Détection d'agrégats temporels et spatiaux

Christophe Demattei

UNIVERSITE MONTPELLIER I
U.F.R. de MEDECINE

Année 2006 N° attribué par la bibliothèque

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER I

Formation Doctorale : Biostatistique

Ecole Doctorale : Information, Structures, Systèmes
Discipline : Mathématiques appliquées et applications des mathématiques

présentée et soutenue publiquement
par

Christophe DEMATTEÏ

Le 20 Novembre 2006

Titre :

DETECTION D'AGREGATS TEMPORELS ET SPATIAUX

Directeurs de thèse : Nicolas MOLINARI
Jean-Pierre DAURES

JURY

M. Avner BAR-HEN	Université Paris 13	Rapporteur
M. Jean-Pierre DAURES	Université Montpellier I	Directeur de thèse
M. Gilles DUCHARME	Université Montpellier II	Examinateur
M. Yann LE STRAT	InVS, Paris	Examinateur
M. Nicolas MOLINARI	Université Montpellier I	Directeur de thèse
Mme Christine THOMAS-AGNAN	Université Toulouse 1	Rapporteur

Remerciements

Je voudrais tout d'abord remercier chaleureusement Nicolas Molinari de m'avoir encadré tout au long de cette thèse. Il m'a témoigné sa confiance en me laissant une grande liberté d'initiative tout en me faisant partager l'étendue de ses compétences, qu'elles soient théoriques ou appliquées. Il a également su me communiquer sa passion pour la recherche. Son soutien a été sans faille, tant sur le plan professionnel qu'humain, et il a su faire preuve d'une grande disponibilité à mon égard. Pour tout cela, je te remercie.

Je tiens également à exprimer mes sincères remerciements à Jean-Pierre Daurès qui a co-encadré cette thèse. Ce fut pour moi un honneur de pouvoir travailler à ses côtés, et ainsi profiter de ses compétences et de son expérience.

Je voudrais aussi exprimer ma profonde gratitude à Christine Thomas-Agnan et Avner Bar-Hen. Malgré un emploi du temps que j'imagine surchargé, ils ont acceptés sans hésiter d'assumer le rôle de rapporteurs. Je leur en suis infiniment reconnaissant et les remercie de l'intérêt qu'ils ont porté à mon travail.

Je tiens enfin à remercier Gilles Ducharme et Yann Le Strat pour avoir spontanément accepté de faire partie de ce jury.

J'ai une pensée pleine de reconnaissance pour toute l'équipe de l'IURC qui m'a chaleureusement accueilli et m'a offert des conditions de travail idéales. Merci à Philippe, Yohann, Yohan, Christel et Vanessa pour votre soutien, vos conseils et pour les bons moments passés ensemble. Merci à Sandy, Françoise, Marie, Christophe C, Christophe B, Murielle, Sylvie, Faïza, Séverine, Pierre, Nadège, Nathalie, Christine, Vivien et Laurent : vous m'avez tous aidés, de près ou de loin, et je vous en suis très reconnaissant.

Sur un plan plus personnel, je tiens à remercier mes parents, ma famille et la famille de Julie pour leur soutien moral et pour toute l'attention qu'ils m'ont témoigné pendant ces trois années. Vous avez toujours cru en moi et je vous dois beaucoup.

Ces remerciements ne seraient pas complets si je ne témoignais pas ma reconnaissance éternelle à Julie. Ton amour, ton soutien inconditionnel, tes encouragements dans les moments difficiles sont pour beaucoup dans ce travail. Je ne te remercierai jamais assez pour tout ce que tu m'as apporté.

Enfin, je dédie cette thèse à ma fille Lilie, dont la venue au monde a illuminé, parfois bruyamment, les derniers mois de ma thèse.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Le terme “cluster” désigne un agrégat inhabituel, réel ou perçu, d'évènements regroupés dans le temps et/ou dans l'espace. Les agrégats (ou clusters) d'évènements de santé, tels que les crises de maladies chroniques, les dates d'occurrence de blessures ou de malformations de naissance, sont souvent reportés aux agences de santé. Quand l'étiologie d'une maladie n'a pas été bien établie, il est parfois requis d'examiner les données pour mettre en évidence un cluster spatial ou temporel et d'établir un lien étiologique avec une exposition. La détection de clusters temporels et/ou spatiaux est utilisée dans plusieurs domaines : la médecine, la cosmologie avec des clusters de galaxies (Szalay et al. [2002]), les sciences sociales et la criminologie (Vinson et Baldry [1999]), l'agronomie et plus encore.

La localisation et la détection d'agrégats d'évènements est une des branches de la statistique en plein développement. Les chercheurs se sont tout d'abord naturellement intéressés au cas de figure temporel, plus simple à traiter car unidimensionnel. Progressivement les méthodes mises en place se sont étoffées, avec la prise en compte de l'évolution de la population à risque par exemple. Ces méthodes ont ensuite été généralisées au cas multidimensionnel, notamment et surtout spatial. Là encore, ces méthodes ont été travaillées afin de tenir compte de la non-homogénéïté éventuelle de la répartition spatiale de la population à risque. Le problème de la détection de clusters spatio-temporels s'est posé plus récemment. Le niveau de complexité des analyses et des interprétations s'accroît une fois de plus avec la prise en compte de dimensions d'unités différentes.

Le perfectionnement des méthodes d'analyse de clusters temporels ou spatiaux a progressivement mis en évidence différents types de tests applicables à différents types de données. Nous pouvons classer les tests en trois grandes catégories. Les tests globaux se contentent de déceler une tendance générale à l'agrégation des données dans le temps ou l'espace sans se soucier de la localisation des clusters éventuels. Dans une toute autre optique, les tests de concentration sont utilisés lorsqu'une information *a priori* permet de pré-spécifier une date ou une coordonnée spatiale, connue pour avoir un lien avec l'incidence de la maladie, et autour de laquelle la recherche d'un agrégat va se focaliser. La troisième catégorie de tests est celle a laquelle nous nous sommes intéressé plus particulièrement dans ce mémoire, aussi bien dans le domaine temporel que spatial. Il s'agit des tests de détection dont le principe est tout d'abord de localiser le ou les agrégats potentiels, puis de tester si ces derniers sont significatifs ou bien simplement le fruit du hasard. Notons que dans le domaine temporel, les

tests de détection sont souvent appelés des tests locaux par opposition aux tests globaux.

Les analyses de clusters peuvent également être classées selon le type de données qu'elles permettent d'étudier. Les deux grandes catégories de données sont définies par leur niveau de résolution. Le domaine d'étude spatial ou temporel analysé est souvent divisé en cellules. On peut citer comme exemples le découpage administratif d'une région géographique et le découpage d'un intervalle de temps en années civiles. Dans ces situations les données, représentées par le nombre d'évènements intervenant dans chaque cellule, sont dites groupées. On parle également de données de comptage. Lorsque la précision des données est plus fine, on parle de données ponctuelles ou encore de données individuelles. Dans ce cas, un évènement peut être défini par ses coordonnées géographiques en spatial, et par la date d'occurrence en temporel. Les deux types de données seront ici utilisées.

Depuis maintenant trois ans, mon travail de recherche au sein de l'équipe du laboratoire de biostatistique de l'IURC s'articule autour de la détection de clusters, principalement spatiaux, mais également temporels. A l'issue de mon travail de synthèse de la littérature sur la détection de clusters spatiaux, deux évidences se sont imposées à moi. Premièrement, la statistique de scan spatiale de Kulldorff [1997] est la méthode de référence dans le domaine et la plus utilisée depuis près de 10 ans. Ensuite, peu de méthodes permettent la détection de clusters de forme arbitraire, quelconque, et surtout aucune de celles qui le permettent ne s'appliquent aux données ponctuelles.

Le principe de la statistique de Kulldorff est de scanner le domaine d'étude en utilisant des cercles et de déterminer celui qui maximise le test du rapport de la vraisemblance. Cette méthode est très puissante et s'applique aussi bien sur des données groupées que ponctuelles. Par contre, et même si elle a été étendue à la détection de clusters de forme elliptique, elle a l'inconvénient de fixer *a priori* la forme des agrégats potentiels. Plusieurs versions non-paramétriques ont été développées afin que les clusters localisés puissent avoir des formes différentes du cercle ou d'une ellipse, mais ces méthodes ne s'appliquent que sur des données groupées. Or, avec le développement de technologies puissantes permettant d'obtenir une résolution spatiale de plus en plus précise, et avec la mise en place des systèmes d'information géographique, il est utile de disposer de méthodes permettant de prendre en compte la précision des données ponctuelles.

Une partie non négligeable de ma thèse a donc été consacrée à la mise au point d'une méthode de détection de clusters spatiaux :

- qui s'applique sur des données ponctuelles,
- qui prend en compte la répartition spatiale de la population à risque,
- qui permet de repérer plusieurs agrégats, s'ils existent,
- et dont la forme n'est pas prédéfinie.

Le principe de base de cette approche est inspiré de celle élaborée par Molinari et al. [2001] dans le cadre temporel : après avoir ordonné les points selon un certain critère (temporel

ou spatial), définissant ainsi une trajectoire, les points regroupés au sein d'un cluster sont successifs sur cette trajectoire et proches (temporellement ou spatialement). Les agrégats potentiels sont repérés par les portions de la trajectoire dont la distance moyenne d'un point à son plus proche voisin est faible. Les portions, délimitées par des points de cassure, sont déterminées par des modèles à changements structurels multiples.

Nous nous sommes également intéressés à la détection de clusters temporels. Nous avons tout d'abord appliqué la statistique de scan temporelle dans le cadre d'une étude des effets des rayonnements ionisants sur l'apparition de cas de cancer. Puis nous nous sommes attachés à contourner l'utilisation de simulations auxquelles la plupart des méthodes de détection de cluster temporels font appel dans la phase d'inférence. Nous avons ainsi adapté l'inégalité de Bernstein [1946] dans le but de calculer une borne supérieure pour la p-valeur des clusters potentiels localisés par la méthode de Molinari et al. [2001]. Enfin, nous avons récemment travaillé sur la prise en compte de données temporelles incomplètes, fréquentes lorsque les données portent sur une longue période de suivi. Ces différents points font l'objet de la première partie qui est divisée en deux chapitres. Le premier se propose de faire un état des lieux des principales méthodes de détection de clusters temporels existantes. Le second présente les contributions originales dans ce domaine.

La seconde partie de cette thèse est elle aussi scindée en deux chapitres. Elle est consacrée au traitement de données spatiales, toujours dans le cadre de la détection de clusters d'évènements. Le premier chapitre présente les différentes versions de la statistique de scan spatiale. Cette dernière est appliquée à l'étude du lien entre rayonnements ionisants et répartition spatiale des cas de cancer. La méthode sur données ponctuelles évoquée plus haut fait l'objet du deuxième chapitre de cette partie. Après une présentation détaillée des aspects théoriques, l'approche est tout d'abord étudiée au travers de simulations, puis illustrée sur différents jeux de données réelles. Cette approche a été implémentée en langage R et a fait l'objet d'un package nommé SPATCLUS mis en ligne sur le site du CRAN, le réseau d'archives complet de R. Ce package est décrit en annexe.

Nous avons également eu l'idée d'appliquer la détection de clusters d'évènements dans un contexte différent de celui des données groupées ou individuelles précédemment définies. Le traitement des données obtenues par Imagerie par Résonance Magnétique fonctionnelle est ainsi abordé dans la troisième partie. La méthode d'analyse standard de ce type bien particulier de données, caractérisées notamment par leur grand nombre, est appelée Statistical Parametric Mapping. Sa description fait l'objet du premier chapitre de cette partie. Cette méthode permet d'obtenir une carte d'activation statistique pour chaque sujet d'un protocole. Les pics d'activation sont alors aisément localisables. Le second chapitre présente l'application de la méthode sur données ponctuelles aux données en trois dimensions obtenues par le regroupement des pics de tous les sujets. Cette approche originale permet de détecter des clusters de pics d'activation communs à tous les sujets.

La dernière partie de cette thèse aborde la détection de clusters spatio-temporels. Après une présentation des quelques méthodes existantes dans le domaine, et une application de la statistique de scan spatio-temporelle ayant permis de mettre en évidence une mini-épidémie dermatologique en Inde, la perspective d'extension de la méthode spatiale sur données ponctuelles au cas spatio-temporel est décrite. Cette partie ne constitue en rien un travail abouti mais se contente de présenter des perspectives possibles dans ce domaine.

Première partie

Détection d'agrégats temporels

Introduction

La question de savoir si des évènements de santé sont agrégés dans le temps a reçu une attention considérable depuis les années soixante. La revue de littérature introductive qui suit se concentre sur les tests qui ont marqué l'histoire de la détection de clusters temporels : le test EMM, le test de Larsen, celui de Grimson et enfin celui de Tango. La statistique de scan temporel, méthode référence, fera l'objet d'une section séparée. Une revue plus complète de la littérature sur le sujet peut être trouvée dans le manuscrit de thèse de Christophe Bonaldi [2003].

L'origine d'une méthodologie rigoureuse pour la détection de clusters temporels provient certainement de Elderer et al. [1964]. Ils ont mis au point une statistique du Chi-2, également appelée EMM (initiales du nom des auteurs) basée sur la division de la période étudiée en intervalles égaux. L'indicateur d'agrégation est défini comme étant le nombre maximal de cas dans un des ces intervalles. Les auteurs ont calculé la probabilité d'observer ce nombre maximal d'évènements sous l'hypothèse de répartition équiprobable des cas dans chacune des cellules. Ce test permet d'analyser conjointement plusieurs séries temporelles, comme par exemple des séries recueillies en différents lieux géographiques : la somme des fréquences maximales des différentes séries, une fois centrée réduite, suit asymptotiquement une loi normale, et son carré suit une loi du Chi-2 à un degré de liberté. Cette approche dépend cependant du choix initial du nombre d'intervalles, ce qui a permis de mettre en avant l'importance de l'unité de temps considérée dans ce type d'analyses. Elle ne permet pas de prendre en compte la variation de la population à risque. Cet inconvénient peut être contourné en considérant plusieurs séries définies sur des périodes où la population peut être considérée comme constante. Par contre, la différence de la taille de la population entre les différentes séries n'a pas d'influence sur le résultat du test.

Le test de Larsen et al. [1973] permet de tester la présence d'un seul cluster (on parle de cluster unimodal). Cette approche est elle aussi basée sur une division de la période d'étude, et permet également l'analyse de plusieurs séries temporelles. Autres points communs : le choix *a priori* du découpage et le biais inhérent à une évolution de la population à risque. La comparaison avec le test EMM s'arrête là puisque le test de Larsen est un test global, qui permet uniquement de déceler une tendance générale à l'agrégation sans la localiser. La statistique de test mesure l'écart entre le rang des intervalles contenant au moins un évènement et le rang de la cellule centrale. Une valeur faible de la statistique dénote donc la

présence d'un cluster central, tandis que des valeurs élevées résultent soit d'une répartition uniforme des cas, soit de la présence de clusters multiples. Les résultats du test de Larsen doivent donc être interprétés avec une grande précaution.

Le test temporel de Grimson [1991] est une adaptation au cas temporel d'une procédure de recherche de clusters spatiaux. Comme précédemment, la période étudiée est découpée en cellules, parmi lesquelles des cellules à haut risque sont définies de façon arbitraire. On peut par exemple utiliser la définition de Larsen en prenant les intervalles présentant au moins un évènement. L'approche de Grimson propose de tester l'hypothèse nulle de répartition aléatoire des cellules à haut risque contre l'hypothèse alternative d'adjacence de ces cellules (qui forment ainsi un cluster). Contrairement aux méthodes précédentes, la généralisation du test à plusieurs séries temporelles permet de déceler si les clusters identifiés sur chaque série ont tendance à apparaître en même temps. Le test de Grimson a également l'avantage de ne pas être influencé par des variations de la population à risque. Par contre les résultats ont l'inconvénient de dépendre fortement de la façon dont les cellules à haut risque sont définies.

Enfin, toujours dans le cas de données groupées par intervalles de temps égaux, Tango [1984] a proposé un indice qui permet de mesurer le niveau d'agrégation des observations. Cet indice s'exprime en fonction des fréquences relatives de cas dans chaque cellule et d'une mesure de proximité entre les cellules. La valeur de la statistique de test est d'autant plus élevée que les évènements sont concentrés dans un nombre restreint de cellules. Sous l'hypothèse de répartition uniforme des cas, la statistique suit asymptotiquement une loi du Chi-2. Signalons que la généralisation de l'indice au cas spatial (Tango [1995]) prend en compte l'inhomogénéïté spatiale de la population à risque, et a donc permis par transposition au cas temporel, d'adapter l'indice initial pour la prise en compte de l'évolution de la population dans le temps. Le test de Tango considère toute l'information disponible dans chaque cellule, ce qui présente un avantage par rapport aux méthodes décrites ci-dessus qui n'utilisent qu'une information binaire (absence/présence ou encore fréquence maximale observée). Le test présente également l'avantage d'être très puissant lorsque le cluster se situe au centre de la période étudiée. Il est moins efficace lorsque les données sont agrégées au début ou à la fin de la période.

Cette partie dédiée à la détection de clusters temporels est constituée de deux chapitres. Le premier se propose de présenter et appliquer deux approches existantes, dont la méthode de référence de la statistique de scan temporel à fenêtres variables. Le deuxième chapitre sera consacré à deux apports originaux qui tentent de régler certains problèmes rencontrés dans ce type d'analyses.

