de rapport de vraisemblance généralisé (TRVG) rejette $\mathcal{H}_0$ (A.1) au profit de $\mathcal{H}_1$ (A.2) lorsque la statistique de scan bidimensionnelle discrète à fenêtre de taille fixe $m_1 \times m_2$, $S(m1, m_2, N_1, N_2)$, excède un seuil τ déterminé à partir de $\mathbb{P}(S(m1, m_2, N_1, N_2) > \tau | \mathcal{H}_0) = \alpha$ où α correspond au risque de première espèce associé au test.*

Preuve. *Le preuve suit un raisonnement identique au cas unidimensionnel. La fonction de vraisemblance sous $\mathcal{H}_0$, $L_{\mathcal{H}_0}$, a pour expression*

$$L_{\mathcal{H}_0} = \prod_{i=1}^{N_1} \prod_{j=1}^{N_2} \binom{n}{x_{ij}} p_0^{x_{ij}} (1 - p_0)^{n - x_{ij}}.$$

La fonction de vraisemblance $L_{\mathcal{H}_1}(t, s)$, a pour expression

$$L_{\mathcal{H}_1}(t, s) = \prod_{i=1}^{t-1} \prod_{j=1}^{s-1} \binom{n}{x_{ij}} p_0^{x_{ij}} (1 - p_0)^{n - x_{ij}} \prod_{i=t}^{t+m_1-1} \prod_{j=s}^{s+m_2-1} \binom{n}{x_{ij}} p_1^{x_{ij}} (1 - p_1)^{n - x_{ij}} \times \prod_{i=t+m_1}^{N_1} \prod_{j=s+m_2}^{N_2} \binom{n}{x_{ij}} p_0^{x_{ij}} (1 - p_0)^{n - x_{ij}}.$$

Le rapport de vraisemblance $LR(t,m)$ a donc pour expression

$$LR(t,s,m_1,m_2) = \frac{\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_1^{x_{ij}}(1-p_1)^{n-x_{ij}}}{\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}}},$$

et son logarithme, $LLR(t,s,m_1,m_2)$,

$$\begin{aligned} LLR(t,s,m_1,m_2) = &\log \prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_1^{x_{ij}}(1-p_1)^{n-x_{ij}} \\ &- \log \prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}}, \end{aligned}$$

qui se simplifie

$$LLR(t,s,m_1,m_2) = \sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}\left[x_{ij}\log\left(\frac{p_1}{p_0}\right) + (n-x_{ij})\log\left(\frac{1-p_1}{1-p_0}\right).\right]$$

Tout comme dans le cas unidimensionnel, posons $C_1 = \log\left(\frac{p_1}{p_0}\right)$ et $C_2 = \log\left(\frac{1-p_1}{1-p_0}\right)$. Comme $p_1 > p_0$ alors $C_1 > 0$ et $C_2 < 0$.

$$LLR(t,s,m_1,m_2) = C_1\sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}x_{ij} + C_2\sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}(n-x_{ij}).$$

Posons

$$\nu_{ts} = \sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}x_{ij},$$

ν_{ts} correspond au nombre de cas observés dans la fenêtre de taille $[t,t+m_1-1]\times[s,s+m_2-1]$.

$$LLR(t,s,m_1,m_2) = C_1\nu_{ts} + C_2(m_1m_2n - \nu_{ts}).$$

Pour m_1 et m_2 fixés, et comme $C_1 > 0$ et $C_2 < 0$, le $LLR(t,m)$ est une fonction monotone croissante en ν_{ts}. Par conséquent, le test de rapport de vraisemblance rejette $\mathcal{H}_0$ pour une valeur de ν_{ts} aussi grande que possible, à la savoir la statistique de scan bidimensionnelle discrète à fenêtre fixe de taille $m_1\times m_2$:

$$S(m_1,m_2,N_1,N_2) = \max_{\substack{1\le t\le N_1-m_1+1\\ 1\le s\le N_2-m_2+1}} \nu_{t,s}.$$

2 Modèle de Poisson

Soit $[0,N_1]\times[0,N_2]$ une région rectangulaire, $N_1,N_2\in\mathbb{N}$. Soit $\{X_{ij}\}$ une famille de variables aléatoires de Poisson $\mathcal{P}(\lambda_k)$, indépendantes :

$$\mathbb{P}(X_{ij}=x_{ij}) = \frac{e^{-\lambda_k}\lambda_k^{x_{ij}}}{x_{ij}!} \qquad 1\le i\le N_1,\ 1\le j\le N_2,$$

où

$\forall t,s,m_1,m_2\in\mathbb{N},\ 1\le m_1<N_1,\ 1\le m_2<N_2,\ 1\le t\le N_1-m_1+1,\ 1\le s\le N_2-m2+1$

$$k = \begin{cases} 0 & \text{si} \quad \{i,j\}\in[0,N_1]\times[0,N_2]\backslash[t,t+m_1-1]\times[s,s+m_2-1] \\ 1 & \text{si} \quad \{i,j\}\in[t,t+m_1-1]\times[s,s+m_2-1] \end{cases}$$