Chapitre 1

Approches existantes

Nous allons illustrer la détection de clusters temporels par la description de deux méthodes, l'une globale à l'instar des tests de Larsen, Grimson et Tango, et l'autre locale, comme le test EMM et le test du scan développé plus loin.

La première a été choisie pour son caractère historique et original. C'est une adaptation d'un théorème de Lucien Le Cam de 1958 sur la division d'un intervalle par des points pris au hasard. Nous avons utilisé les résultats de ce théorème, obtenus sous l'hypothèse de répartition uniforme des points, pour mettre au point un test global. Ce théorème est peu intuitif. C'est la raison pour laquelle nous commencerons par l'expliquer et le détailler. Nous présenterons également des exemples théoriques avant de passer à l'application du théorème au cas concret qui nous intéresse.

La deuxième approche est le test de référence de détection de clusters temporels. La statistique de scan temporel à fenêtres variables est une méthode efficace pour détecter des clusters temporels. Sa version actuelle (Nagarwalla [1996] et Kulldorff et Nagarwalla [1995]) est le fruit de nombreuses évolutions. Elle a été intégrée dans un cadre plus général incluant la statistique de scan spatial par Kulldorff [1997]. Les auteurs utilisent une statistique de scan avec des fenêtres de taille variable, permettant ainsi à la taille des agrégats (longueur de l'intervalle de temps) de ne pas être choisie *a priori*. Le test utilisé est celui du rapport de vraisemblance généralisée de l'hypothèse nulle de distribution uniforme versus l'hypothèse alternative d'agrégation non aléatoire. L'extension de Kulldorff [1997] permet la détection de clusters de maladie dans le cas de populations non homogènes. Cette méthode sera appliquée à la détection de clusters de cancers dans 9 départements français afin d'étudier les effets des rayonnements ionisants.

1.1 Un théorème sur la division d'un intervalle par des points pris au hasard

1.1.1 Cadre général

Pour tout entier $n > 1$, soient $(X_{n,j})_{j=1,\ldots,n-1}$ une suite de variables aléatoires indépendantes de répartition uniforme sur le segment $[0;1]$. Pour $j = 1, \ldots, n$, on définit

$$U_{n,j} = n(X_{n,(j)} - X_{n,(j-1)}) \tag{1.0}$$

où les $X_{n,(j)}$ correspondent à la statistique d'ordre des $X_{n,j}$, avec les conventions $X_{n,(0)} = 0$ et $X_{n,(n)} = 1$. Les $U_{n,j}$ correspondent donc aux intervalles successifs séparant $n-1$ points pris au hasard sur le segment $[0;n]$.

L'objet du théorème de Le Cam [1958] est de caractériser, sous certaines hypothèses, le comportement asymptotique des sommes du type

$$W_n = \sum_{j=1}^{n} g_n(U_{n,j})$$

où $\{g_n\}$ est une suite de fonctions de Baire (fonctions $\mathcal{C}^1$) sur l'intervalle $[0;+\infty]$.

Pour cela il définit

$$S_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (V_j - 1) \text{ et } T_n = \sum_{j=1}^{n} g_n(V_j),$$

où les V_j's sont des variables *i.i.d.* de loi exp(1) (loi exponentielle de paramètre $\lambda = 1$). On peut penser (c'est en tout cas ce que Le Cam affirme) que les distributions asymptotiques de W_n et T_n seront voisines.

1.1.2 Enoncé du théorème

Théorème 1 Supposons que lorsque n tend vers l'infini la distribution $\mathcal{L}[T_n, S_n]$ tende vers une limite $\mathcal{L}[T, S]$.

Alors la fonction caractéristique

$$\varphi_{(T,S)}(t,s) = E[\exp(itT + isS)]$$

de (T, S) est de la forme

$$\varphi_{(T,S)}(t,s) = \omega(t)\ \psi(t,s) \text{ avec } \log\psi(t,s) = -\frac{1}{2}\left(s^2 + 2Bst + C^2t^2\right). \tag{1.0}$$

En outre la distribution $\mathcal{L}[W_n]$ tend vers la loi de fonction caractéristique

$$\varphi_W(t) = E[\exp(itW)] = \omega(t)\exp\left(-\frac{1}{2}\left(C^2 - B^2\right)t^2\right). \tag{1.0}$$

La démonstration, longue et fastidieuse, de ce théorème est donnée dans Le Cam [1958].

1.1.3 Quelques réflexions sur le théorème

La bonne compréhension du théorème et de sa puissance nécessite quelques approfondissements. Il faut notamment comprendre pourquoi la loi exp(1) a été choisie pour les V_j, pourquoi la présence de S_n est nécessaire dans ce théorème, et enfin ce qu'implique le fait que les $U_{n,j}$ soient indicés par n.

1.1.3.1 Explication intuitive du choix de la loi des V_j's

Afin de mieux comprendre le choix de la loi exponentielle de paramètre 1 pour les V_j, nous allons montrer que la loi asymptotique des $U_{n,j}$ est exp(1) (convergence en loi). Il nous faut pour cela montrer que, pour tout $x > 0$ fixé, la fonction de répartition des $U_{n,j}$, notée $F_n(x)$, tend vers celle de la loi exp(1) qui vaut $1 - e^{-x}$.

Les n intervalles $Y_j = X_{n,(j)} - X_{n,(j-1)}$ suivent une loi $\beta(1, n-1)$ (David [1980]) d'espérance $\frac{1}{n}$ et de densité

$$h_n(y) = (n-1)(1-y)^{n-2} I_{[0;1]}(y).$$

En utilisant la formule de changement de variable pour les densités, on en déduit que les $U_{n,j}$ admettent pour densité la fonction

$$f_n(x) = \frac{n-1}{n}\left(1 - \frac{x}{n}\right)^{n-2} I_{[0;n]}(x).$$

Ainsi, on montre facilement par une intégration par partie que

$$F_n(x) = \int_0^x f_n(t)dt = 1 - \left(1 - \frac{x}{n}\right)^{(n-1)} I_{[0;n]}(x).$$

Or

$$\left(1 - \frac{x}{n}\right)^{(n-1)} = e^{(n-1)ln\left(1-\frac{x}{n}\right)} \underset{n\to\infty}{\longrightarrow} e^{-x},$$

et ainsi $F_n(x) \underset{n\to\infty}{\longrightarrow} 1 - e^{-x}$.

1.1.3.2 Explication de la présence de S_n

La raison de la présence de S_n dans la caractérisation de la distribution asymptotique de W_n est troublante au premier abord. Elle est en fait évidente une fois que l'on a remarqué que $\sum_{i=1}^n U_{n,j} = n$ (par construction des $U_{n,j}$), ce qui implique que $\sum_{i=1}^n (U_{n,j} - 1) = 0$. Ainsi si on remplace V_j par $U_{n,j}$ dans l'expression de S_n on obtient 0. Autrement dit, la loi de W_n est tout simplement la loi de T_n conditionnellement à $S_n = 0$, propriété que Le Cam utilise dans la démonstration de son théorème.

1.1.3.3 Application dans un cas particulier : $g_n(x) = \frac{1}{\sqrt{n}}(x-1)$

Dans ce cas particulier, nous avons $T_n = S_n$. D'après le théorème central limite (TCL),

$$\mathcal{L}[T] = \mathcal{L}[S] = \mathcal{N}(0,1)$$

et, puisque $cov(T,S) = var(S) = 1$,

$$\mathcal{L}[T,S] = \mathcal{N}_2(0,V) \text{ où } V = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Autrement dit,

$$\log \varphi_{T,S}(t,s) = \log \varphi_S(t+s) = -\frac{1}{2}\left(s^2 + 2st + t^2\right).$$

L'application directe du théorème nous indique $B = 1$ et $C^2 = 1$. Par conséquent

$$\log \varphi_W(t) = -\frac{1}{2}\left(C^2 - B^2\right)t^2 = 0,$$

ou encore $\varphi_W(t) = 1$, ce qui implique que $W = 0$. Ce résultat est cohérent avec la contrainte

$$W_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_{n,j} - 1) = 0.$$

1.1.4 Applications théoriques

1.1.4.1 Exemple 1

Proposition 1 Considérons $g_n(x) = I_{E_n}(x)$ où E_n est un ensemble mesurable de $[0;+\infty]$ tel que

$$n\int_{E_n} e^{-x}dx \underset{n\to\infty}{\longrightarrow} \lambda.$$

Alors Ł(T_n) et Ł(W_n) tendent vers la loi de Poisson d'espérance λ.

$T_n = \sum_{i=1}^{n} I_{E_n}(V_j)$ est le nombre de V_j qui tombent dans l'intervalle E_n donc

$$P(T_n = k) = C_n^k P(V_j \in E_n)^k P(V_j \notin E_n)^{n-k}$$

$$= \frac{n!}{k!(n-k)!}\left[\int_{E_n} e^{-x}dx\right]^k \left[1 - \int_{E_n} e^{-x}dx\right]^{n-k}.$$

Par hypothèse,

$$\int_{E_n} e^{-x}dx \underset{n\to\infty}{\sim} \frac{\lambda}{n}.$$

Ainsi,

$$P(T_n = k) \underset{n\to\infty}{\sim} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Puisque $\left(1 - \frac{\lambda}{n}\right)^{n-k} \underset{n\to\infty}{\longrightarrow} e^{-\lambda}$ et $\frac{n!}{(n-k)!} \underset{n\to\infty}{\sim} n^k$,

$$P(T_n = k) \underset{n\to\infty}{\longrightarrow} \frac{\lambda^k e^{-\lambda}}{k!}.$$

T_n converge donc en loi vers une distribution de Poisson de paramètre λ. Comme cette loi asymptotique n'a pas de composante Laplacienne (Gaussienne), le théorème de Le Cam nous permet de conclure que W_n a la même distribution asymptotique.

1.1.4.2 Exemple 2

Considérons $g_n(x) = \frac{1}{\sqrt{n}}(x^2 - 2)$.
Rappelons que $S_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (V_j - 1)$ et $T_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (V_j^2 - 2)$.

D'après le TCL, $\mathcal{L}[S_n, T_n]$ tend vers la loi normale $\mathcal{N}_2(0, V)$ où

$$V = \begin{pmatrix} 1 & 4 \\ 4 & 20 \end{pmatrix}$$

En effet :
$E[V_j] = var[V_j] = 1$,
$E[V_j^2] = var[V_j] + E[V_j]^2 = 2$,
$E[V_j^4] = \int_0^{+\infty} x^4 e^{-x} dx = [-x^4 e^{-x}]_0^{+\infty} + 4 \int_0^{+\infty} x^3 e^{-x} dx = 0 + 4 \times \left(0 + 3 \int_0^{+\infty} x^2 e^{-x} dx\right) =$
$12 \times 2 = 24$ donc $var[V_j^2] = 24 - 4 = 20$,
et enfin $cov(S_n, T_n) = cov(V_j - 1, V_j^2 - 2) = E[V_j^3] - E[V_j^2] - 2E[V_j] + 2 = 6 - 2 - 2 + 2 = 4$.

Donc $log\ \varphi(t, s) = -\frac{1}{2}(s^2 + (2 \times 4)st + 20t^2)$. On en déduit que $C^2 = 20$ et $B^2 = 16$ et donc que la fonction caractéristique de W (loi limite de W_n) est $exp(-2t^2)$, ce qui signifie que

$$\mathcal{L}[W_n] \underset{n \to \infty}{\longrightarrow} \mathcal{N}(0, 4).$$

Afin de bien comprendre les mécanismes qui entrent en jeu dans la distribution asymptotique de W_n, commençons par calculer les moments d'ordre 1 et 2 de $U_{n,j}^2$.

$$E[U_{n,j}] = 1 \text{ et } var(U_{n,j}) = \frac{n-1}{n+1} = E[U_{n,j}^2] - E[U_{n,j}]^2.$$

Donc

$$E[U_{n,j}^2] = \frac{n-1}{n+1} + 1 = 2\frac{n}{n+1} \underset{n \to \infty}{\longrightarrow} 2.$$

$$E[U_{n,j}^4] = \int_0^n \frac{n-1}{n} \left(1 - \frac{x}{n}\right)^{n-2} x^4 dx \underset{IPP}{=} \underbrace{\left[-\left(1 - \frac{x}{n}\right)^{n-1} x^4\right]_0^n}_{=0} + 4 \int_0^n \left(1 - \frac{x}{n}\right)^{n-1} x^3 dx.$$

Par le changement de variable $y = x/n$, on obtient

$$E[U_{n,j}^4] = 4n^4 \int_0^1 (1 - y)^{n-1} y^3 dy.$$

On reconnaît à une constante près la densité d'une loi $\beta(4, n)$ dont l'intégrale vaut 1. On obtient alors

$$E[U_{n,j}^4] = 4n^4 \frac{\Gamma(n)\Gamma(4)}{\Gamma(n+4)} = 4n^4 \frac{(n-1)!3!}{(n+3)!} = 24 \frac{n^4}{(n+3)(n+2)(n+1)n} \underset{n \to \infty}{\longrightarrow} 24.$$

On en déduit que

$$var(U_{n,j}^2) = E[U_{n,j}^4] - E[U_{n,j}^2]^2 \underset{n\to\infty}{\longrightarrow} 20.$$

Si les $U_{n,j}^2$ avaient des moments d'ordre 1 et 2 fixes (respectivement égaux à 2 et 20), on pourrait alors appliquer le TCL et la distribution asymptotique de W_n serait la loi $\mathcal{N}(0, 20)$ (celle de T_n). Le fait que les moments dépendent de n et que les $U_{n,j}$ soient liées par la contrainte $\sum_{i=1}^{n}(U_{n,j} - 1) = 0$ rend caduque l'application du TCL, et diminue la variance asymptotique de W_n qui passe de 20 à 4.

Cette constatation est encore plus flagrante dans le cas $g_n(x) = \frac{1}{\sqrt{n}}(x-1)$ précédemment étudié puisque la variance asymptotique passe de 1 pour T_n à 0 pour W_n.

1.1.5 Utilisation en tant que test global de détection de cluster

La fonction étudiée dans le paragraphe précédent, $g_n(x) = \frac{1}{\sqrt{n}}(x^2-2)$, a été utilisée pour illustrer l'application du théorème de Le Cam en tant que test de détection global de clusters temporels. Nous avons vu qu'avec cette fonction la statistique W_n suit asymptotiquement, sous l'hypothèse de répartition uniforme, une loi normale de moyenne nulle et d'écart-type 2. Ce résultat peut être utilisé pour déterminer si une série temporelle d'évènements présente une tendance à l'agrégation, sans pour autant chercher à connaître la fenêtre temporelle où le cluster se situe.

Pour cela, on détermine de façon usuelle la probabilité qu'une variable normale centrée d'écart-type 2 soit supérieure en valeur absolue à la valeur de W_n. Si cette probabilité est inférieure à un seuil α prédéfini, les données sont agrégées. Dans ce qui suit, $\alpha = 0.05$.

1.1.5.1 Etude de puissance

Afin d'étudier la puissance de ce test, nous avons simulé des échantillons de n évènements compris entre 0 et 1 suivant des mélanges de lois uniformes de la forme $\frac{n-n_c}{n} \times \mathcal{U}([0,1]) + \frac{n_c}{n} \times \mathcal{U}(C)$, avec $C = [0.7, 0.8]$. Afin d'étudier différentes situations d'agrégation, nous avons fait varier le nombre total d'évènements n (de 50 à 1000) ainsi que le nombre n_c d'évènements simulés dans C (de 0 à 30% de n). Le rapport entre n et n_c permet de connaître la densité à l'intérieur du cluster simulé comparativement à celle en dehors du cluster. Par exemple, pour $n = 50$, n_c a successivement été pris égal à 0, 5, 10 et 15, ce qui équivaut respectivement à une densité relative γ approximativement de 1, 2, 3.5 et 5. Il est à noter que le cas $n_c = 0$ correspond à la situation de répartition uniforme. 1000 réplicats ont été générés dans chacune de ces situations.

La Table 1.1 présente la puissance du test dans les différentes situations envisagées. Ces résultats sont illustrés par la Figure 1.1.

Nous obtenons un taux acceptable de détection de la présence d'un cluster pour des tailles d'échantillons supérieures à 100 et une densité relative de cas supérieure à 2. Par ailleurs, le risque de première espèce se rapproche de α lorsque la taille d'échantillon aug-

n ╲ γ	1	2	3.5	5
50	0.031	0.053	0.137	0.369
100	0.033	0.071	0.263	0.672
500	0.040	0.214	0.882	1.000
1000	0.049	0.348	0.994	1.000

TAB. 1.1 – Résultats de l'étude de puissance dans l'utilisation du théorème de Le Cam en tant que test global de détection de cluster. La fonction utilisée est $g_n(x) = \frac{1}{\sqrt{n}}(x^2 - 2)$.

mente. L'analyse de ces résultats met en avant la nécessité d'avoir des échantillons de grande taille si l'on veut disposer d'un test puissant.

1.1.5.2 Les données de Knox

Les données de Knox constituent un jeu de données classiquement utilisé pour comparer les méthodes de détection de clusters. Elles seront d'ailleurs utilisées à plusieurs reprises pour illustrer les différentes méthodes temporelles présentées dans ce chapitre.