Les X_{ij} correspondent au nombre d'évènements observés dans la sous-région élémentaire $[i-1,i]\times[j-1,j]$. Supposons m_1 et m_2 fixés et les paramètres p_k connus. On souhaite tester l'hypothèse nulle $\mathcal{H}_0$ selon laquelle les X_{ij} sont *i.i.d.* $\mathcal{P}(\lambda_0)$:

$$\mathcal{H}_0 : \lambda_0 = \lambda_1 \tag{A.3}$$

Contre une hypothèse alternative $\mathcal{H}_1$ supportant un cluster d'évènements dans une fenêtre de taille $m_1 \times m_2$:

$$\mathcal{H}_1 : \lambda_1 > \lambda_0 \tag{A.4}$$

Proposition A.2. *Le test de rapport de vraisemblance généralisé (TRVG) rejette $\mathcal{H}_0$ (A.3) au profit de $\mathcal{H}_1$ (A.4) lorsque la statistique de scan bidimensionnelle discrète à fenêtre de taille fixe $m_1 \times m_2$, $S(m1, m_2, N_1, N_2)$, excède un seuil τ déterminé à partir de $\mathbb{P}(S(m1, m_2, N_1, N_2) > \tau|\mathcal{H}_0) = \alpha$ où α correspond au risque de première espèce associé au test.*

Preuve. *Le preuve suit un raisonnement identique au cas unidimensionnel. La fonction de vraisemblance sous $\mathcal{H}_0$, $L_{\mathcal{H}_0}$, a pour expression*

$$L_{\mathcal{H}_0} = \prod_{i=1}^{N_1}\prod_{j=1}^{N_2} \frac{e^{-\lambda_0}\lambda_0^{x_{ij}}}{x_{ij}!}.$$

La fonction de vraisemblance $L_{\mathcal{H}_1}(t,s)$, a pour expression

$$L_{\mathcal{H}_1}(t,s) = \prod_{i=1}^{t-1}\prod_{j=1}^{s-1} \frac{e^{-\lambda_0}\lambda_0^{x_{ij}}}{x_{ij}!} \prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1} \frac{e^{-\lambda_1}\lambda_1^{x_{ij}}}{x_{ij}!} \prod_{i=t+m_1}^{N_1}\prod_{j=s+m_2}^{N_2} \frac{e^{-\lambda_0}\lambda_0^{x_{ij}}}{x_{ij}!}.$$

Le rapport de vraisemblance $LR(t,m)$ a donc pour expression

$$LR(t,s,m_1,m_2) = \frac{\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1} \frac{e^{-\lambda_1}\lambda_1^{x_{ij}}}{x_{ij}!}}{\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1} \frac{e^{-\lambda_0}\lambda_0^{x_{ij}}}{x_{ij}!}},$$

et son logarithme, $LLR(t,s,m_1,m_2)$,

$$LLR(t,s,m_1,m_2) = \sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1} \left[(-\lambda_1 + x_{ij}\log(\lambda_1)) - (\lambda_0 + x_{ij}\log(\lambda_0))\right],$$

qui se simplifie

$$LLR(t,s,m_1,m_2) = m_1m_2(\lambda_0-\lambda_1) + C\sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1} x_{ij},$$

où $C = \log\left(\frac{\lambda_1}{\lambda_0}\right)$. Posons

$$\nu_{ts} = \sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1} x_{ij},$$

où ν_{ts} correspond au nombre de cas observés dans la fenêtre de taille $[t,t+m_1-1]\times[s,s+m_2-1]$.

$$LLR(t,s,m_1,m_2) = m_1m_2(\lambda_0-\lambda_1) + C\nu_{ts}$$

Comme $\lambda_1 > \lambda_0$, $C > 0$ et $m_1m_2(\lambda_0-\lambda_1) < 0$. Pour m_1 et m_2 fixés le $LLR(t,m)$ est une fonction monotone croissante en ν_{ts}. Par conséquent, le test de rapport de vraisemblance rejette $\mathcal{H}_0$ pour une valeur de ν_{ts} aussi grande que possible, à la savoir la statistique de scan bidimensionnelle discrète à fenêtre fixe de taille $m_1 \times m_2$:

$$S(m_1,m_2,N_1,N_2) = \max_{\substack{1\le t\le N_1-m_1+1\\ 1\le s\le N_2-m_2+1}} \nu_{t,s}.$$

Annexe B

Formule d'interpolation linéaire

Si N_1 and N_2 ne sont pas multiples de m_1 and m_2 nous avons besoin d'une interpolation afin de donner une approximation de $\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq n)$. Posons $\mathcal{R}_1 = N_1^1 \times N_2^1$ la plus grande région rectangulaire contenue dans $\mathcal{R} = N_1 \times N_2$ et posons $\mathcal{R}_2 = N_1^2 \times N_2^2$ la plus petite région rectangulaire contenant $\mathcal{R}$.

Pour un $n \in \mathbb{N}$ donné, posons $p_1 = \mathbb{P}(S(m_1, m_2, N_1^1, N_2^1) \leq n)$ correspondant à la valeur de la distribution de la statistique de scan en scannant $\mathcal{R}_1$ et $p_2 = \mathbb{P}(S(m_1, m_2, N_1^2, N_2^2) \leq n)$ la valeur de la distribution de la statistique de scan en scannant $\mathcal{R}_2$. De manière évidente, $p_1 \geq p_2$ et $\mathbb{P}(S(m_1, m_2, N_1, N_2) \leq n) \in [p_1, p_2]$.

Posons $p^* = \mathbb{P}(S(m_1, m_2, N_1, N_2) \leq n)$ la valeur estimée de la distribution de la statistique de scan en scannant $\mathcal{R}$ obtenue par interpolation linéaire comme suit (voir Figure B.1). L'équation de

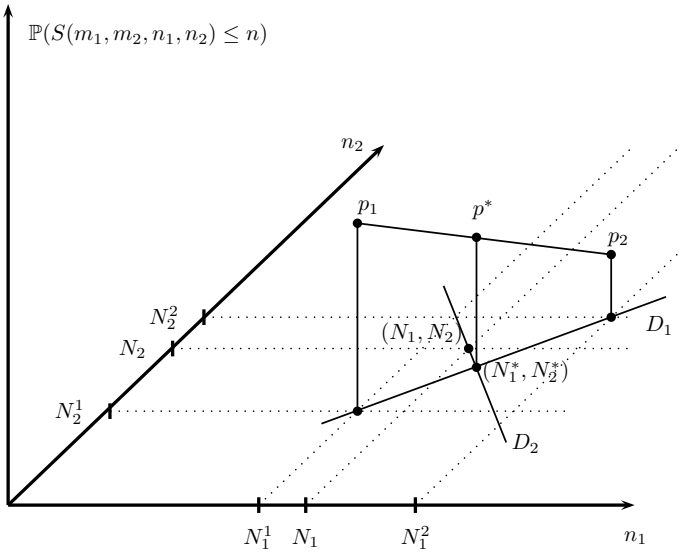


FIGURE B.1 – Approximation de p^* par interpolation linéaire.

la droite D_1 passant par les points (N_1^1, N_2^1) et (N_1^2, N_2^2) a pour expression

$$n_2 = An_1 + B,$$

où

$$A = \frac{N_2^2 - N_2^1}{N_1^2 - N_1^1}$$

et

$$B = N_2^1 - N_1^1 \left(\frac{N_2^2 - N_2^1}{N_1^2 - N_1^1} \right).$$

Posons (N_1^*, N_2^*) le point sur D_1 qui est le plus proche de (N_1, N_2). Clairement, (N_1^*, N_2^*) est le point défini par l'intersection de D_1 avec la droite perpendiculaire à D_1 passant par (N_1, N_2), nommée D_2. Des calculs élémentaires donnent les expressions suivantes pour N_1^* et N_2^* :

$$N_1^* = \frac{C - B}{A + \dfrac{1}{A}},$$

$$N_2^* = \frac{A^2(C - B)}{A^2 + 1} + B,$$

avec $C = N_2 + \frac{1}{A} N_1$. En utilisant le théorème de Thalès, on obtient finalement

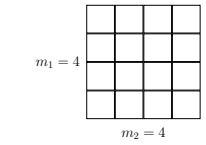
$$p^* = p_1 - (p_1 - p_2) \frac{N_1^1 - N_1^*}{N_1^1 - N_1^2}.$$

Annexe C

Forme de la fenêtre : applications numériques supplémentaires

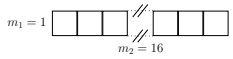
TABLE C.1 – Approximations de $\mathbb{P}(S \leqslant n)$: Scan d'une région $N_1 \times N_2 = 42 \times 42$ avec différentes formes de fenêtres.

Scanning window shape | **Approximation and approximation error for $\mathbb{P}(S \leqslant n)$**



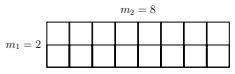
Square
$m_1 = m_2 = 4,\ K = L = 12,\ N = 10^6$

	$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$	
n	(2.16)	(2.23)	n	(2.16)	(2.23)	n	(2.16)	(2.23)
3	0.985482	0.005324	13	0.941050	0.015285	13	0.907245	0.028243
4	0.999735	0.000978	14	0.982274	0.005066	14	0.971505	0.007914
5	0.999999	7.99E-10	15	0.995073	0.002207	15	0.992204	0.003505



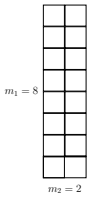
Rectangle
$m_1 = 1,\ m_2 = 16,\ K = 48,\ L = 3,\ N = 10^6$

	$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$	
n	(2.16)	(2.23)	n	(2.16)	(2.23)	n	(2.16)	(2.23)
3	0.990571	0.003023	13	0.963984	0.006572	13	0.957232	0.021233
4	0.999807	0.000488	14	0.990747	0.003057	14	0.987906	0.005119
5	1.000000	0.000000	15	0.998023	0.001356	15	0.996412	0.002316



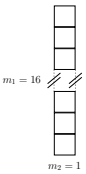
Rectangle
$m_1 = 2,\ m_2 = 8,\ K = 24,\ L = 6,\ N = 10^6$

	$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$	
n	(2.16)	(2.23)	n	(2.16)	(2.23)	n	(2.16)	(2.23)
3	0.982114	0.004750	13	0.951232	0.010641	13	0.917379	0.034018
4	0.999779	0.000748	14	0.988426	0.004261	14	0.976583	0.008009
5	1.000000	0.000000	15	0.996092	0.002124	15	0.993762	0.003004



Rectangle
$m_1 = 8,\ m_2 = 2,\ K = 6,\ L = 24,\ N = 10^6$

	$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$	
n	(2.16)	(2.23)	n	(2.16)	(2.23)	n	(2.16)	(2.23)
3	0.987399	0.005148	13	0.953775	0.016608	13	0.918798	0.019169
4	0.999867	0.000676	14	0.985515	0.004862	14	0.978260	0.006486
5	1.000000	0.000000	15	0.997181	0.001674	15	0.994439	0.003073