Ce jeu de données est constitué des dates de naissance de 35 enfants présentant des malformations de l'oesophage et de la trachée, observées dans un hôpital de Birmingham (Royaume Uni) entre 1950 et 1955. La durée totale de la période d'étude était de 2191 jours. Knox [1959] et Weinstock [1981] avec la statistique de scan, puis Nagarwalla [1996] avec la statistique de scan avec fenêtres de taille variable ont utilisé ces données pour illustrer leurs approches (se reporter à la section suivante pour une description de ces méthodes). Molinari et al. [2001] a également analysé ces données avec une méthode de régression pour la détection de clusters multiples (voir section 2.1). Toutes ces méthodes ont trouvé le même cluster de 15 cas dans l'intervalle de temps de 258 jours allant du jour 1233 au jour 1491. L'existence d'un second cluster de 7 cas en 125 jours [2049-2174] est controversée.

Dans cet exemple, $n - 1 = 35$. Notons $y_{n,j}$ la date de l'occurence de l'évènement j en nombre de jours depuis le 1^{er} janvier 1950. La série temporelle est définie sur l'intervalle [0-2191]. La Figure 1.2 représente les délais observés entre 2 évènements successifs. Nous avons ramené cette série au segment $[0, 1]$ en définissant $x_{n,j} = y_{n,j}/2191$ pour $j = 1, \ldots, n - 1$.

Nous avons ensuite calculé les délais $u_{n,j}$, définis dans l'équation 1.1.1, entre 2 évènements successifs lorsque ces derniers sont rapportés au segment $[0, n]$. La somme $w_n = \sum_{j=1}^{n} g_n(u_{n,j})$ calculée sur ces données avec $g_n(x) = \frac{1}{\sqrt{n}}(x^2-2)$ vaut 5.54. Cette valeur est le fractile d'ordre 99.72% d'une loi normale $\mathcal{N}(0, 2)$. On obtient donc $p = 0.0028 < 0.025$ (risque de première espèce bilatéral de 5%) : on peut rejeter l'hypothèse de répartition aléatoire des évènements de la série temporelle de Knox.

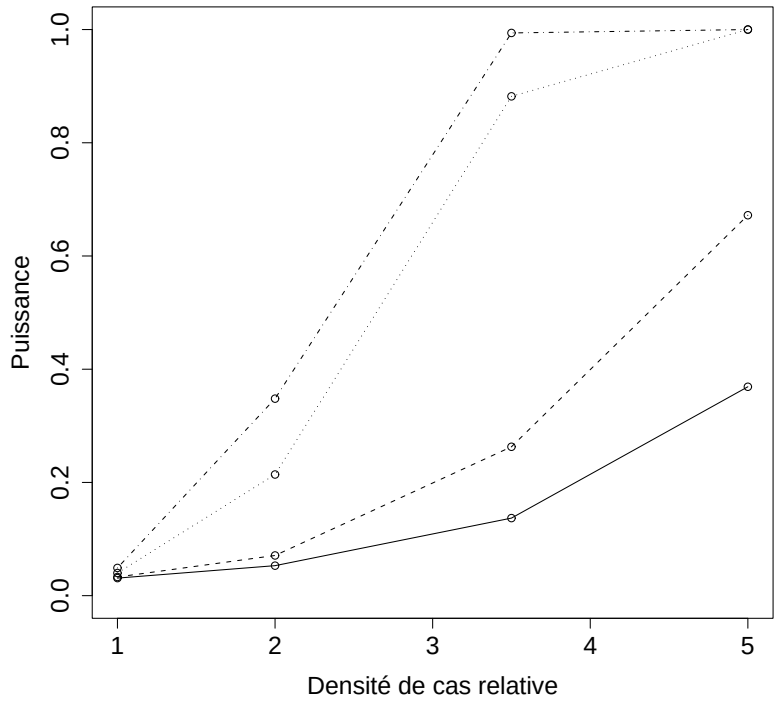


FIG. 1.1 – Résultats de l'étude de puissance dans l'utilisation du théorème de Le Cam en tant que test global de détection de cluster. Pour une taille d'échantillon n donnée, l'évolution de la puissance en fonction de la densité du cluster simulé est représentée par une ligne continue pour $n = 50$, des tirets pour $n = 100$, des pointillés pour $n = 500$ et une ligne mixte pour $n = 1000$.

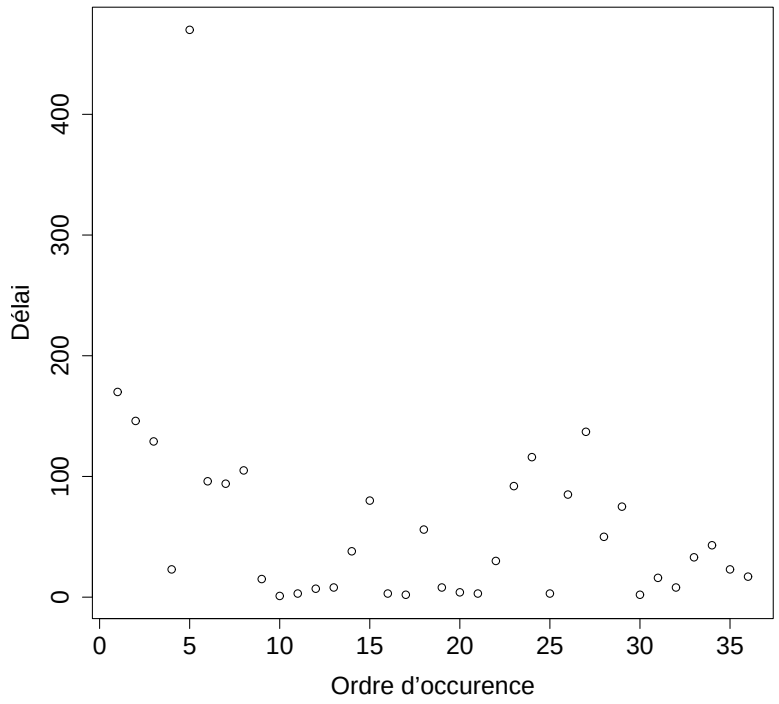


FIG. 1.2 – Délais séparant deux évènements successifs de la série temporelle de Knox en nombre de jours. L'axe des abscisses représente l'ordre d'apparition des évènements.

L'échantillon de Knox est un échantillon de très petite taille en comparaison de la taille minimale requise pour l'application du théorème de Le Cam. Ces résultats sont donc à considérer avec beaucoup de précautions. La conclusion d'agrégation de ces données n'est cependant pas en contradiction avec celles obtenues avec les méthodes précédemment citées.

1.1.6 Commentaires

Malgré la petite taille de la série temporelle de Knox, l'application du théorème de Le Cam en tant que test global a permis de mettre en évidence une tendance générale au regroupement temporel sur des données connues pour être agrégées. L'utilisation de cette approche met cependant en évidence la frustration qui résulte de l'application des tests globaux : l'observation de la Figure 1.2 nous laisse entrevoir la présence d'un cluster entre les $9^{\text{ième}}$ et $22^{\text{ième}}$ évènements observés. Mais avec ce type de test, on ne peut que constater que les données sont agrégées sans pouvoir donner la localisation du ou des regroupements. Les méthodes dites locales, dont la plus connue est présentée dans la section suivante, permettent de remédier à cette limitation.

1.2 La statistique de scan temporel

La statistique de scan a connu de nombreuses améliorations depuis plusieurs décennies. Elle est née des travaux de Naus [1965] qui a calculé la probabilité qu'un sous-intervalle de longueur fixée contiennent n ou plus évènements parmi le nombre total d'évènements. La statistique de scan peut être vue comme une version continue du test EMM de Elderer et al. [1964] que nous avons vu précédemment. Naus [1966] a montré que la statistique de scan était en fait un test de rapport de vraisemblance généralisé de l'hypothèse nulle d'un risque uniforme contre une hypothèse alternative d'un type particulier. Jusqu'alors, l'utilisation de la statistique de scan imposait de choisir *a priori* la taille de la fenêtre scannante. Afin de se rapprocher de la taille de fenêtre optimale, Nagarwalla [1996] a proposé une statistique de scan à fenêtre variable. Enfin, la dernière évolution de la statistique de scan temporel est venue de la détection de clusters dans le cas spatial. Kulldorff et Nagarwalla [1995] ont généralisé la statistique de scan à la dimension 2, puis Kulldorff [1997] a adapté cette approche de façon à prendre en compte l'évolution spatiale ou temporelle de la population à risque. Ainsi, deux des inconvénients majeurs de la statistique de scan initiale, à savoir le choix *a priori* d'une taille de fenêtre et la non prise en compte de l'évolution de la population à risque, ont progressivement été effacés.

La méthode de Kulldorff [1997], bien que proposée dans le cas spatial, s'applique pour autant aussi bien sur des données spatiales que temporelles. Elle permet de prendre en compte les variations temporelles et spatiales de la population à risque. Dans le cas temporel, les données analysées sont les délais d'apparition des cas (depuis le début de l'étude). Dans le cas spatial, ce sont les incidences de cas de cancer dans chaque cellule spatiale (départements ou régions par exemple) pendant la période d'étude qui sont analysées. Dans les deux cas, le modèle utilisé est un modèle de Poisson ou un modèle de Bernouilli. Le premier est

utilisé lorsque la population de fond reflète la population à risque. Le nombre de cas est alors très faible en comparaison à la taille de la population. Le second modèle s'emploie lorsque les données sont de type cas / contrôle (on parle également de données binaires). Typiquement, les données sont dans ce cas constituées d'individus auxquels on attribue le statut de malade ou de non malade. Sous l'hypothèse nulle, le nombre de cas attendus (à chaque temps ou dans chaque cellule spatiale) est proportionnel à la taille de la population à risque. En temporel, la population n'a pas besoin d'être spécifiée de façon continue au cours du temps, mais seulement à un ou plusieurs temps de repère. Pour les temps situés entre ces temps de repère, une interpolation linéaire est effectuée à partir de la population aux temps de repère immédiatement précédent et immédiatement suivant.

1.2.1 Description de la méthode

La statistique de scan spatiale de Kulldorff [1997] sera présentée de manière approfondie dans la partie consacrée à la détection de clusters spatiaux. Nous allons nous contenter ici d'une brève description de cette approche en se concentrant sur le point du vue temporel, ce qui nous donnera une vue générale du principe de la méthode.

Le scan de Kulldorff utilise une fenêtre (ou zone) flexible à la fois pour sa localisation temporelle et pour sa taille (durée) : les dates de début et de fin d'une fenêtre sont variables. On impose généralement une taille maximum pour la fenêtre égale à 50% de la période d'étude. En effet, un cluster d'une taille plus grande indiquerait la présence d'un taux anormalement faible à l'extérieur du cluster plutôt qu'un taux anormalement élevé à l'intérieur.

Pour chaque zone scannante Z, définie par sa localisation et sa taille, l'hypothèse alternative est que le taux de cas est plus élevé à l'intérieur de la fenêtre qu'à l'extérieur. Sous l'hypothèse de Poisson, la fonction de vraisemblance pour une fenêtre donnée est proportionnelle à :

$$\left(\frac{c_Z}{c_e}\right)^{c_Z} \left(\frac{C - c_Z}{C - c_e}\right)^{C - c_Z} I(Z),$$

où C est le nombre total de cas, c_Z est le nombre de cas à l'intérieur de la fenêtre considérée, et c_e est le nombre de cas attendus dans la fenêtre sous l'hypothèse de distribution uniforme. $I()$ est une fonction indicatrice qui vaut 1 lorsque la fenêtre Z a plus de cas que le nombre attendu sous l'hypothèse nulle, et 0 sinon.

Le nombre de cas attendus sous l'hypothèse nulle dans la fenêtre Z est

$$c_e = n_Z \times \frac{C}{N},$$

où n_Z est la taille moyenne de la population à l'intérieur de Z et N est la population moyenne sur l'ensemble de la période étudiée. Dans le cas où une covariable ayant k modalités est utilisée pour ajuster l'analyse, le nombre de cas attendus dans la fenêtre Z s'écrit

$$c_e = \sum_{j=1}^{k} c_e^j = \sum_{j=1}^{k} \left(n_Z^j \times \frac{C^j}{N^j} \right),$$

où n_Z^j, N^j et C^j sont définis comme précédemment mais pour la strate j de la covariable. c_e^j est le nombre de cas possédant la caractéristique j attendus dans la fenêtre Z. Si par exemple l'analyse est stratifiée sur le sexe, le nombre de cas attendus dans une fenêtre Z sera la somme du nombre de cas de sexe féminin attendus dans Z et du nombre de cas de sexe masculin attendus dans Z, ces derniers étant étant calculés à partir du nombre total de cas, de la taille totale de la population, et de la population de Z chez les femmes d'une part, et chez les hommes d'autre part.

La fonction de vraisemblance est maximisée sur toutes les localisations et tailles de fenêtre. Celle qui a la vraisemblance maximum constitue le cluster le plus probable. C'est le cluster qui a la probabilité la plus faible d'être apparu par hasard. La statistique de test utilisée est celle du rapport de la vraisemblance. Sa distribution sous l'hypothèse nulle est obtenue en répétant le même exercice analytique sur un grand nombre de réplicats aléatoires des données générés sous l'hypothèse nulle. La p-value est obtenue par un test d'hypothèse de Monte Carlo (Dwass [1957]), en comparant le rang du maximum de vraisemblance des données réelles aux maximum de vraisemblance des données générées aléatoirement. Si on note ce rang R, alors $p = \frac{R}{1+\#\text{réplicats}}$. Dans ce qui suit, le nombre de réplicats considéré est 9999.

1.2.2 Analyses de clusters de cancers et rayonnements ionisants

Cette section présente les résultats de l'étude **"Analyses de clusters de cancers et rayonnements ionisants (en vue de disposer d'hypothèses de travail de type étiologique)"** qui a été financée par le Service de Radioprotection d'EDF. Elle a été menée en collaboration avec les registres du cancer des départements du Calvados, du Doubs, de l'Hérault, de l'Isère, de la Manche, du Bas-Rhin, du Haut-Rhin, de la Somme et du Tarn. Cette étude se propose de mettre en évidence l'éventuel effet des rayonnements ionisants sur l'apparition de certains cancers, en particuliers les cancers de la thyroïde, les leucémies et les tumeurs cérébrales.

Les études épidémiologiques des effets de rayonnements ionisants sur la santé portent essentiellement sur des populations ayant reçu des radiations dans un contexte particulier. Historiquement, l'étude la plus importante est le suivi de la cohorte des 200000 survivants des bombardements d'Hiroshima et Nagasaki en 1945. D'autres études plus récentes portaient sur des sujets ayant reçu une radiation dans un but thérapeutique ou diagnostique,

ou chez des sujets exposés professionnellement. Hormis ces contextes très particuliers, il est très difficile de mettre en évidence qu'un cancer observé est radio-induit, d'autant plus que les cancers imputables aux radiations sont, d'un point de vue anapathologique, identiques à ceux observés dans la population générale. Il est donc particulièrement intéressant d'essayer de mesurer les effets cancérigènes des radiations sur des populations indirectement exposées. Ce type de problématique est aujourd'hui essentiel car il s'inscrit dans un principe de préoccupation croissant des populations comme des autorités publiques. Nous proposons alors de répondre à cette question par une approche d'analyse de clusters des principaux cancers pouvant être induits par les rayonnements ionisants. Ceci permettra de mettre en évidence des zones ou des périodes à risque qui pourront par la suite être l'objet d'enquêtes étiologiques spécifiques.

Les analyses de détection et de localisation de clusters ont été effectuées dans les cas temporel et spatial. L'analyse de la partie spatiale a nécessité une localisation géographique exacte des communes. Le but de ces analyses était d'étudier la répartition temporelle et spatiale des cas de cancer.

Nous allons ici présenter les résultats des analyses temporelles effectuées avec la "version temporelle" de la statistique de scan de Kulldorff [1997] décrite dans la section précédente. Les données analysées sont décrites. Les résultats obtenus pour chaque type de cancer sont ensuite présentés. Les graphiques illustrant les résultats obtenus sont présentés en Annexe B. Afin de ne pas alourdir inutilement le nombre de pages en annexe seuls les résultats graphiques concernant deux départements ont été reproduits : le Tarn (département "témoin") et La Manche (accueillant deux sites nucléaires). La totalité de ces graphiques sont disponibles dans le rapport d'étude de Dematteï et al. [2005]. Les analyses spatiales seront détaillées dans la section 3.3.

1.2.2.1 Données patients

Les registres des cancers sont des structures chargées de colliger tous les cas de cancer survenant parmi la population résidant dans un territoire donné (un département en général). Un registre peut être soit général (il recense tous les cancers) soit spécialisé (il ne s'intéresse qu'à certains types de cancers). En France métropolitaine il existe 24 registres : 12 registres généraux et 12 registres spécialisés, couvrant un peu plus de 10% de la population. Ils sont regroupés au sein du Réseau français des registres des cancers (FRANCIM) qui a été créé en 1991. Les données recueillies par ces registres permettent notamment d'orienter la politique de prévention, de repérer certains facteurs de risque et d'évaluer les besoins en terme de prise en charge initiale de cette maladie.

Les données patients ont été recueillies par les registres généraux de 9 départements :

3 départements accueillant au moins un site nucléaire : l'Isère (38) avec la centrale de St Alban, le Haut-Rhin (68) avec la centrale de Fessenheim et la Manche (50) avec la centrale de Flamanville et l'usine de La Hague.

6 départements "témoins" : le calvados (14), le Doubs (25), l'Hérault (34), le Haut-Rhin (68), la Somme (80) et le Tarn (81).

Les données concernent 3 types de cancer : cancer du système nerveux central (SNC), cancer hématologique et cancer de la thyroïde.

Les périodes d'études ne sont pas identiques entre les départements. La table 1.2 résume cette information et présente le nombre de communes par département.