Rectangle
$m_1 = 16,\ m_2 = 1,\ K = 3,\ L = 48,\ N = 10^6$

	$X_{ij} \sim \mathcal{B}(1, 0.01)$			$X_{ij} \sim \mathcal{B}(5, 0.05)$			$X_{ij} \sim \mathcal{P}(0.25)$	
n	(2.16)	(2.23)	n	(2.16)	(2.23)	n	(2.16)	(2.23)
3	0.990953	0.003519	13	0.965199	0.014372 8	13	0.958368	0.009658
4	0.999764	0.000521	14	0.991177	0.003537	14	0.986935	0.004355
5	1.000000	0.000000	15	0.997826	0.001518	15	0.996679	0.002277

Bibliographie

[Aamodt et al., 2008] Aamodt, G., Jahnsen, J., Bengtson, M. B., Moum, B., and Vatn, M. H. (2008). Geographic distribution and ecological studies of inflammatory bowel disease in southeastern norway in 1990-1993. *Inflamm Bowel Dis*, 14(7) :984–91.

[Abrams et al., 2010] Abrams, A. M., Kleinman, K., and Kulldorff, M. (2010). Gumbel based p-value approximations for spatial scan statistics. *Int J Health Geogr*, 9 :61.

[Aldous, 1989] Aldous, D. J. (1989). *Probability approximations via the Poisson clumping heuristic*. Springer-Verlag New York.

[Alm, 1983] Alm, S. E. (1983). On the distribution of scan statistic of poisson process. In Gut, A. and Helst, L., editors, *Probability and Mathematical Statistics*, pages 1–10. Upsalla University Press.

[Alm, 1997] Alm, S. E. (1997). On the distributions of scan statistics of a two-dimensional poisson process. *Advances in Applied Probability*, pages 1–18.

[Alm, 1998] Alm, S. E. (1998). Approximation and simulation of the distributions of scan statistics for poisson processes in higher dimensions. *Extremes*, 1(1) :111–126.

[Alt and Vach, 1991] Alt, K. and Vach, W. (1991). The reconstruction of "genetic kinship" in prehistoric burial complexes — problems and statistics. In Bock, H.-H. and Ihm, P., editors, *Classification, Data Analysis, and Knowledge Organization*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 299–310. Springer Berlin Heidelberg.

[Amarioarei and Preda, 2013] Amarioarei, A. and Preda, C. (2013). Approximations for the distribution of three dimensional scan statistic. *Methodology and Computing in Applied Probability*, page In press.

[Amin et al., 2010] Amin, R., Bohnert, A., Holmes, L., Rajasekaran, A., and Assanasen, C. (2010). Epidemiologic mapping of florida childhood cancer clusters. *Pediatr Blood Cancer*, 54(4) :511–8.

[Anderson and Titterington, 1997] Anderson, N. H. and Titterington, D. M. (1997). Some methods for investigating spatial clustering, with epidemiological applications. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 160(1) :87–105.

[Assunção et al., 2006] Assunção, R., Costa, M., Tavares, A., and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25(5) :723–742.

[Balakrishnan and Koutras, 2001] Balakrishnan, N. and Koutras, M. V. (2001). *Runs and scans with applications*. Wiley.

[Barnard, 1963] Barnard, G. (1963). Discussion of professor bartlett's paper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25B(294).

[Baron et al., 2005] Baron, S., Turck, D., Leplat, C., Merle, V., Gower-Rousseau, C., Marti, R., Yzet, T., Lerebours, E., Dupas, J. L., Debeugny, S., Salomez, J. L., Cortot, A., and Colombel, J. F. (2005). Environmental risk factors in paediatric inflammatory bowel diseases : a population based case control study. *Gut*, 54(3) :357–63.

[Barreiro-de Acosta et al., 2011] Barreiro-de Acosta, M., Alvarez Castro, A., Souto, R., Iglesias, M., Lorenzo, A., and Dominguez-Munoz, J. E. (2011). Emigration to western industrialized countries : A risk factor for developing inflammatory bowel disease. *J Crohns Colitis*, 5(6) :566–9.

[Besag and Clifford, 1991] Besag, J. and Clifford, P. (1991). Sequential monte carlo p-values. *Biometrika*, 78(2) :301–304.

[Besag and Diggle, 1977] Besag, J. and Diggle, P. (1977). Simple monte carlo tests for spatial pattern. *Applied Statistics*, pages 327–333.

[Besag and Newell, 1991] Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 154(1) :143–155.

[Besag et al., 1991] Besag, J., York, J., and Molli©, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43 :1–21.

[Bithell, 2007] Bithell, J. (2007). The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, 14(21-22) :2309–2322.

[Boutsikas and Koutras, 2003] Boutsikas, M. V. and Koutras, M. V. (2003). Bounds for the distribution of two-dimensional binary scan statistics. *Probability in the Engineering and Informational Sciences*, 17 :509–525.

[Bresenham, 1977] Bresenham, J. (1977). A linear algorithm for incremental digital display of circular arcs. *Communications of the ACM*, 20(2) :100–106.

[Chen and Glaz, 1996] Chen, J. and Glaz, J. (1996). Two-dimensional discrete scan statistics. *Statistics and probability letters*, 31(1) :59–68.

[Chen and Glaz, 1999] Chen, J. and Glaz, J. (1999). Approximations for the distribution and the moments of discrete scan statistics. *Scan Statistics and Applications*, pages 27–66.

[Chen and Glaz, 2002] Chen, J. and Glaz, J. (2002). Approximations for a conditional two-dimensional scan statistic. *Statistics and probability letters*, 58(3) :287 – 296.

[Chen and Glaz, 2009] Chen, J. and Glaz, J. (2009). *Approximations for Two-Dimensional Variable Window Scan Statistics*. Springer.

[Cho, 2008] Cho, J. H. (2008). The genetics and immunopathogenesis of inflammatory bowel disease. *Nat Rev Immunol*, 8(6) :458–66.

[Chouraki et al., 2011] Chouraki, V., Savoye, G., Dauchet, L., Vernier-Massouille, G., Dupas, J. L., Merle, V., Laberenne, J. E., Salomez, J. L., Lerebours, E., Turck, D., Cortot, A., Gower-Rousseau, C., and Colombel, J. F. (2011). The changing pattern of crohn's disease incidence in northern france : a continuing increase in the 10- to 19-year-old age bracket (1988-2007). *Aliment Pharmacol Ther*, 33(10) :1133–42.

[Colombel et al., 1996] Colombel, J. F., Grandbastien, B., Gower-Rousseau, C., Plegat, S., Evrard, J. P., Dupas, J. L., Gendre, J. P., Modigliani, R., Belaiche, J., Hostein, J., Hugot, J. P., van Kruiningen, H., and Cortot, A. (1996). Clinical characteristics of crohn's disease in 72 families. *Gastroenterology*, 111(3) :604–7.

[Conley et al., 2005] Conley, J., Gahegan, M., and Macgill, J. (2005). A genetic approach to detecting clusters in point data sets. *Geographical Analysis*, 37(3) :286–314.

[Cosnes, 2010] Cosnes, J. (2010). Smoking, physical activity, nutrition and lifestyle : environmental factors and their impact on ibd. *Dig Dis*, 28(3) :411–7. Cosnes, Jacques Switzerland Basel, Switzerland Dig Dis. 2010 ;28(3) :411-7. Epub 2010 Sep 30.

[Cosnes et al., 2011] Cosnes, J., Gower-Rousseau, C., Seksik, P., and Cortot, A. (2011). Epidemiology and natural history of inflammatory bowel diseases. *Gastroenterology*, 140(6) :1785–94.

[Cousens et al., 2001] Cousens, S., Smith, P. G., Ward, H., Everington, D., Knight, R. S., Zeidler, M., Stewart, G., Smith-Bathgate, E. A., Macleod, M. A., Mackenzie, J., and Will, R. G. (2001). Geographical distribution of variant creutzfeldt-jakob disease in great britain, 1994-2000. *Lancet*, 357(9261) :1002–7.

[Cox, 1955] Cox, D. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164.

[Cressie, 1991] Cressie, N. (1991). *Statistics for spatial data*. J. Wiley.

[Cuzick and Edward, 1990] Cuzick, J. and Edward, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1) :73–104.

[D'Aignaux et al., 2002] D'Aignaux, J. H., Cousens, S. N., Delasnerie-Laupretre, N., Brandel, J. P., Salomon, D., Laplanche, J. L., Hauw, J. J., and Alperovitch, A. (2002). Analysis of the geographical distribution of sporadic creutzfeldt-jakob disease in france between 1992 and 1998. *Int J Epidemiol*, 31(2) :490–5.

[DeChello and Sheehan, 2007] DeChello, L. M. and Sheehan, T. J. (2007). Spatial analysis of colorectal cancer incidence and proportion of late-stage in massachusetts residents : 1995-1998. *Int J Health Geogr*, 6 :20.

[Declercq et al., 2010] Declercq, C., Gower-Rousseau, C., Vernier-Massouille, G., Salleron, J., Balde, M., Poirier, G., Lerebours, E., Dupas, J. L., Merle, V., Marti, R., Duhamel, A., Cortot, A., Salomez, J. L., and Colombel, J. F. (2010). Mapping of inflammatory bowel disease in northern france : spatial variations and relation to affluence. *Inflamm Bowel Dis*, 16(5) :807–12.

[Demattei et al., 2007] Demattei, C., Molinari, N., Daurès, J.-P., et al. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data. *Computational Statistics and Data Analysis*, 51(8) :3931–3945.

[Diggle, 2003] Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns*. Hodder Arnold.

[Diggle and Chetwynd, 1991] Diggle, P. and Chetwynd, A. G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 47(3) :1155–1163.

[Duczmal and Assuncao, 2004] Duczmal, L. and Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45(2) :269–286.

[Duczmal et al., 2008] Duczmal, L., Cançado, A. L. F., and Takahashi, R. H. C. (2008). Delineation of irregularly shaped disease clusters through multiobjective optimization. *Journal of Computational and Graphical Statistics*, 17(1) :243–262.

[Duczmal et al., 2006] Duczmal, L., Kulldorff, M., and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2) :428–442.

[Duerr et al., 2006] Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee, A. T., Gregersen, P. K., Barmada, M. M., Rotter, J. I., Nicolae, D. L., and Cho, J. H. (2006). A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, 314(5804) :1461–3.

[Dwass, 1957] Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1) :181–187.

[Ekbom et al., 1991] Ekbom, A., Zack, M., Adami, H. O., and Helmick, C. (1991). Is there clustering of inflammatory bowel disease at birth ? *Am J Epidemiol*, 134(8) :876–86.

[Fevre et al., 2001] Fevre, E. M., Coleman, P. G., Odiit, M., Magona, J. W., Welburn, S. C., and Woolhouse, M. E. (2001). The origins of a new trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern uganda. *Lancet*, 358(9282) :625–8.