Département	Période d'étude	Nb communes
Calvados (14)	1978/1/1 - 1998/12/31	705
Doubs (25)	1978/6/1 - 2001/12/31	594
Hérault (34)	1986/1/1 - 2002/12/31	343
Isère (38)	1979/1/1 - 1999/12/31	533
Manche (50)	1994/1/1 - 1999/12/31	602
Bas-Rhin (67)	1975/1/1 - 1999/12/31	526
Haut-Rhin (68)	1988/1/1 - 1999/12/31	377
Somme (80)	1982/2/1 - 1998/12/31	783
Tarn (81)	1982/4/1 - 1999/12/31	324

TAB. 1.2 – Période d'étude, nombre de communes et d'habitants par département

1.2.2.2 Données temporelles et population des départements

La donnée permettant l'analyse temporelle est la date initiale du diagnostic. Comme nous allons le voir, le jour du diagnostic contient de nombreuses données manquantes. Le mois a donc été choisi comme précision de la date de diagnostic. Les quelques observations ayant un mois de diagnostic manquant n'ont pas été prises en compte dans les analyses.

Les analyses de localisation et détection de clusters temporels nécessitent, pour chaque département, la prise en compte de la taille de la population ainsi que son évolution au cours du temps. Ce sont les résultats des recensements de population de 1975, 1982, 1990 et 1999 qui ont été utilisés (Figure 1.3). Ces chiffres proviennent du Site sur le REcensement en France (le S.RE.F).

Dans ce qui suit nous présentons, pour chacun des trois types de cancer, un récapitulatif sur les données manquantes puis les résultats des analyses de détection de clusters temporels. Les histogrammes illustrant ces résultats sont présentés en annexe B.1.

1.2.2.3 Cancer du SNC

Les données sur le cancer du SNC comportent 7194 observations. Le jour de diagnostique est manquant pour 3378 d'entre elles (47%). 67 observations ayant un mois de diagnos-

Département	1975	1982	1990	1999
Calvados (14)	560967	589559	618468	647520
Doubs (25)	471082	477163	484770	498186
Hérault (34)	648000	709000	795000	896900
Isère (38)	860339	936771	1016227	1091105
Manche (50)	451662	465948	479630	481512
Bas-Rhin (67)	882121	915676	953053	1023638
Haut-Rhin (68)	635209	650372	671319	706225
Somme (80)	538462	544570	547825	556145
Tarn (81)	338024	339345	342741	342364

FIG. 1.3 – Résultats des recensements de 1975 à 1999 pour les différents départements.

tique non renseigné (0.9%) n'ont pas été prises en compte dans les analyses. La Figure 1.4 synthétise ces données par département.

La Figure 1.5 résume les résultats des analyses temporelles pour chaque département obtenus par la méthode du scan temporel. Les dates sont données sous la forme $AAAA/MM/JJ$. Les lignes grisées correspondent à un cluster significatif (p-value $\leqslant 5\%$).

Département	Nombre d'observations	Jour diagniostique manquant	Mois diagnostique manquant	Nombre d'observations utilisées pour l'analyse
Calvados (14)	524	0	0	524
Doubs (25)	667	0	0	667
Hérault (34)	1001	28	28	973
Isère (38)	1992	1992	36	1956
Manche (50)	188	0	0	188
Bas-Rhin (67)	1337	1231	2	1335
Haut-Rhin (68)	550	52	1	549
Somme (80)	542	75	0	542
Tarn (81)	393	0	0	393

FIG. 1.4 – Récapitulatif du nombre d'observations pour le cancer du SNC.

Département	Période d'étude	Cluster potentiel	Nombre de cas	Nombre de cas attendus	p-value
Calvados (14)	1978/1/1 – 1998/12/31	1990/2/1 - 1998/12/31	296	230.08	0.001
Doubs (25)	1978/1/1 - 2001/12/31	1995/10/1 - 1996/2/29	26	11.96	0.189
Hérault (34)	1986/1/1 – 2002/12/31	1996/8/1 – 2002/12/31	506	389.62	0.001
Isère (38)	1979/1/1 – 1999/12/31	1994/8/1 - 1999/12/31	771	526.14	0.001
Manche (50)	1994/1/1 - 1999/12/31	1994/5/1 - 1994/6/30	11	5.17	0.835
Bas-Rhin (67)	1975/1/1 - 1999/12/31	1990/1/1 - 1999/12/31	640	557.07	0.005
Haut-Rhin (68)	1988/1/1 - 1999/12/31	1994/12/1 - 1995/2/28	21	11.34	0.830
Somme (80)	1982/1/1 - 1998/12/31	1991/4/1 - 1991/6/30	21	7.99	0.050
Tarn (81)	1982/1/1 - 1999/12/31	1994/9/1 - 1999/11/30	164	118.45	0.001

FIG. 1.5 – Résultats des analyses temporelles pour le cancer du SNC.

1.2.2.4 Cancers hématologiques

Les données sur les cancers hématologiques comportent 36682 observations. Le jour de diagnostique est manquant pour 13145 d'entre elles (35.8%). 592 observations ayant un mois de diagnostique non renseigné (1.6%) n'ont pas été prises en compte dans les analyses. La Figure 1.6 synthétise ces données par département.

La Figure 1.7 résume les résultats des analyses temporelles pour chaque département obtenus par la méthode du scan temporel.

Département	Nombre d'observations	Jour diagniostique manquant	Mois diagnostique manquant	Nombre d'observations utilisées pour l'analyse
Calvados (14)	2963	160	0	2963
Doubs (25)	3698	0	0	3698
Hérault (34)	4899	308	308	4591
Isère (38)	5805	5805	225	5580
Manche (50)	1303	0	0	1303
Bas-Rhin (67)	6822	6179	19	6803
Haut-Rhin (68)	2869	329	22	2847
Somme (80)	2474	245	1	2473
Tarn (81)	2192	0	0	2192

FIG. 1.6 – Récapitulatif du nombre d'observations pour les cancers hématologiques.

Département	Période d'étude	Cluster potentiel	Nombre de cas	Nombre de cas attendus	p-value
Calvados (14)	1978/1/1 - 1999/12/31	1992/1/1 - 1998/12/31	1266	977.49	0.001
Doubs (25)	1978/1/1 - 2001/12/31	1991/1/1 - 2001/11/30	2367	1740.7	0.001
Hérault (34)	1986/1/1 – 2002/12/31	1997/2/1 – 2002/11/30	2085	1672.78	0.001
Isère (38)	1979/1/1 - 1999/12/31	1992/11/1 - 1999/12/31	2579	1976.33	0.001
Manche (50)	1994/1/1 - 1999/12/31	1997/1/1 - 1999/10/31	716	619.21	0.001
Bas-Rhin (67)	1975/1/1 - 1999/12/31	1990/2/1 - 1999/7/31	3062	2693.85	0.001
Haut-Rhin (68)	1988/1/1 - 1999/12/31	1994/1/1 - 1999/12/31	1731	1444.89	0.001
Somme (80)	1982/1/1 - 1998/12/31	1991/5/1 - 1991/6/30	40	24.44	0.689
Tarn (81)	1982/1/1 - 1999/12/31	1993/3/1 - 1999/7/31	1074	804.29	0.001

FIG. 1.7 – Résultats des analyses temporelles pour les cancers hématologiques.

1.2.2.5 Cancer de la thyroïde

Les données sur le cancer de la thyroïde comportent 4789 observations. Le jour de diagnostique est manquant pour 1742 d'entre elles (36.4%). 61 observations ayant un mois de diagnostique non renseigné (1.3%) n'ont pas été prises en compte dans les analyses. La Figure 1.8 synthétise ces données par département.

La Figure 1.9 résume les résultats des analyses temporelles pour chaque département obtenus par la méthode du scan temporel.

Département	Nombre d'observations	Jour diagniostique manquant	Mois diagnostique manquant	Nombre d'observations utilisées pour l'analyse
Calvados (14)	591	0	0	591
Doubs (25)	541	0	0	541
Hérault (34)	698	39	39	659
Isère (38)	993	993	22	971
Manche (50)	200	0	0	200
Bas-Rhin (67)	776	679	0	776
Haut-Rhin (68)	267	15	0	267
Somme (80)	289	16	0	289
Tarn (81)	434	0	0	434

FIG. 1.8 – Récapitulatif du nombre d'observations pour le cancer de la thyroïde.

Département	Période d'étude	Cluster potentiel	Nombre de cas	Nombre de cas attendus	p-value
Calvados (14)	1978/1/1 - 1998/12/31	1989/4/1 - 1998/12/31	437	283.28	0.001
Doubs (25)	1978/6/1 - 2001/12/31	1991/6/1 - 2001/12/31	373	251.29	0.001
Hérault (34)	1986/1/1 – 2002/12/31	1995/10/1 – 2002/9/30	359	286.75	0.001
Isère (38)	1979/1/1 - 1999/12/31	1995/9/1 - 1999/10/31	398	201.35	0.001
Manche (50)	1994/1/1 - 1999/12/31	1999/3/1 - 1999/5/31	14	8.49	0.993
Bas-Rhin (67)	1975/1/1 - 1999/12/31	1995/9/1 - 1999/12/31	231	142.34	0.001
Haut-Rhin (68)	1988/1/1 - 1999/12/31	1998/1/1 - 1999/9/30	68	39.82	0.005
Somme (80)	1982/2/1 - 1998/12/31	1996/12/1 - 1998/12/31	63	37.02	0.018
Tarn (81)	1982/4/1 - 1999/12/31	1991/3/1 - 1999/12/31	299	221.27	0.001

FIG. 1.9 – Résultats des analyses temporelles pour le cancer de la thyroïde.

1.2.2.6 Conclusion sur l'analyse temporelle

La quasi-totalité des départements présentent un agrégat temporel pour les trois types de cancer. Ces clusters se situent indifféremment dans les départements "témoins" et les départements ASN (Accueillant une Structure Nucléaire) : les cancers du SNC sont agrégés pour deux tiers des départements "témoins" et ASN, et la totalité des départements présentent une agrégation des cancers hématologiques (sauf un département "témoin", la Somme) et des cancers de la thyroïde (sauf un département ASN, La Manche). La répartition temporelle des cas de cancers ne semble donc pas être affectée par les rayonnements ionisants.

Par ailleurs, les clusters détectés se situent tous (à une exception près) à la fin de la période d'étude. Ceci est caractéristique des données temporelles lorsqu'elles présentent une tendance : l'incidence de cas augmente linéairement avec le temps. On peut expliquer ce phénomène par une amélioration de la qualité et de l'exhaustivité du recueil des cas de cancer : pour chaque département, le début de la période d'étude correspond généralement à la création du registre.

Ces résultats nous incitent à effectuer une analyse spatiale portant sur l'ensemble de la période d'étude. Ce regroupement des données temporelles nous permettra d'obtenir une puissance plus élevée pour cette analyse spatiale dont les résultats sont présentés dans la section 3.3.

Notons pour finir que cet exemple met en avant le problème posé par le traitement des données incomplètes (jour et/ou mois non renseignés). Le grand nombre d'observations incomplètes pour le jour du diagnostic nous a tout d'abord contraint à considérer le mois comme précision temporelle, ce qui en soit n'est finalement pas gênant compte tenu de l'étendue de la période totale d'observation. Par contre, face au problème des observations pour lesquelles le mois du diagnostic n'était pas renseigné, nous avons ici adopté la stratégie qui consiste à ne pas inclure ces données dans les analyses. Nous avons donc implicitement supposé une répartition temporelle uniforme de cette censure. L'approche présentée dans la section 2.2 permet de prendre en compte les données incomplètes.

Chapitre 2

Développements récents

Nous proposons ici deux approches que nous avons récemment développées qui tentent de contourner deux inconvénients des méthodes existantes, à savoir l'utilisation de simulations dans la phase d'inférence d'une part et la non prise en compte des données incomplètes d'autre part.

2.1 Calcul de p-valeurs dans la détection de clusters temporels multiples

Comme nous l'avons vu dans la partie précédente, la significativité du test du scan temporel est fournie par des simulations de Monte Carlo. Plus récemment, Molinari et al. [2001] ont proposé une méthode permettant de détecter plusieurs agrégats temporels. Cet algorithme est basé sur une transformation simple des données. Il détermine une fenêtre de temps contenant un excès d'évènements et scanne la période d'étude avec des fenêtres dont la position varie de façon continue. De plus, cette approche est fonctionnelle lorsque la taille de la population à risque varie au cours du temps. La présence d'un ou plusieurs agrégats est déterminée par bootstrap, autrement dit par l'utilisation de simulations comme pour le scan temporel. Ce dernier point constitue la principale faiblesse de ces méthodes.

Nous proposons ici de dépasser cette difficulté en testant la significativité des clusters sans utiliser d'échantillons simulés. Cette approche (Dematteï et Molinari [2006a,b]) est basée sur une application de l'inégalité établie par Bernstein [1946] pour la somme de variables aléatoires indépendantes. Cette inégalité est adaptée à la détection de clusters temporels multiples, ce qui nous permet d'obtenir une borne supérieure de la p-valeur du test de la significativité du ou des clusters.

2.1.1 Localisation des clusters potentiels

Dans cette section nous rappelons la méthode de Molinari et al. [2001]. Elle est basée sur une transformation des données permettant d'obtenir des valeurs qui correspondent au temps (la distance) séparant deux évènements successifs. Sous l'hypothèse H_0 d'absence

de clusters (distribution uniforme), ces valeurs peuvent être estimées par une constante, la distance moyenne. Sous l'hypothèse alternative, un modèle constant par morceaux améliore l'ajustement.

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) représentant les temps d'occurrence de n évènements dans un intervalle $(0, T)$. Sans perte de généralité, nous pouvons considérer que $T = 1$. Supposons que $x_1, \ldots, x_n$ sont tirés au hasard dans l'intervalle unité $(0, 1)$. Notons $x_{(1)}, \ldots, x_{(n)}$ les distances ordonnées séparant chacun de ces points de l'origine. Soient enfin $y_i = x_{(i)} - x_{(i-1)}$ pour $i = 1, \ldots, n$ (par convention $x_{(0)} = 0$).

Considérons maintenant les couples $(k, y_k)_{k=1,\ldots,n}$. Sous l'hypothèse d'absence de cluster, une fonction de régression appropriée pour modéliser ces données est la fonction constante $f(t) = \overline{d} = \dfrac{1}{n}\sum_{k=1}^{n} y_k$.

Si au contraire on désire déterminer la présence de m points de cassure ($m+1$ régimes), la fonction de régression à considérer est $f(t) = \sum_{j=1}^{m+1} \overline{d}_{[n_{j-1}+1;n_j]} \times I_{[n_{j-1}+1;n_j]}(t)$ où $n_1, \ldots, n_m$ sont les m points de cassure, avec la convention $n_0 = 0$ et $n_{m+1} = n$. La notation $\overline{d}_{[i;j]}$ $(1 \leqslant i < j \leqslant n)$ dénote la moyenne de y_t pour t dans $[i; j]$, et $I_{[i;j]}(t) = 1$ si $t \in [i; j]$ et 0 sinon.

Les points de cassure sont estimés par la résolution du problème des moindres carrés sous contraintes

$$\min_{(0<n_1<\ldots<n_m<n)} \quad \sum_{k=1}^{n} \left(y_k - f(k)\right)^2, \tag{2.0}$$

et nous notons $(\hat{n}_1, \ldots, \hat{n}_m)$ la solution. Une méthode permettant de calculer ces estimations de manière efficace est présentée dans Bai et Perron [2003a]. Elle est basée sur une programmation algorithmique dynamique.

L'idée générale de la méthode de Molinari et al. [2001] est basée sur l'hypothèse que les points inclus dans un cluster sont consécutifs et que les distances qui leur sont associées sont plus faibles que celles associées aux points à l'extérieur du cluster (la densité de point est plus élevée à l'intérieur du cluster). Ainsi les clusters potentiels sont localisés par les portions ayant une distance moyenne faible $\overline{d}_{[\hat{n}_{j-1}+1;\hat{n}_j]}$. Dans l'approche originale, le modèle avec m points de cassure est testé versus l'hypothèse d'absence de cluster H_0 en utilisant des réplicats obtenus par bootstrap. Dans la section suivante une nouvelle approche est proposée qui permet de dépasser cette limitation.

Cette approche basée sur l'inégalité de Bernstein est établie à m fixé. Il faut donc au préalable déterminer le meilleur modèle et le nombre de points de cassure de ce modèle peut ensuite être logiquement considéré comme valeur de m . Plusieurs méthodes permettent d'effectuer la sélection du meilleur modèle. Molinari et al. [2001] proposent d'utiliser le critère d'information d'Akaike. Le modèle retenu est celui qui présente la valeur du critère la plus faible. Nous pouvons également utiliser le test du double maximum de Bai et Perron [1998] qui sera présenté en détail dans le chapitre 4. Les deux procédures ont conduit à la sélection des mêmes modèles dans les applications sur données réelles présentées à la fin de cette section.

2.1.2 Inégalités et test

Pour un nombre de points de cassure m donné, nous proposons de tester la significativité pour chaque portion entre deux points de cassure, disons $\hat{n}_k+1$ et $\hat{n}_{k+1}$. Soit $N = \hat{n}_{k+1}-\hat{n}_k$. Afin de simplifier les notations, nous renommons $(Y_i)_{i=1}^{N}$ la série de distances $(Y_i)_{i=\hat{n}_k+1}^{\hat{n}_{k+1}}$.