[Fu and Lou, 2003] Fu, J. and Lou, W. (2003). *Distribution theory of runs and patterns and its applications*. World Scientific Publishing Co.

[Gearry et al., 2010] Gearry, R. B., Richardson, A. K., Frampton, C. M., Dodgshun, A. J., and Barclay, M. L. (2010). Population-based cases control study of inflammatory bowel disease risk factors. *J Gastroenterol Hepatol*, 25(2) :325–33.

[Genin et al., 2013] Genin, M., Duhamel, A., Preda, C., Fumery, M., Savoye, G., Peyrin-Biroulet, L., Salleron, J., Lerebours, E., Vasseur, F., Cortot, A., Colombel, J.-F., and Gower-Rousseau, C. (2013). Space-time clusters of crohn's disease in northern france. *Journal of Public Health*, pages 1–8.

[George et al., 2001] George, M., Wiklund, L., Aastrup, M., Pousette, J., Thunholm, B., Saldeen, T., Wernroth, L., Zaren, B., and Holmberg, L. (2001). Incidence and geographical distribution of sudden infant death syndrome in relation to content of nitrate in drinking water and groundwater levels. *Eur J Clin Invest*, 31(12) :1083–94.

[Glaz, 1992] Glaz, J. (1992). Approximations for tail probabilities and moments of the scan statistic. *Computational Statistics and Data Analysis*, 14(2) :213 – 227.

[Glaz and Balakrishnan, 1999] Glaz, J. and Balakrishnan, N. (1999). *Scan Statistics and Applications*. Statistics for Industry and Technology. Birkhäuser Boston.

[Glaz and Naus, 1991] Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *Annals of applied probability*, 1(2) :306–318.

[Glaz et al., 2001] Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan Statistics*. Springer.

[Glaz et al., 2009] Glaz, J., Pozdnyakov, V., and Wallenstein, S. (2009). *Scan Statistics : Methods and Applications*. Birkhaüser Boston.

[Glaz and Zhang, 2004] Glaz, J. and Zhang, Z. (2004). Multiple window discrete scan statistics. *Journal of Applied Statistics*, 31(8) :967–980.

[Glaz and Zhang, 2006] Glaz, J. and Zhang, Z. (2006). Maximum scan score-type statistics. *Statistics and probability letters*, 76(13) :1316 – 1322.

[Gower-Rousseau et al., 1994] Gower-Rousseau, C., Salomez, J. L., Dupas, J. L., Marti, R., Nuttens, M. C., Votte, A., Lemahieu, M., Lemaire, B., Colombel, J. F., and Cortot, A. (1994). Incidence of inflammatory bowel disease in northern france (1988-1990). *Gut*, 35(10) :1433–8.

[Green et al., 2006] Green, C., Elliott, L., Beaudoin, C., and Bernstein, C. N. (2006). A population-based ecologic study of inflammatory bowel disease : searching for etiologic clues. *Am J Epidemiol*, 164(7) :615–23 ; discussion 624–8.

[Grimson and Rose, 1991] Grimson, R. and Rose, R. (1991). A versatile test for clustering and a proximity analysis of neurons. *Methods Inf Med*, 30(4) :299–303.

[Haiman, 1999] Haiman, G. (1999). First passage time for some stationary processes. *Stochastic Processes and their Applications*, 80(2) :231 – 248.

[Haiman, 2000] Haiman, G. (2000). Estimating the distributions of scan statistics with high precision. *Extremes*, 3(4) :349 – 361.

[Haiman, 2007] Haiman, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *Journal of statistical planning and inference*, 137(3) :821–828.

[Haiman and Preda, 2002] Haiman, G. and Preda, C. (2002). A new method for estimating the distribution of scan statistics for a two-dimensional poisson process. *Methodology and Computing in Applied Probability*, 4 :393 – 407.

[Haiman and Preda, 2006] Haiman, G. and Preda, C. (2006). Estimation for the distribution of two-dimensional discrete scan statistics. *Methodology and Computing in Applied Probability*, 8(3) :373–382.

[Hjalmars et al., 1996] Hjalmars, U. L. F., Kulldorff, M., Gustafsson, G., and Nagarwalla, N. (1996). Childhood leukaemia in sweden : using gis and spatial scan statistic for cluster detection. *Statistics in Medicine*, 15(7-9) :707–715.

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

[Huang et al., 2007] Huang, L., Kulldorff, M., and Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, 63(1) :109–18.

[Huntington and Naus, 1975] Huntington, R. and Naus, J. (1975). A simpler expression for k th nearest neighbor coincidence probabilities. *The Annals of Probability*, 3(5) :894–896.

[Janson, 1984] Janson, S. (1984). Bounds on the distributions of extremal values of a scanning process. *Stochastic processes and their applications*, 18(2) :313–328.

[Jennings et al., 2005] Jennings, J. M., Curriero, F. C., Celentano, D., and Ellen, J. M. (2005). Geographic identification of high gonorrhea transmission areas in baltimore, maryland. *Am J Epidemiol*, 161(1) :73–80.

[Jockel, 1986] Jockel, K. (1986). Finite sample properties and asymptotic efficiency of monte carlo tests. *The annals of Statistics*, 14(1) :336–347.

[Jung et al., 2007] Jung, I., Kulldorff, M., and Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine*, 26(7) :1594–607.

[Jung et al., 2010] Jung, I., Kulldorff, M., and Richard, O. J. (2010). A spatial scan statistic for multinomial data. *Statistics in Medicine*, 29(18) :1910–8.