N est une variable aléatoire qui dépend de m et plus généralement de l'échantillon $X_1, \ldots, X_n$. D'une manière générale, nous ne pouvons pas utiliser N directement, mais nous devons calculer toutes les probabilités conditionnellement à N. Cette difficulté est surmontée lorsqu'une autre réalisation $\tilde{X}_1, \ldots, \tilde{X}_n$ est connue. Le nombre N d'évènements qui tombent dans une portion donnée est alors calculé sur cet autre échantillon, ceci afin de supprimer la dépendance entre N et $X_1, \ldots, X_n$. Dans ce qui suit, nous supposons ceci.

En supposant que les X_i sont i.i.d. uniformes $U(0,1)$, les variables $X_{(1)}, \ldots, X_{(n)}$ sont distribuées suivant n statistiques d'ordre ayant une loi uniforme $U(0,1)$ pour parent. Dans ce cas, $X_{(i)}$ suit une distribution beta $\beta(i, n-i+1)$ et $Y_i = X_{(i)} - X_{(i-1)}$, la distance (le temps) entre les évènements successifs $X_{(i-1)}$ et $X_{(i)}$, a une distribution beta $\beta(1,n)$. Ainsi, l'hypothèse nulle de distribution uniforme peut être écrite H_0 : "*la moyenne des Y_i sur une portion est égale à la moyenne d'une variable distribuée suivant une $\beta(1,n)$*". Pour chaque portion, nous proposons de tester H_0 versus H_1 : "*la moyenne des Y_i est plus petite que la moyenne d'une variable $\beta(1,n)$*" en utilisant les inégalités suivantes. L'hypothèse H_1 dénote la présence d'un cluster sur la portion considérée.

Proposition 2 Soient $(Y_i)_{i=1}^{N}$ des variables aléatoires indépendantes ayant une distribution $\beta(1,n)$ avec $n \geqslant N$. Pour tout $i \in \{1, \ldots, N\}$, définissons $Z_i = (n+1)Y_i$. Soient encore $T = \frac{1}{N}\sum_{i=1}^{N} Z_i$ et $t > 0$. Nous avons alors

$$\mathbb{P}\left(T \leqslant 1 - \frac{t}{N}\right) \leqslant \exp\left(-\frac{t^2}{\frac{2nN}{n+2} + \frac{2t}{3}}\right). \tag{2.0}$$

Preuve. Puisque $Y_i \sim \beta(1,n)$, Y_i est non-négative avec $\mathbb{E}[Y_i] = \frac{1}{n+1}$ et $\mathrm{Var}(Y_i) = \frac{n}{(n+1)^2(n+2)}$. Ainsi, pour $i = 1, \ldots N$, les variables aléatoires $Z_i = (n+1)Y_i$ sont indépendantes, non-négatives, avec $\mathbb{E}[Z_i] = 1$ et $\mathrm{Var}(Z_i) = (n+1)^2\mathrm{Var}(Y_i) = \frac{n}{n+2}$.

Puisque Z_i est non-négative et $\mathbb{E}[Z_i] = 1$, nous sommes en mesure d'appliquer l'inégalité de Bernstein [1946] aux variables aléatoires $1 - Z_i$:

$$\mathbb{P}\left(\sum_{i=1}^{N}(1 - Z_i) - \mathbb{E}\left[\sum_{i=1}^{N}(1 - Z_i)\right] \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^{N}\mathrm{Var}(Z_i) + \frac{2t}{3}}\right). \tag{2.0}$$

De plus, nous avons clairement $\sum_{i=1}^{N}(1 - Z_i) - \mathbb{E}\left[\sum_{i=1}^{N}(1 - Z_i)\right] = N(1 - T)$. Par conséquent (2.1.2) devient

$$\mathbb{P}\left(N(1 - T) \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{\frac{2nN}{n+2} + \frac{2t}{3}}\right),$$

comme voulu. □

La Proposition 2 fournit une borne supérieure pour la probabilité que la moyenne des Y_i soit plus petite qu'un seuil donné.

Le corollaire suivant permet de rendre ce résultat effectif en contrôlant le risque d'erreur de type I, noté α.

Corollaire 1 Sous les hypothèses de la Proposition 2, nous avons pour tout $\alpha \in (0, 1)$

$$\mathbb{P}\left(T \leqslant 1 - \frac{t_\alpha}{N}\right) \leqslant \alpha \text{ avec } t_\alpha = -\frac{\ln(\alpha)}{3} + \sqrt{\left(\frac{\ln(\alpha)}{3}\right)^2 - \frac{2nN\ln(\alpha)}{n+2}}. \tag{2.0}$$

Preuve. La preuve est évidente. □

En fixant l'erreur de type I à α, le Corollaire 1 permet de spécifier le seuil $1 - t_\alpha/N$ associé à cette valeur de α. Ainsi, sous H_0, la probabilité que la moyenne d'une portion de taille N soit sous le seuil est plus faible que α. Si la moyenne est effectivement sous le seuil, nous pouvons rejeter H_0 avec un risque d'erreur de type I plus petit que α, et ceci nous fournit une procédure conservative pour tester la significativité d'une portion donnée. L'application de cette procédure à chaque cluster potentiel permet de détecter plusieurs clusters. Il est important de souligner qu'elle permet d'éviter l'utilisation du bootstrap ou des méthodes de Monte Carlo pour effectuer l'inférence.

Comme résultat dérivé de la Proposition 2, nous donnons la p-valeur correspondant à la distance moyenne observée d'une portion, notée u.

Corollaire 2 Sous les hypothèses de la Proposition 2, nous avons pour tout $u < 1$

$$\mathbb{P}(T \leqslant u) \leqslant p_u \text{ avec } p_u = \exp\left(-\frac{N(1-u)^2}{\frac{2n}{n+2} + \frac{2(1-u)}{3}}\right). \tag{2.0}$$

Preuve. Il suffit d'appliquer (2) avec $t = N(1-u)$ et de remarquer que $t > 0$ puisque $u < 1$. □

Une autre façon d'utiliser la Proposition 2 est de fixer le seuil u dans le Corollaire 2 à la distance moyenne observée d'une portion donnée, ce qui fournit une p-valeur p_u. Si $p_u \leqslant \alpha$, nous pouvons rejeter H_0 au risque α et la portion représente un cluster temporel significatif.

Notons également que le seuil $1 - t_\alpha/N$ est négatif lorsque

$$\frac{1}{3N} + \frac{n}{N(n+2)} > -\frac{1}{2\ln(\alpha)}.$$

Dans ce cas, le seuil ne peut jamais être atteint par la distance moyenne.

2.1.3 Applications

2.1.3.1 Simulations

Nous avons appliqué cette approche avec $\alpha = 0.05$ à un échantillon de 100 temps d'occurrence d'évènements. Les temps ont été simulés par le mélange de lois uniformes $0.5 \times \mathcal{U}(0, 100) + 0.25 \times \mathcal{U}(25, 35) + 0.25 \times \mathcal{U}(60, 80)$. Ce mélange contient deux clusters potentiels. Le premier ($C_1 = [25, 35]$) a une densité élevée et contient N_1 évènements. Le second, appelé $C_2 = [60, 80]$, a une densité deux fois plus faible et contient N_2 évènements. N_1 et N_2 ont été déterminés sur un réplicat de l'échantillon dans le but de supprimer la dépendance entre le nombre d'évènements dans une portion donnée et l'échantillon analysé. La représentation de la régression pour le modèle avec $m = 4$ points de cassure est présentée sur la Figure 2.1. Dans C_1, la moyenne des distances Z_i est plus faible que le seuil et $p_{u_1} = 0.008 < 0.05$. Dans C_2, la moyenne des distances Z_i est supérieure au seuil et $p_{u_2} = 0.11$. Nous obtenons ainsi un cluster significatif (C_1) qui correspond à la portion avec une densité d'évènements élevée.

2.1.3.2 Application aux données de Knox

Les données de Knox sont présentées dans la section 1.1.5.2. Rappelons qu'elles sont constituées de 35 cas observés sur une période de 2191 jours, et que la plupart des méthodes révèlent la présence d'un cluster bien établi entre les jours 1233 et 1491, et d'un autre plus controversé entre les jours 2049 et 2174.

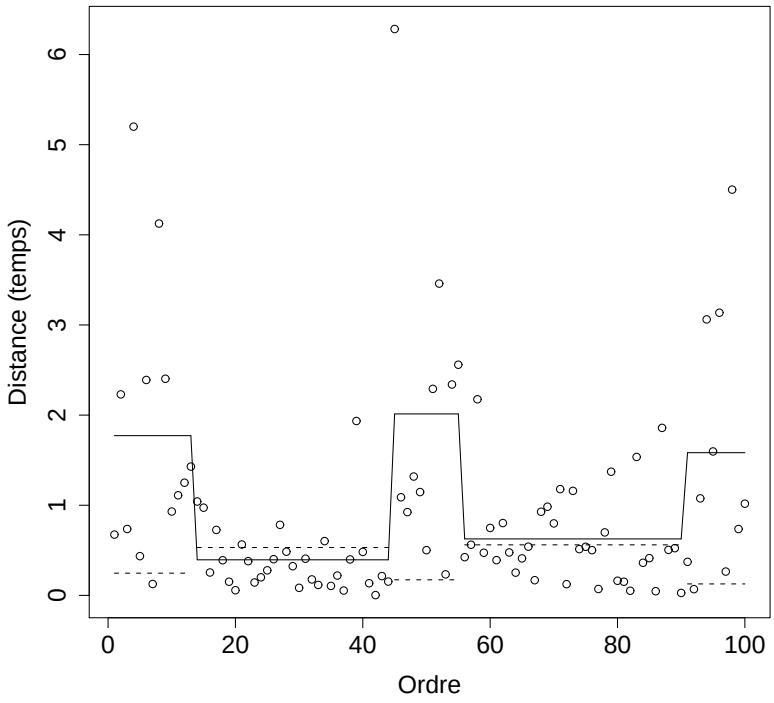


FIG. 2.1 – Résultats de la simulation. C_1 correspond à la seconde portion (qui inclut les ordres 20 et 40), et contient N_1 évènements avec une distance moyenne u_1. C_2 correspond à la quatrième portion (qui inclut les ordres 60 et 80), et contient N_2 évènements avec une distance moyenne u_2. Les lignes en pointillés représentent les seuils $1 - t/N$ calculés pour chaque portion. Pour C_1, le seuil est supérieur à u_1, ce qui signifie que $p_{u_1} < 0.05$ et que C_1 est un cluster significatif. Pour C_2, le seuil est inférieur à u_2, ce qui se signifie que l'on ne peut conclure que C_2 est significatif.

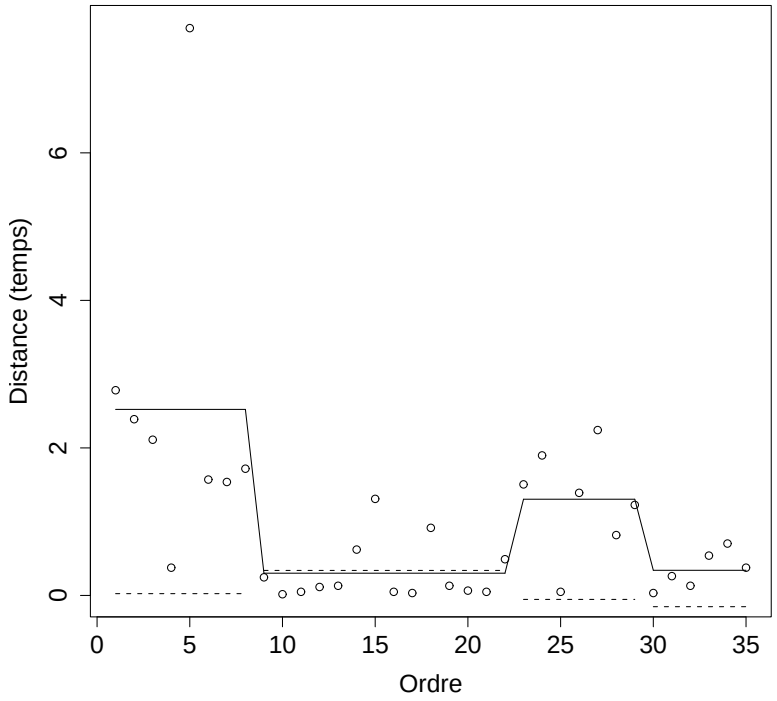


FIG. 2.2 – Données de Knox : représentation de la régression. Les lignes en pointillés représentent les seuils $1 - t/N$ calculés pour chaque portion. Le premier cluster potentiel est significatif.

Le modèle avec 3 points de cassure a été sélectionné, avec deux clusters potentiels. Le premier est le cluster bien établi [1233-1491] avec 15 cas en 258 jours. La distance moyenne est inférieure au seuil et $p = 0.045$. Le second cluster potentiel est [2049-2174]. Ce cluster n'est pas significatif ($p = 0.33$). Ces résultats sont concordant avec ceux obtenus par Molinari et al. [2001] et par l'utilisation de la statistique de scan ($p = 0.019$ pour le premier agrégat).

2.1.3.3 Données d'admission de cas d'hémoptysie au CHU de Nice

Ce jeu de données est constitué de 62 cas d'hémoptysie admis au CHU de Nice du 1^{er} au 31 décembre 1995. Comme cela était proposé par Molinari et al. [2001], la population à risque a été modifiée pour prendre en compte son évolution de 0.72% par an et l'affluence de 55000 touristes en été.

Le modèle avec 2 points de cassure a été sélectionné. Le cluster potentiel est [58; 87] comme dans Molinari et al. [2001]. Cependant, la p-valeur associée à cette portion est 0.13. Notre approche ne permet pas de conclure que ce cluster potentiel est significatif. La statistique de scan a donné le même résultat ($p = 0.29$). Ces deux résultats sont en contradiction avec

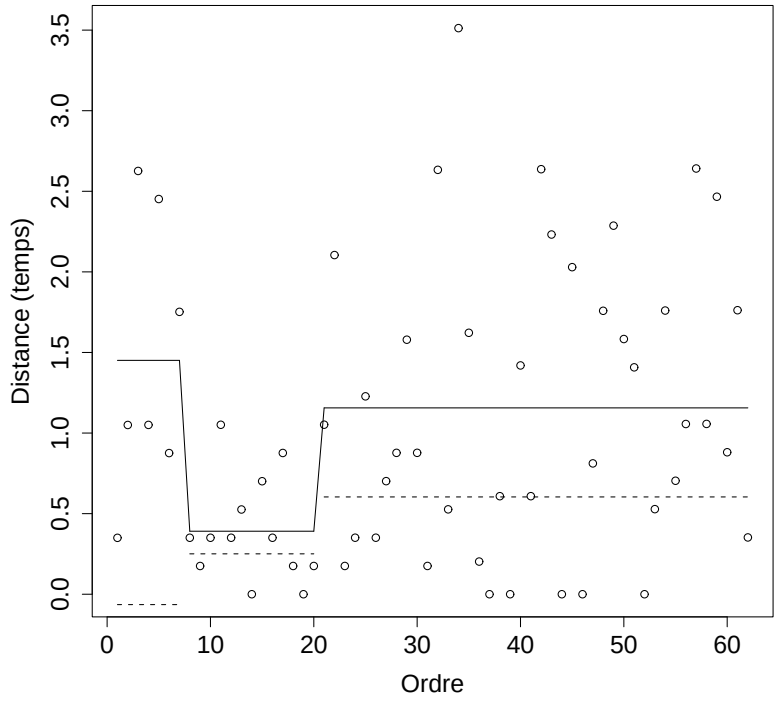


FIG. 2.3 – Donnée d'hémoptysie : représentation de la régression. Les lignes en pointillés représentent les seuils $1 - t/N$ calculés pour chaque portion. Le cluster potentiel n'est pas significatif.

ceux obtenus par Molinari et al. [2001] ($p = 0.02$).

2.1.4 Discussion

Dans cette section, nous avons modifié la méthode originale en évitant l'utilisation de simulations par bootstrap.

La première façon de contourner ces simulations consiste à utiliser le test du double maximum, initialement proposée dans le domaine de l'économie par Bai et Perron [1998]. Cette procédure, conçue pour les changements structurels multiples, est particulièrement utile pour sélectionner le meilleur modèle dans les problèmes de détection de clusters multiples.

Notre approche est une application directe de l'inégalité de Bernstein pour la somme de variables aléatoires bornées. Cette méthode fournit une borne supérieure pour la p-valeur. L'inégalité permet ainsi de détecter des clusters d'une manière conservative. Cette méthode a également l'avantage d'être très flexible. Premièrement, elle peut localiser plusieurs clusters potentiels. Par ailleurs, elle permet de tester l'hypothèse de distribution uniforme pour chaque cluster séparément.

Ce dernier point est très bien illustré sur l'exemple simulé puisque le meilleur modèle contient deux clusters potentiels : la méthode de Molinari et al. [2001] peut seulement tester le modèle dans sa globalité et détecte ainsi soit deux clusters soit aucun cluster, alors que la présente méthode permet d'affirmer qu'un seul des deux clusters potentiels est significatif.

Une perspective de ce travail est l'adaptation de l'inégalité de Bernstein à la méthode de détection de clusters spatiaux multiples de Dematteï et al. [2006a]. Cependant, cette extension à la dimension 2 n'est pas immédiate puisque les variables (correspondant à la distance d'un point à son plus proche voisin) sont faiblement dépendantes dans ce cas (les distances dépendent de la trajectoire déjà effectuées jusqu'à ce point). Des inégalités existent pour la somme de variable faiblement dépendantes. Pour les appliquer au domaine spatial, l'essentiel du travail consistera à caractériser la dépendance entre les variables de distance.

2.2 Détection de clusters temporels sur données incomplètes

Une donnée temporelle incomplète est un évènement dont on sait qu'il est intervenu dans un intervalle de temps mais dont la date exacte est inconnue. On peut par exemple ne connaître que le mois et l'année d'apparition d'un évènement alors que le jour est spécifié pour les autres évènements de la série temporelle. Deux options immédiates sont alors envisageables dans l'optique d'une détection de cluster : supprimer les données incomplètes (en supposant une censure uniforme) ou travailler sur la précision temporelle disponible pour toutes les observations. Ce qui suit a pour objectif de prendre en compte toute l'information temporelle disponible en incluant les dates incomplètes. Ce travail est né d'une collaboration avec le Professeur Antonio Ciampi de l'université McGill à Montréal. Un article est en cours de rédaction.

2.2.1 Problème et notations

Nous présentons ici un exemple afin d'illustrer le problème des données temporelles incomplètes. Les données consistent en 4899 dates de diagnostic de cas de leucémie reportés dans le département de l'Hérault entre le 01/01/1986 et le 31/12/2002 (ce sont les données de cancers hématologiques précédemment analysées dans la partie 1.2.2). Dans la plupart des cas, les dates sont disponibles sous la forme usuelle jour/mois/année. Cependant, pour 308 cas, seule l'année est connue. Les données (groupées par années) sont représentées sur la Figure 2.4. L'objectif est de détecter, s'ils existent, les clusters temporels de cas, c'est à dire une concentration de cas anormalement élevée dans un sous-intervalle de la période d'observation.

La façon la plus simple d'analyser ces données serait, comme cela a été fait dans la partie 1.2.2, de supprimer les 308 dates incomplètes, et de ne travailler que sur les 4591

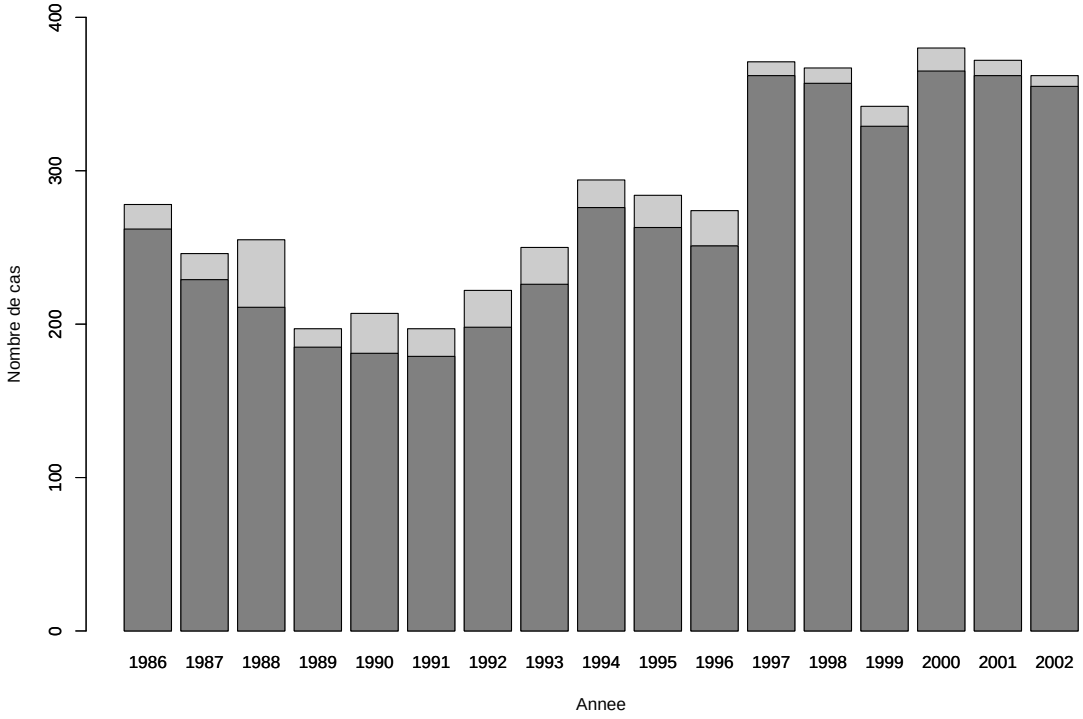


FIG. 2.4 – Histogramme du nombre de cas de leucémies par année. Le nombre de données complètes est représenté en gris foncé, et le nombre de données incomplètes est représenté en gris clair.

dates complètes. Cette stratégie est généralement adoptée lorsque l'hypothèse de répartition uniforme de la censure est admise. La détection de cluster peut alors être efficacement effectuée par la méthode du scan temporel qui permet de prendre en compte l'évolution de la population à risque pendant la période d'étude.

L'application de cette approche à ces données est rapportée sur la Figure 2.5. L'existence d'un cluster à la fin de la période d'observation est évidente, même après ajustement sur l'évolution de la taille de la population représentée par une augmentation linéaire. Cependant, on peut remarquer à partir de la Figure 2.4 qu'il y a plus de dates incomplètes au début qu'à la fin de la période d'étude (par exemple, 17% de données incomplètes en 1988 et 3% en 2002). Ceci peut suggérer que ce cluster pourrait très bien être un artefact dû à des pourcentages de censure différents d'une année à l'autre.

Soient n le nombre total d'évènements observés, n_C le nombre de dates complètement observées, et n_I le nombre de dates incomplètement observées ($n = n_C + n_I$). Nous notons $t_1, \ldots, t_{n_C}$ les dates complètes et $\tau_1, \ldots, \tau_{n_I}$ les intervalles sur lesquels les évènements interviennent pour les dates incomplètes. Dans notre exemple, les limites d'un intervalle sont le premier et le dernier jour de l'année dans laquelle on sait que l'évènement est apparut.

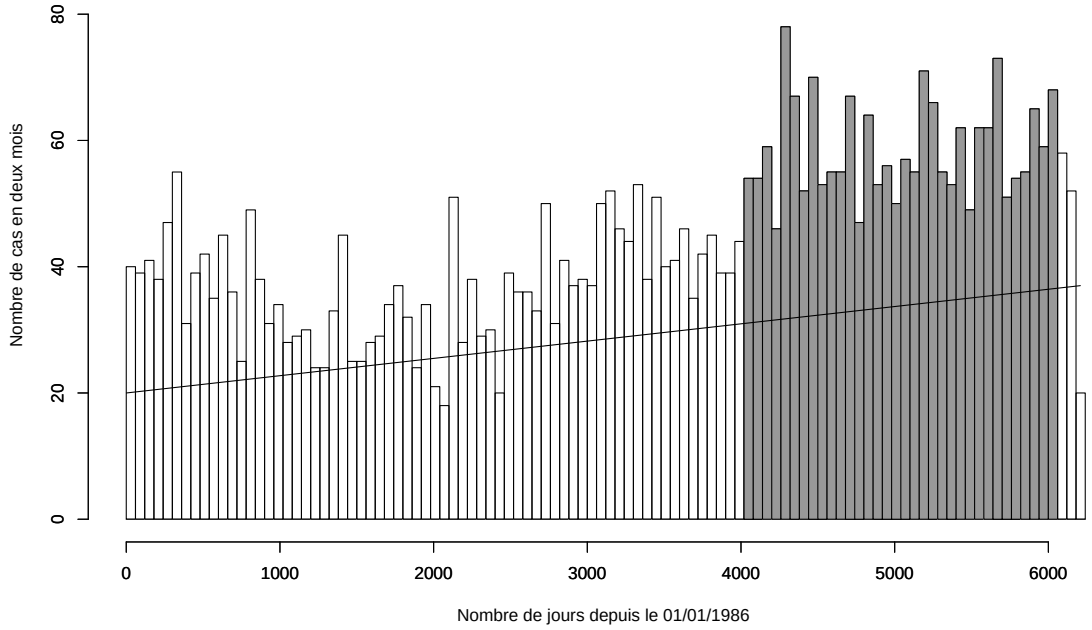


FIG. 2.5 – Histogramme du nombre de cas de leucémies par mois. Le cluster détecté par la statistique de scan temporel est représentée en gris foncé. Seules les données complètes ont été utilisées dans cette analyse. La ligne diagonale représente l'évolution linéaire de la population au cours du temps.

Nous supposons que nos données sont générées par un processus de Poisson non homogène de fonction d'intensité $\lambda(t)$. Alors la contribution à la log-vraisemblance des cas complets est

$$\sum_{i=1}^{n_C} log\lambda\left(t_i\right) - \int_0^T \lambda(t)dt,$$

où T est la borne supérieure de l'intervalle d'observation.

Kulldorff [1997] ne prend en compte que les données complètes. Nous supposons un modèle de Poisson, ce qui est raisonnable lorsque le nombre d'évènements est très petit en comparaison à la taille de la population.

2.2.2 Méthodes

La clé de notre proposition est une approche pour l'estimation de la fonction d'intensité, qui prend en compte les données incomplètes. Une fois que la fonction d'intensité est estimée, nous proposons de l'utiliser pour imputer les données incomplètes et ensuite calculer la statistique de scan aux données complétées. Dans ce qui suit, après une revue de l'approche

de Green [1995] et de la simulation d'évènement temporels à partir de la fonction d'intensité estimée, nous présentons deux approches pour l'imputation.

La méthode de Green permet d'obtenir une estimation $\lambda_1(t)$ du vrai taux $\lambda(t)$ d'occurrence. Cette estimation est appelée le taux moyen d'occurrence postérieur par Green dans son article. L'approche de Green est basée sur un modèle bayésien pour l'analyse de points de rupture multiples et permet de calculer $\lambda_1(t) = E\left[\lambda(t) \mid t_1, \ldots, t_{n_C}\right]$ en utilisant un échantillonneur MCMC (Markov Chain Monte Carlo) à sauts réversibles.

La méthode du "thinning", proposée par Lewis et Shedler [1979], permet de générer n_I observations suivant un processus de Poisson non homogène de taux $\lambda_1(t)$. Soit λ^* tel que $\lambda_1(t) < \lambda^*$ pour tout t. Générons une réalisation w d'un processus de Poisson homogène de taux λ^*. Gardons ensuite w avec la probabilité $\frac{\lambda_1(t)}{\lambda^*}$, sinon supprimons le. Répétons ceci jusqu'à ce que n_I observations soient retenues. Ces n_I observations forment alors une réalisation suivant $\lambda_1(t)$.

Une approche naïve, que nous appellerons imputation simpliste, consisterait à imputer les n_I données incomplètes en simulant simplement une seule fois n_I temps ordonnés suivant l'intensité $\lambda_1(t)$. Cependant, si nous faisions cela, nous ignorerions une importante source de variabilité due à la génération aléatoire. Pour prendre ce dernier point en compte, nous avons développé deux approches. Dans les deux cas nous commençons par générer n_I temps ordonnés K fois.

Dans la première approche, dénommée ici imputation simple, nous extrayons un ensemble unique de n_I temps ordonnés en prenant comme premier temps le mode des K premiers temps, et ainsi de suite. Puis la statistique de scan temporel est calculée une seule fois sur les données complétées en ajoutant cet ensemble unique de n_I temps imputés aux données complètes.

Dans la seconde approche, appelée imputation multiple, nous créons K' jeux de données complets, un pour chaque ensemble de n_I temps ordonnés, et nous calculons sur chacun d'entre eux la statistique de scan temporel, sa p-valeur, et l'intervalle contenant le cluster. A partir de la distribution des bornes inférieures et supérieures de l'intervalle, nous pouvons estimer la variation de la localisation du cluster. De plus, nous pouvons calculer la probabilité qu'un temps donné appartienne au vrai cluster.

Dans ce qui suit, $M = 999$ réplicats ont été générés sous l'hypothèse nulle pour le calcul de la p-valeur de la statistique de scan par la méthode de Monte Carlo.

2.2.3 Applications

2.2.3.1 Données de Knox

Nous avons illustré ces approches sur les données de Knox, présentées dans la section 1.1.5.2.

Sur ce jeu de données, nous rendons 11 observations incomplètes en éliminant une date toute les trois dates. Ainsi, les données consistent en 24 dates complètes (en jours) et 11 cas additionnels sans aucune date mais dont on sait qu'ils sont intervenus pendant la période d'étude. La fonction d'intensité $\lambda_1(t)$ a tout d'abord été estimée à partir des données complètes. Puis les données incomplètes ont été imputées par la procédure du "thinning".

La Figure 2.6 montre 100 estimations de la fonction $\lambda_2(t)$ obtenue en censurant les données une seule fois, puis en complétant les données incomplètes 100 fois par l'imputation simpliste et enfin en estimant la fonction d'intensité pour chaque jeu de données complété de façon simpliste. Cette figure illustre clairement la variabilité ignorée par l'imputation simpliste.

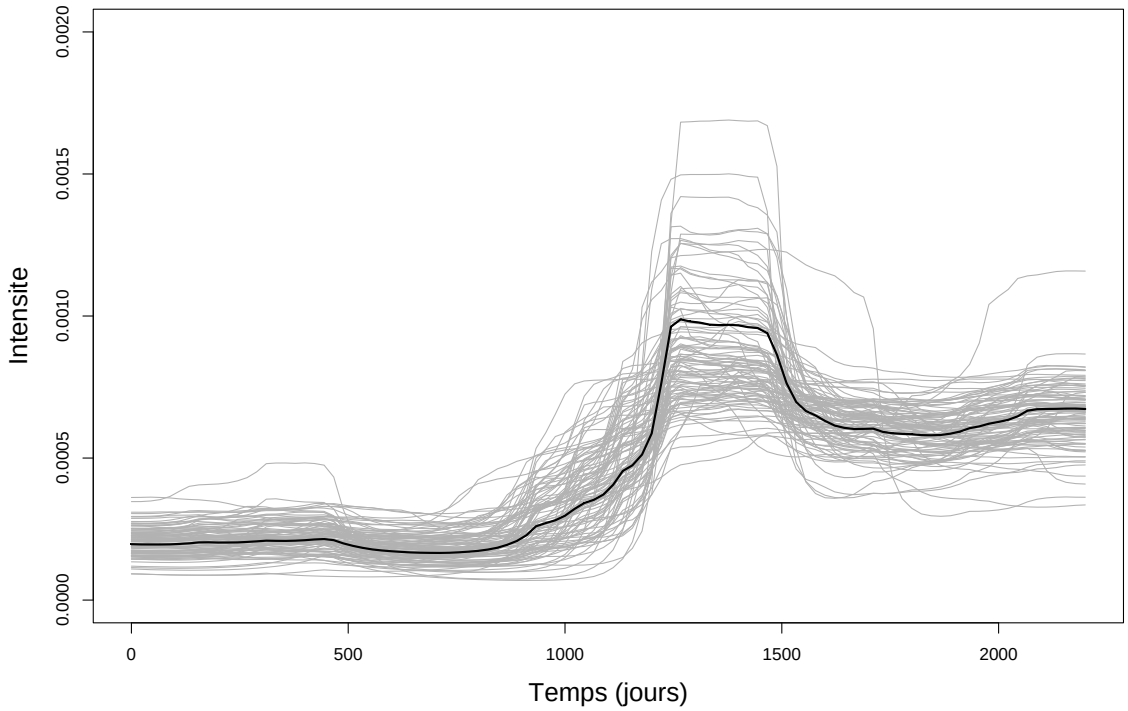


FIG. 2.6 – Illustration de la variabilité de la procédure de "thinning". La courbe noire représente la vraie fonction d'intensité des données de Knox. Chacune des 100 courbes grises représente l'intensité des données complétées, après une censure homogène de 30 pour cent des données.

Nous avons utilisé $K = 10000$ pour l'approche par imputation simple. La Figure 2.7 représente la distribution pour chacune des $n_I = 11$ données imputées (K valeurs ont été générées pour chaque). Les données complétées ont été obtenues en ajoutant le vecteur des n_I modes aux données complètes. Sur ces données complétées, la statistique de scan a détecté un cluster significatif de 28 cas entre les jours 1209 et 2175. Ce cluster regroupe le cluster bien établi $[1233 - 1491]$ et celui plus controversé $[2049 - 2174]$.