[Klassen et al., 2005] Klassen, A. C., Kulldorff, M., and Curriero, F. (2005). Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *Int J Health Geogr*, 4(1) :1.

[Kuehl and Loffredo, 2006] Kuehl, K. S. and Loffredo, C. A. (2006). A cluster of hypoplastic left heart malformation in baltimore, maryland. *Pediatr Cardiol*, 27(1) :25–31. Kuehl, K S Loffredo, C A Pediatr Cardiol. 2006 Jan-Feb ;27(1) :25-31.

[Kulldorff, 1997] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6) :1481–1496.

[Kulldorff, 1999] Kulldorff, M. (1999). Spatial scan statistics : models, calculations, and applications. *Scan statistics and applications*, pages 303–322.

[Kulldorff, 2006] Kulldorff, M. (2006). Tests of spatial randomness adjusted for an inhomogeneity. *Journal of the American Statistical Association*, 101(475) :1289–1305.

[Kulldorff, 2011] Kulldorff, M. (2011). Satscan v9.1.1 : Software for the spatial and space-time scan statistics.

[Kulldorff et al., 1998] Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A., and Key, C. R. (1998). Evaluating cluster alarms : a space-time scan statistic and brain cancer in los alamos, new mexico. *Am J Public Health*, 88(9) :1377–80.

[Kulldorff et al., 1997] Kulldorff, M., Feuer, E. J., Miller, B. A., and Freedman, L. S. (1997). Breast cancer clusters in the northeast united states : a geographic analysis. *Am J Epidemiol*, 146(2) :161–70.

[Kulldorff et al., 2009] Kulldorff, M., Huang, L., and Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *Int J Health Geogr*, 8 :58.

[Kulldorff et al., 2006] Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in medicine*, 25(22) :3929–3943.

[Kulldorff and Nagarwalla, 1995] Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters : detection and inference. *Statistics in Medicine*, 14(8) :799–810.

[Kulldorff et al., 2003] Kulldorff, M., Tango, T., and Park, P. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4) :665–684.

[Kulpa, 1979] Kulpa, Z. (1979). On the properties of discrete circles, rings, and disks. *Computer graphics and image processing*, 10(4) :348–365.

[Lawson, 1993] Lawson, A. B. (1993). On the analysis of mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 156(3) :363–377.

[Lawson, 2006] Lawson, A. B. (2006). Disease cluster detection : a critique and a bayesian proposal. *Statistics in Medicine*, 25(5) :897–916.

[Lerebours et al., 2007] Lerebours, E., Gower-Rousseau, C., Merle, V., Brazier, F., Debeugny, S., Marti, R., Salomez, J. L., Hellot, M. F., Dupas, J. L., Colombel, J. F., et al. (2007). Stressful life events as a risk factor for inflammatory bowel disease onset : a population-based case–control study. *The American journal of gastroenterology*, 102(1) :122–131.

[Lerner, 2007] Lerner, A. (2007). Aluminum is a potential environmental factor for crohn's disease induction. *Annals of the New York Academy of Sciences*, 1107(1) :329–345.

[Leroyer et al., 2001] Leroyer, A., Hemon, D., Nisse, C., Bazerques, J., Salomez, J. L., and Haguenoer, J. M. (2001). Environmental exposure to lead in a population of adults living in northern france : lead burden levels and their determinants. *Sci Total Environ*, 267(1-3) :87–99.

[Loader, 1991] Loader, C. R. (1991). Large-deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, pages 751–771.

[Marriott, 1979] Marriott, F. (1979). Barnard's monte carlo tests : How many simulations ? *Applied Statistics*, pages 75–77.

[Michelozzi et al., 2002] Michelozzi, P., Capon, A., Kirchmayer, U., Forastiere, F., Biggeri, A., Barca, A., and Perucci, C. A. (2002). Adult and childhood leukemia near a high-power radio station in rome, italy. *Am J Epidemiol*, 155(12) :1096–103.

[Moller and Waagepetersen, 2003] Moller, J. and Waagepetersen, R. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Taylor & Francis.

[Moran, 1950] Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, pages 17–23.

[Mostashari et al., 2003] Mostashari, F., Kulldorff, M., Hartman, J., Miller, J., and Kulasekera, V. (2003). Dead bird clusters as an early warning system for west nile virus activity. *Emerging infectious diseases*, 9(6) :641.

[Nagarwalla, 1996] Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in Medicine*, 15(7-9) :845–850.

[Naus, 1974] Naus, J. (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association*, 69(347) :810–815.

[Naus, 1965a] Naus, J. I. (1965a). Clustering of random points in two dimensions. *Biometrika*, 52(1/2) :263–267.

[Naus, 1965b] Naus, J. I. (1965b). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60 :493–517.

[Naus, 1966] Naus, J. I. (1966). Power comparison of two tests of non-random clustering. *Technometrics*, 8(3) :493–517.

[Naus, 1982] Naus, J. I. (1982). Approximations for distributions of scan statistics. *Journal of the American Statistical Association*, 77(377) :177–183.

[Neff and Naus, 1980] Neff, N. and Naus, J. (1980). The distribution of the size of the maximum cluster of points on a line. In *IMS Series of Selected tables in Mathematical Statistics*, volume VI. Providence, American Mathematical Society.

[Neyman and Scott, 1958] Neyman, J. and Scott, E. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–43.

[Ng et al., 2012] Ng, S. C., Woodrow, S., Patel, N., Subhani, J., and Harbord, M. (2012). Role of genetic and environmental factors in british twins with inflammatory bowel disease. *Inflamm Bowel Dis*, 18(4) :725–36.

[Openshaw et al., 1987] Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System*, 1(4) :335–358.

[Ozdenerol et al., 2005] Ozdenerol, E., Williams, B. L., Kang, S. Y., and Magsumbol, M. S. (2005). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. *Int J Health Geogr*, 4 :19.

[Patefield, 1981] Patefield, W. (1981). Algorithm as 159 : an efficient method of generating random r× c tables with given row and column totals. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(1) :91–97.

[Patil and Taillie, 2004] Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11(2) :183–197.

[Pedigo et al., 2011] Pedigo, A., Aldrich, T., and Odoi, A. (2011). Neighborhood disparities in stroke and myocardial infarction mortality : a gis and spatial scan statistics approach. *BMC Public Health*, 11 :644.

[Ranta et al., 1996] Ranta, J., Pitkaeniemi, J., Karvonen, M., Virtala, E., Rusanen, J., Colpaert, A., Naukkarinen, A., and Tuomilehto, J. (1996). Detection of overall space –time clustering in a non-uniformly distributed population. *Statistics in medicine*, 15(23) :2561–2572.

[Riis et al., 2007] Riis, L., Vind, I., Vermeire, S., Wolters, F., Katsanos, K., Politi, P., Freitas, J., Mouzas, I. A., O'Morain, C., Ruiz-Ochoa, V., Odes, S., Binder, V., Munkholm, P., Moum, B., Stockbrugger, R., and Langholz, E. (2007). The prevalence of genetic and serologic markers in an unselected european population-based cohort of ibd patients. *Inflamm Bowel Dis*, 13(1) :24–32.

[Sahajpal et al., 2004] Sahajpal, R., Ramaraju, G., and Bhatt, V. (2004). Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. In *Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on*, pages 35–40. IEEE.

[Sankoh et al., 2001] Sankoh, O. A., Ye, Y., Sauerborn, R., Muller, O., and Becher, H. (2001). Clustering of childhood mortality in rural burkina faso. *Int J Epidemiol*, 30(3) :485–92.

[Silva et al., 2009] Silva, I., Assunçao, R., and Costa, M. (2009). Power of the sequential monte carlo test. *Sequential Analysis*, 28(2) :163–174.

[Stone, 2006] Stone, R. (2006). Investigations of excess environmental risks around putative sources : statistical problems and a proposed test. *Statistics in Medicine*, 7(6) :649–660.

[Stoyan and Stoyan, 1994] Stoyan, D. and Stoyan, H. (1994). *Fractals, random shapes, and point fields : methods of geometrical statistics*. Wiley.

[Tango, 1995] Tango, T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, 14(21-22) :2323–2334.

[Tango, 2000] Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in medicine*, 19(2) :191–204.

[Tango and Takahashi, 2005] Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *Int J Health Geogr*, 4 :11. Tango, Toshiro Takahashi, Kunihiko England Int J Health Geogr. 2005 May 18 ;4 :11.

[Tango and Takahashi, 2012] Tango, T. and Takahashi, K. (2012). A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Statistics in Medicine*. Tango, Toshiro Takahashi, Kunihiko Stat Med. 2012 Jul 16. doi : 10.1002/sim.5478.

[Thia et al., 2008] Thia, K. T., Loftus, E. V., J., Sandborn, W. J., and Yang, S. K. (2008). An update on the epidemiology of inflammatory bowel disease in asia. *Am J Gastroenterol*, 103(12) :3167–82.

[Thomas and Carlin, 2003] Thomas, A. and Carlin, B. P. (2003). Late detection of breast and colorectal cancer in minnesota counties : an application of spatial smoothing and clustering. *Statistics in Medicine*, 22(1) :113–27.

[Turnbull et al., 1990] Turnbull, B., Iwano, E., Burnett, W., Howe, H., and Clark, L. (1990). Monitoring for clusters of disease : application to leukemia incidence in upstate new york. *American Journal of Epidemiology*, 132(supp1) :136–143.

[Walter, 1994] Walter, S. (1994). A simple test for spatial pattern in regional health data. *Statistics in Medicine*, 13(10) :1037–1044.

[Walther, 2010] Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics*, 38(2) :1010–1033.

[Weisent et al., 2011] Weisent, J., Rohrbach, B., Dunn, J., and Odoi, A. (2011). Detection of high risk campylobacteriosis clusters at three geographic levels. *Geospatial Health*, 6(1) :65–76.

[Wu et al., 2012] Wu, S., Wu, F., Hong, R., and He, J. (2012). Incidence analyses and space-time cluster detection of hepatitis c in fujian province of china from 2006 to 2010. *PLoS ONE*, 7(7) :e40872.

[Wylie et al., 2005] Wylie, J. L., Cabral, T., and Jolly, A. M. (2005). Identification of networks of sexually transmitted infection : a molecular, geographic, and social network analysis. *J Infect Dis*, 191(6) :899–906.

[Zhang and Glaz, 2008] Zhang, Z. and Glaz, J. (2008). Bayesian variable window scan statistics. *Journal of Statistical Planning and Inference*, 138(11) :3561 – 3567. ¡ce :title¿Special Issue in Honor of Junjiro Ogawa (1915 - 2000) : Design of Experiments, Multivariate Analysis and Statistical Inference¡/ce :title¿.

[Zhang and Kulldorff, 2010] Zhang, Z. and Kulldorff, M. (2010). Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 2010.