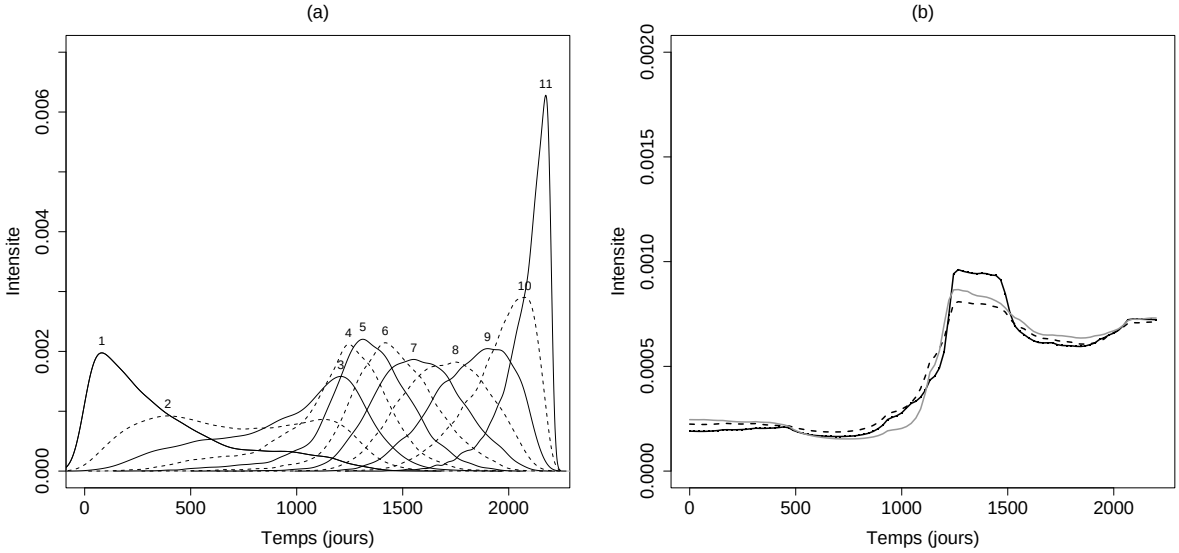


FIG. 2.7 – Illustration de l'approche par imputation simple. **(a) : distribution des données imputées.** La courbe 1 en trait plein (avec un 1 placé au dessus de son mode) représente l'histogramme lissé des $K = 10000$ premiers temps générés. La courbe 2 représente l'histogramme lissé des seconds temps générés, et ainsi de suite. Les courbes sont alternativement représentées en trait plein et en pointillés. **(b) : fonction de taux avant et après imputation.** La courbe noire en trait plein représente la vraie fonction d'intensité des données de Knox, avant censure ($n = 35$). La courbe noire en pointillés représente l'intensité estimée sur les données complètes, après censure ($n_C = 24$). Enfin, la courbe grise est la fonction d'intensité estimée sur les données complétées, après imputation simple ($n = n_C + n_I = 35$).

$K' = 1000$ jeux de données complètes ont été utilisé pour l'approche par imputation multiple. La statistique de scan a été appliquée sur chaque jeu de données. La Figure 2.8 (a) représente le nombre de fois qu'un temps tombe dans l'intervalle contenant le cluster le plus probable, ainsi que la distribution des bornes inférieures et supérieures de cet intervalle. La Figure 2.8 (b) représente la distribution des p-valeurs de la statistique de scan.

Le nombre de fois qu'une date tombe dans le cluster le plus probable peut être vu comme la probabilité pour une date d'appartenir au vrai cluster. Le cluster bien connu $[1233-1491]$ correspond précisément à l'ensemble des temps qui ont une probabilité supérieure à 0.8

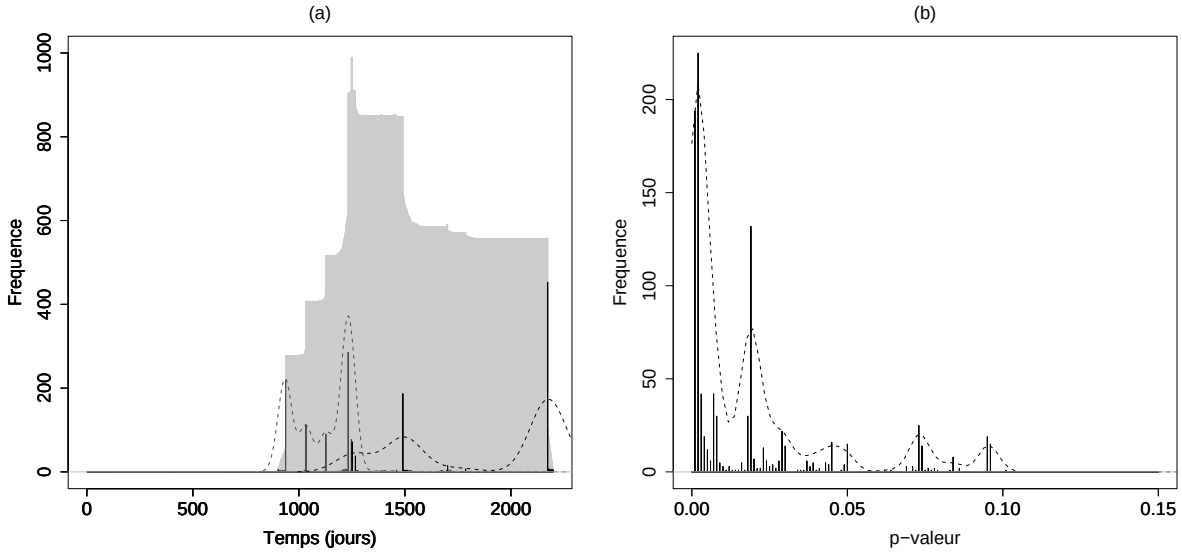


FIG. 2.8 – Illustration de l'approche par imputation multiple. **(a) : localisation du cluster.** L'histogramme du nombre de fois qu'un temps tombe dans le cluster est représenté en gris clair. En gris foncé : la courbe en pointillés représente l'histogramme lissé des $K' = 1000$ bornes inférieures du cluster, et les lignes verticales représentent leur histogramme exact. En noir : la courbe en pointillés représente l'histogramme lissé des K' bornes supérieures, et les lignes verticales représentent leur histogramme exact. **(b) : p-valeurs.** La courbe en pointillés est l'histogramme lissé des K' p-valeurs de la statistique de scan. Les lignes verticales représentent leur histogramme exact.

d'appartenir au vrai cluster. Par ailleurs, 90% des clusters potentiels sont significatifs ($p \leqslant 0.05$).

2.2.3.2 Jeu de données des leucémies

Dans cet exemple illustratif, l'approche simpliste a été appliquée par année : l'intensité $\lambda_1(t)$ a été estimée à partir des dates complètes pour chaque année, et les dates incomplètes ont été imputées suivant cette intensité. Ensuite, la statistique de scan a été appliquée aux données complétées composées des dates complétées pour toutes les années. Le cluster le plus probable, représenté sur la Figure 2.9 est localisé au début de la période d'observation - [1986/2/1 − 1988/5/31] - et est significatif ($p \leqslant 0.001$). Ce résultat illustre très bien l'effet de la distribution non homogène des dates incomplètes. En effet, avant l'imputation, le cluster était localisé à la fin de la période d'étude (voir la Figure 2.5) tandis que le pourcentage de dates incomplètes était plus élevé au début (voir la Figure 2.4). La localisation du cluster a été complètement modifiée en imputant les dates incomplètes.

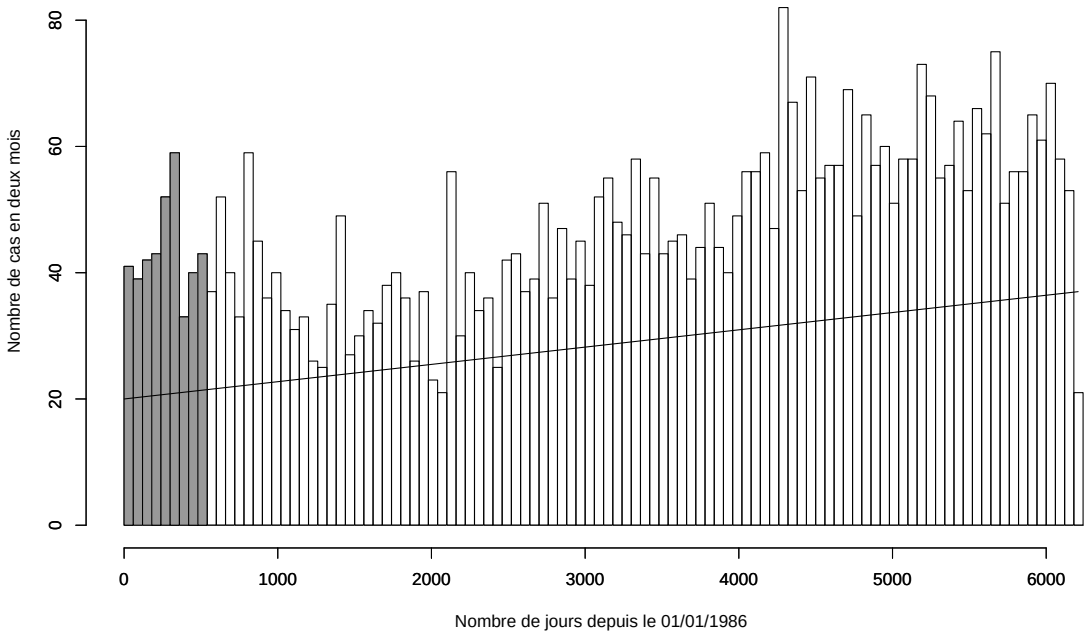


FIG. 2.9 – Histogramme du nombre de cas de leucémies par mois. Le cluster détecté par la statistique de scan temporel est représentée en gris foncé. Les dates incomplètes ont été imputées dans cette analyse. La ligne diagonale représente l'évolution de la population au cours du temps.

2.2.4 Discussion

Nous avons pu constater sur l'exemple des leucémies qu'en cas de censure non-uniforme, ne pas prendre en compte l'information, même incomplète, de toutes les données pouvait conduire à une conclusion erronée. Les approches par imputation simple et multiple proposées ici permettent d'étendre la méthode de référence, la statistique de scan à fenêtres variables, à la détection de clusters temporels en présence de données incomplètes.

Nous avons ici choisi d'estimer la fonction de taux et d'imputer les données incomplètes suivant cette fonction. Une approche plus classique de tenir compte de la présence de données partiellement censurées aurait été d'utiliser un algorithme de type EM (Dempster et al. [1977]). Il serait intéressant de comparer les résultats obtenus par l'application de cet algorithme avec ceux que nous avons obtenus ici.

Discussion

La détection de clusters a tout naturellement débuté dans le cas unidimensionnel. Nous venons de voir que la détection de clusters temporels a connu une évolution spectaculaire depuis les années 50 pour aboutir à la méthode bien établie du scan temporel à fenêtres variables avec prise en compte de la variation de la population à risque.

Nous avons tenté d'apporter une petite pierre à cet édifice en proposant deux approches permettant, pour l'une, d'éviter l'utilisation de simulations lors de l'étape d'inférence, pour l'autre, d'intégrer dans les analyses des données dont l'information temporelle est partielle.

La partie suivante traite du cas de figure bidimensionnel. L'analyse des données spatiales a été abordée par les différents auteurs en partant des résultats obtenus dans le domaine temporel.

Deuxième partie

Détection d'agrégats spatiaux

Introduction

Les nombreux tests proposés dans le but de savoir si les évènements sont agrégés dans l'espace peuvent être classés selon leur objectif. Celui des *tests globaux* ou *généraux* (Ripley [1977]; Whittemore et al. [1987]; Cuzick et Edwards [1990]; Besag et Newell [1991]; Tango [1995, 2000]) est d'analyser la tendance globale à l'agrégation de l'incidence d'une maladie dans une région d'étude, en ne prêtant aucune attention à la localisation du cluster. Avec les *tests de détection* de clusters (Turnbull et al. [1990]; Kulldorff et Nagarwalla [1995]; Kulldorff [1997, 1999a]), les agrégats potentiels peuvent à la fois être localisés et leur significativité être testée. Finalement, les *tests de concentration* (Cuzick et Edwards [1990]; Besag et Newell [1991]; Diggle et al. [1999]; Allard et Fraley [1997]) sont utilisés lorsqu'une localisation pré-spécifiée est supposée être liée à l'incidence de la maladie étudiée.

Kulldorff [2002] a récemment proposé un cadre général qui regroupe une grande majorité des tests. En utilisant ce cadre, une définition mathématique peut être attribuée à chacune des trois classes de tests. Etant donné que parmi la grande variété de tests existants, beaucoup d'entre eux sont identiques, il était important de disposer de critères permettant de distinguer les tests différents. Le cadre général rend cette distinction possible. Chaque test (unique) peut être précisément identifié par la définition de plusieurs éléments : un ensemble de centroïdes (chaque localisation de cas et/ou individus de la population par exemple), des aires autour d'eux (type, forme, taille et distance utilisée), une mesure devant être calculée pour chaque aire (nombre de cas en excédent ou fonction de vraisemblance par exemple), une quantification résumant les mesures correspondant aux aires de taille différente mais partageant le même centroïde, et des aires centrées autour de centroïdes différents (sommation en ligne ou pondérée, ou maximum). De plus, ce cadre permet de créer de nouveaux tests en choisissant une combinaison encore non utilisée de ces éléments de définition. Tous les tests proposés dans la littérature ne rentrent pas dans ce cadre, et c'est justement le cas de la méthode qui sera présenté à la fin de cette partie.

Parmi les méthodes de détection de clusters, la *statistique de scan spatial* (Kulldorff et Nagarwalla [1995]; Kulldorff [1997]) est devenue la plus populaire. Le principe de cette méthode est de balayer l'aire d'étude en utilisant des fenêtres de balayage ayant une forme prédéfinie (généralement des cercles) et de déterminer ensuite celle qui regroupe un nombre anormalement élevé de cas en utilisant le test du rapport de la log-vraisemblance. L'ensemble de ces fenêtres (zones candidates pour l'agrégat le plus probable) représente l'espace

de paramètres réduit et est noté Ω_0.

La statistique de scan spatial est très puissante, notamment lorsque le vrai cluster est de forme circulaire. Cependant, cette méthode a une puissance faible pour détecter des agrégats de forme irrégulière. Ceci est dû à l'utilisation de fenêtres de balayage circulaires (le logiciel *SaTScan*, désormais largement utilisé est disponible gratuitement - Kulldorff et Information Managements Services [2004]). Cette méthode a été généralisée pour la détection de clusters de forme elliptique par Kulldorff et al. [2006]. Trois travaux récents de Patil et Taillie [2004], Duczmal et Assunção [2004] et Tango et Takahashi [2005] sur la détection d'agrégats de forme arbitraire ont été proposés dans le but de dépasser cette limitation de la statistique de scan spatial. Ces trois méthodes ont le même principe de base : appliquer la statistique de scan sur un espace de paramètres réduit Ω_0 qui ne soit pas restreint aux clusters de forme régulière. Dans les trois cas, la significativité des agrégats est obtenue par l'intermédiaire de simulations de Monte Carlo.

La méthode de Patil et Taillie [2004], connue comme la *statistique de scan ULS*, suggère une nouvelle approche permettant de réduire la liste des zones candidates Z. Dans ce but, $\Omega_0 = \Omega_{ULS}$ est composé de tous les sous-ensembles connectés de niveau supérieur (Upper Level Set) d'une surface constante par morceaux définie sur la tessélation (subdivision en cellule du domaine étudié) par les taux ajustés (nombre de cas / population de la cellule). Ω_{ULS} a la structure d'un arbre et est données-dépendant, ce qui implique de le calculer à nouveau pour chaque réplicat des données lors de la simulation de la distribution sous l'hypothèse nulle. La taille de Ω_{ULS} correspond au nombre de noeuds dans l'arbre et ne dépasse pas le nombre de cellules dans la tessélation.

Duczmal et Assunção [2004] ont développé une autre méthode, que nous appellerons la *statistique de scan SA*, et qui ne restreint pas le cluster à avoir une forme géométrique fixée. L'espace de paramètres Ω_0 permet aux agrégats potentiels d'être n'importe quel sous-ensemble de cellules adjacentes. Dans le but d'analyser seulement les sous-ensembles les plus prometteurs, une stratégie de recuit simulé (Simulated Annealing) est utilisée. La routine de surveillance de base est répétée plusieurs fois (jusqu'à ce que 99% de toutes les cellules aient été visitées au moins une fois) avec différent sous-ensembles de départ : le cluster trouvé par la statistique du scan spatial et plusieurs cellules seules choisies aléatoirement.

Enfin, la statistique de scan spatial flexible de Tango et Takahashi [2005] modifie l'ensemble des fenêtres devant être scannées en ajoutant les régions connectées (voisines) aux fenêtres scannantes circulaires. Ce test a une puissance plus élevée que la statistique de scan circulaire lorsque le vrai cluster est non-circulaire. Les auteurs ont précisé que la statistique de scan spatial flexible ne peut être appliquée qu'à des données de comptage.

Toutes les méthodes présentées ci-dessus ont été conçues pour des données de comptage, c'est à dire pour lesquelles les cas sont disponibles à un niveau agrégé, défini par des

cellules de comptage. Même si la statistique de scan basée sur des cercles est également appliquable à des données individuelles, ce n'est pas le cas pour les statistiques de scan récentes permettant de détecter des agrégats de forme arbitraire, puisque le critère d'adjacence des cellules n'est pas défini pour ce type de données. Comme cela a été expliqué par Lawson [2001], l'utilisation des données individuelles a des avantages et des inconvénients non négligeables. D'un côté, la relation entre une localisation exacte et l'étiologie de la maladie peut être incertaine (maladie liée au travail, ou déménagement des cas). La localisation exacte, souvent donnée par l'adresse, n'est pas toujours disponible étant donné de possibles problèmes de confidentialité. D'un autre côté, ce type de données fournit une information spatiale détaillée qui pourrait être perdue lorsque des données de comptage sont utilisées. Dans ce sens, les données de comptage sont une approximation des données individuelles. Ainsi, si les données individuelles sont disponibles, l'auteur recommande de ne pas perdre d'information spatiale en utilisant des données de comptage, et d'analyser ce niveau individuel de résolution des données.

Il nous faut également aborder le problème de la non-homogéneïté de la population sous-jacente. En effet, avec les évènements rares, un domaine d'étude vaste est nécessaire pour examiner l'agrégation spatiale des données. Le problème est que, dans ce cas, la population sous-jacente évolue dans l'espace. Etant donné une non homogéneïté naturelle de l'espace, la population sous-jacente n'est pas constante. Le nombre d'évènements suit la même règle. La statistique de scan spatial est calculée en utilisant la taille de la population dans chaque région et prend ainsi en compte la non homogénéïté dans la population sous-jacente. Nous verrons que la méthode présentée dans ce chapitre ajuste elle aussi les calculs sur une éventuelle non-homogéneïté de la population.

Cette partie est divisée en deux chapitres. Le premier décrit les différentes versions existantes de la statistique de scan spatial et illustre l'application du scan circulaire sur des données de cancer. La méthode présentée ensuite dans le chapitre 4 traite des évènements ponctuels de $\mathbb{R}^2$, telles que les coordonnées spatiales de l'occurrence de cas de maladie ou les positions géographiques d'individus. Cette nouvelle approche permet de localiser et détecter des clusters spatiaux multiples de forme arbitraire. Elle est basée sur une transformation des données et sur un modèle de régression. Ce nouveau test appartient à la classe des tests de détection pour données individuelles.

Chapitre 3

La statistique de scan spatial et ses dérivées

La statistique de scan a été adaptée au cas multidimensionnel (spatial) par Kulldorff et Nagarwalla [1995]. Elle était jusqu'alors essentiellement utilisée pour tester la présence de clusters dans un processus de points unidimensionnel (Naus [1965], Wallenstein [1989], Weinstock [1981]). Dans cette première version spatiale, les fenêtres scannantes envisagées sont, pour des raisons calculatoires, de forme circulaire ce qui confère le caractère paramétrique à la méthode. Cette statistique de scan spatial a ensuite été étendue par Kulldorff [1997] puis par Kulldorff et al. [2006]. La première amélioration porte sur la prise en compte d'une éventuelle non-homogéneïté de la population de fond. La seconde étend la statistique de scan spatial à la détection de clusters de forme elliptique.

L'un des principaux inconvénients des versions paramétriques de la statistique de scan spatial est son manque de flexibilité concernant la forme du cluster localisable. C'est pour contourner cette limitation que des versions non-paramétriques ont récemment vu le jour. En effet, le seul moyen de pouvoir détecter des clusters de forme arbitraire sans pour autant voir le cardinal de Ω_0 augmenter de façon exponentielle avec le nombre de paramètres définissant les zones, est de ne pas considérer cet ensemble comme paramétrique.

Nous commençons par décrire la première version aboutie de la statistique de scan spatial, la version circulaire de Kulldorff [1997]. Nous présenterons ensuite ses versions dérivées, paramétrique d'abord avec l'extension aux fenêtres de forme elliptique, puis non-paramétriques avec les scan appelés ULS, MST, SA et flexible. Une application de la statistique de scan circulaire sur sur des données des registres du cancer de 9 départements français conclura ce chapitre.

3.1 La statistique de scan spatial circulaire

Nous commençons par décrire la statistique de scan circulaire élaborée par Kulldorff et Nagarwalla [1995] et Kulldorff [1997]. Outre l'extension de la statistique de scan au cas multi-dimensionnel, cette méthode permet à la population de fond d'être un processus de Poisson non-homogène ou un processus de Bernouilli avec une intensité proportionnelle à une certaine fonction connue. Le test du rapport de vraisemblance utilisé est différent selon qu'un modèle de Bernouilli ou un modèle de Poisson est considéré. Le premier est utilisé lorsque les témoins sont appariés sur les cas, le second lorsque le nombre de cas est négligeable face à la taille de la population de fond.

3.1.1 Cadre général

Dans l'article de Kulldorff [1997], l'auteur a tenté de traiter le problème de la façon la plus générale possible. La seule exception à cela reste que les analyses sont conditionnelles au nombre total de cas n_A observés sur le domaine d'étude A. L'intensité de la population de fond, qui gouverne la distribution des cas sous l'hypothèse nulle de répartition uniforme, est modélisée par une mesure μ sur A, telle que $\mu(B)$ correspond à la taille de la population de B pour tout $B \subset A$. Lorsque A est un segment et μ la mesure uniforme sur A, on retrouve le cas de figure temporel comme cas particulier.

Lorsque le domaine d'étude est divisé en cellules, les données aggrégées sont représentées par les coordonnées du barycentre de la cellule, le nombre de points et le nombre de cas. Les données individuelles sont représentées par les coordonnées du point et sa valeur (cas / non cas).

Dans ce qui suit, Z désigne une fenêtre qui va se déplacer de façon à couvrir la totalité de A. Ces fenêtres définissent une collection Ω_0 de zones $Z \subset A$. Cet ensemble Ω_0 peut être défini de plusieurs façons. Pour des raisons calculatoires, la méthode a été initialement appliquée exclusivement avec des zones circulaires. En effet, le nombre de paramètres dans ce cas est limité, une zone circulaire étant définie de façon unique par son centre et son rayon. Le rayon peut potentiellement varier continuement de 0 à $+\infty$. Dans la pratique, on le considère comme discret et on le fait varier jusqu'à une limite définie comme étant la longueur maximale séparant deux points du domaine étudié. Les centres se déplacent sur une grille prédéfinie. Par ailleurs, on se limite aux zones regroupant au plus 50% des cas, car un cluster regroupant plus de la moitié des cas serait plutôt révélateur d'un manque de cas à l'extérieur du cluster. Ainsi, le nombre de zones est fini. La zone maximisant le rapport de la vraisemblance constitue le cluster le plus probable, ce qui signifie que c'est le cluster qui a la probabilité la plus faible d'être apparu par hasard.

Soit $L(Z, p, q)$ la vraisemblance d'une zone Z telle que la probabilité pour un point à l'intérieur de Z d'être un cas soit p et q pour un point à l'extérieur de Z. L'hypothèse nulle (absence de cluster dans la zone) est $H_0 : p = q$. L'hypothèse alternative (nombre de cas

dans la zone anormalement élevé) est $H_1 : p > q$. La statistique de test permettant de déterminer si le cluster le plus probable est significatif s'écrit

$$\lambda = \frac{\sup_{Z\in\Omega_0, p>q} L(Z,p,q)}{\sup_{Z\in\Omega_0, p=q} L(Z,p,q)}. \tag{3.0}$$

Nous allons maintenant développer la statistique de test pour les deux modèles envisagés. Rappelons que n_A désigne le nombre total de cas observés et notons n_Z le nombre de cas observés dans la zone Z. Notons N un processus spatial de points où $N(B)$ est le nombre aléatoire de cas inclus dans l'ensemble $B \subset A$.

3.1.2 Le modèle de Bernouilli

Lorsqu'un modèle de Bernouilli est utilisé, chaque unité de mesure correspond à un individu qui peut être dans l'un ou l'autre de deux états, par exemple malade / non malade. Dans ce cas, $N(B)$ suit sous H_0 une loi binomiale $\mathcal{B}in(\mu(B), p)$ pour tout ensemble B. Sous H_1, $N(B) \sim \mathcal{B}in(\mu(B), p)$ pour tout $B \subset Z$, et $N(B) \sim \mathcal{B}in(\mu(B), q)$ pour tout $B \subset Z^c$.

Dans le cas d'un modèle de Bernouilli, la vraisemblance s'exprime comme suit :

$$L(Z,p,q) = p^{n_Z}(1-p)^{\mu(Z)-n_Z} q^{n_A-n_Z}(1-q)^{\mu(A)-\mu(Z)-(n_A-n_Z)}.$$

Si, à Z fixée, on note $L(Z) = \sup_{p>q} L(Z,p,q)$ et $L_0 = \sup_{p=q} L(Z,p,q)$, on obtient

$$L(Z) = \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(\frac{\mu(Z)-n_Z}{\mu(Z)}\right)^{\mu(Z)-n_Z} \times \left(\frac{n_A-n_Z}{\mu(A)-\mu(Z)}\right)^{n_A-n_Z} \left(\frac{\mu(A)-\mu(Z)-(n_A-n_Z)}{\mu(A)-\mu(Z)}\right)^{\mu(A)-\mu(Z)-(n_A-n_Z)}$$

lorsque $\frac{n_Z}{\mu(Z)} > \frac{n_A-n_Z}{\mu(A)-\mu(Z)}$ (nombre de cas dans Z supérieur au nombre de cas attendus sous H_0) et

$$L(Z) = L_0 = \left(\frac{n_A}{\mu(A)}\right)^{n_A} \left(\frac{\mu(A)-n_A}{\mu(A)}\right)^{\mu(A)-n_A} \quad \text{sinon.}$$

Ainsi la statistique de test définit par l'équation 3.1.1 s'écrit maintenant :

$$\lambda = \frac{\sup_{Z\in\Omega_0} L(Z)}{L_0},$$

et l'obtention de sa distribution par le biais de réplicats de Monte Carlo est décrite plus loin.

3.1.3 Le modèle de Poisson

Lorsqu'un modèle de Poisson est utilisé, les points sont générés par un processus de Poisson non-homogène. Dans ce cas, $N(B)$ suit une loi de Poisson $\mathcal{P}(p\mu(B\cap Z)+q\mu(B\cap Z^c))$ pour tout ensemble B. Sous H_0, $N(B) \sim \mathcal{P}(p\mu(B))$ pour tout B.

Dans le cas d'un modèle de Poisson, le calcul de la vraisemblance est un peu plus compliqué. La probabilité d'observer n_A cas dans A est

$$P\left(N(A)=n_A\right)=\frac{e^{-p\mu(Z)-q(\mu(A)-\mu(Z))}\left[p\mu(Z)+q(\mu(A)-\mu(Z))\right]^{n_A}}{n_A!}.$$

La fonction de densité $f(x)$ pour un cas donné d'être localisé en x est

$$f(x)=\frac{p\mu(x)}{p\mu(Z)+q(\mu(A)-\mu(Z))}I_Z(x)+\frac{q\mu(x)}{p\mu(Z)+q(\mu(A)-\mu(Z))}I_{Z^c}(x),$$

où $I_B(x)$ vaut 1 si $x\in B$ et 0 sinon.

La vraisemblance s'écrit alors

$$L(Z,p,q) \;=\; P\left(N(A)=n_A\right)\times\prod_{i=1}^{n_A}f(x_i) \;=\; \frac{e^{-p\mu(Z)-q(\mu(A)-\mu(Z))}}{n_A!}p^{n_Z}q^{n_A-n_Z}\prod_{i=1}^{n_A}\mu(x_i).$$

Pour simplifier les écritures qui vont suivre, adoptons la notation $\Pi_\mu=\prod_{i=1}^{n_A}\mu(x_i)$. Reprenons par ailleurs les mêmes définitions que précédemment pour $L(Z)$ et L_0. On obtient tout d'abord

$$L_0=\frac{e^{-n_A}}{n_A!}\left(\frac{n_A}{\mu(A)}\right)^{n_A}\Pi_\mu.$$

Puis, en notant qu'à Z fixée la vraisemblance atteint son maximum pour

$$p=\frac{n_Z}{\mu(Z)}\text{ et }q=\frac{n_A-n_Z}{\mu(A)-\mu(Z)},$$

on obtient

$$L(Z)=\frac{e^{-n_A}}{n_A!}\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_A-n_Z}{\mu(A)-\mu(Z)}\right)^{n_A-n_Z}\Pi_\mu$$

lorsque $\frac{n_Z}{\mu(Z)}>\frac{n_A-n_Z}{\mu(A)-\mu(Z)}$ et

$$L(Z)=\frac{e^{-n_A}}{n_A!}\left(\frac{n_A}{\mu(A)}\right)^{n_A}\Pi_\mu\quad\text{sinon.}$$

La statistique de test, toujours définie par l'équation 3.1.1 s'écrit donc dans ce cas

$$\lambda=\sup_{Z\in\Omega_0}\frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_A-n_Z}{\mu(A)-\mu(Z)}\right)^{n_A-n_Z}}{\left(\frac{n_A}{\mu(A)}\right)^{n_A}}I\left(\frac{n_Z}{\mu(Z)}>\frac{n_A-n_Z}{\mu(A)-\mu(Z)}\right)$$

s'il existe au moins une zone Z telle que $\frac{n_Z}{\mu(Z)}>\frac{n_A-n_Z}{\mu(A)-\mu(Z)}$ et $\lambda=1$ sinon.

3.1.4 Inférence

La méthode consiste donc dans les deux cas de figure à calculer le rapport de la vraisemblance pour chaque zone Z appartenant à l'ensemble des zones candidates Ω_0. La distribution de λ sous H_0 est obtenue par le biais de simulations de Monte Carlo (Dwass [1957]), ce qui permet de calculer la p-valeur associée au cluster le plus probable et de déterminer si ce cluster est significatif. Pour cela, M réplicats de l'échantillon sous H_0 sont générés. Pour chacun de ces échantillons, la statistique λ est calculée. La p-valeur est alors égale à $\frac{N_\lambda}{M+1}$ où N_λ est le nombre de réplicats où la statistique de test est supérieure à celle de l'échantillon initial. Généralement, on considère $M = 999$ ou $M = 9999$, ce qui permet d'obtenir $M + 1 = 1000$ ou 10000 échantillons en incluant l'échantillon initial des données.

Un exemple d'application du scan circulaire sur données groupées (données de comptage) est présenté dans la section 3.3. La méthode sera ensuite appliquée à la fois sur données groupées et individuelles dans la section 4.8.2.

3.2 Versions dérivées

3.2.1 Le scan elliptique

Récemment Kulldorff et al. [2006] ont adapté la statistique de scan spatial à la détection de clusters de forme elliptique. Le principe de base de la méthode reste le même que dans le cas circulaire. Dans cette nouvelle version, l'ensemble des zones candidates Ω_0 n'est plus seulement défini par un ensemble de centres et de rayons. Il faut leur ajouter l'excentricité (la forme) et l'angle de l'ellipse par rapport à l'axe des abscisses. Le rayon devient la taille de l'ellipse (le demi grand axe). L'excentricité est le ratio de la longueur du demi grand axe sur celle du demi petit axe. Le cercle est donc une ellipse d'excentricité 1. En ce sens, la statistique de scan circulaire est un cas particulier de la statistique de scan elliptique.

L'augmentation du nombre de paramètres rend nécessaire, plus encore que pour la version circulaire, la discrétisation des valeurs prises par les différents paramètres, notamment l'excentricité et l'angle. Les auteurs conseillent de considérer successivement des excentricité de $1, 1.5, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 60$ et 120. Ils conseillent également d'envisager un nombre d'angles égal à trois fois l'excentricité. Enfin et comme précédemment, les centres sont définis par une grille et la taille (anciennement le rayon) varie de 0 à la demi longueur du domaine étudié. Le scan elliptique est bien évidemment plus gourmands en calculs et l'est d'autant plus que l'ellipse est excentrée.

Les études de puissance menées par les auteurs ont permis de mettre en évidence une puissance légèrement supérieure du scan elliptique lorsque le vrai cluster à la forme d'une ellipse allongée. La puissance est comparable lorsque l'excentricité de l'ellipse se rapproche de 1.

Même si l'ellipse est plus flexible que le cercle, l'auteur avoue que cette amélioration impose encore une forme paramétrique aux clusters potentiels et préconise l'utilisation de versions non-paramétriques, ci-après décrites, si on désire détecter des clusters de forme très irrégulières.

3.2.2 Le scan ULS

Avec la statistique de scan ULS (Upper Level Set), Patil et Taillie [2004] proposent de rechercher le cluster le plus probable parmi les éléments d'un ensemble Ω_{ULS} qui est déterminé à partir des données en utilisant le taux d'incidence empirique des cellules. Nous nous trouvons donc bien là dans un cadre non-paramétrique.

Si nous notons, comme précédemment, $\mu(Z)$ et n_Z le nombre de points et le nombre de cas appartenant à une zone Z, le taux d'incidence G_Z de la cellule Z s'écrit

$$G_Z = \frac{n_Z}{\mu(Z)}.$$

On peut ainsi définir une fonction $Z \to G_Z$ sur l'ensemble des cellules qui ne prend qu'un nombre fini de valeurs. Chacune de ces valeurs g détermine un ensemble de niveau supérieur (un ULS)

$$U_g = \{Z : G_Z \geqslant g\}.$$

Ces ULSs ne sont pas forcément connexes. L'ensemble Ω_{ULS} des zones candidates est défini comme l'ensemble des ULSs connexes. La détermination des ensembles connexes nécessite la définition d'un critère d'adjacence entre cellules. Un des critères envisageable est : *deux cellules sont adjacentes si leur frontière commune a une longueur strictement positive.*

Une des conséquences de la dépendance aux données de l'approche ULS est la nécessité de recalculer l'ensemble Ω_{ULS} pour chaque réplicat de Monte Carlo afin d'obtenir la p-valeur du cluster le plus probable. Comme pour les versions paramétriques, M réplicats sont générés sous l'hypothèse nulle. La statistique de test

$$\lambda = \frac{\sup_{Z \in \Omega_{ULS}, p>q} L(Z,p,q)}{\sup_{Z \in \Omega_{ULS}, p=q} L(Z,p,q)}. \tag{3.0}$$

est calculée sur les données et les M réplicats. La p-valeur est déterminée par le rang de la valeur de λ calculée sur les données parmi le vecteur des $M+1$ valeurs. L'obtention de la p-valeur nécessite donc de calculer $M+1$ ensembles Ω_{ULS}.

Le scan ULS a été appliqué dans l'analyse de la répartition des pharmacies à Montpellier, présenté dans la section 4.8.2.

TAB. 3.1 – Algorithme de détermination du MST

Initialisation du MST à l'ensemble vide ;
    Ajoût d'une cellule v arbitraire au MST ;
    Tant que il reste des cellules encore non incluses au MST **faire**
        Trouver le coût minimum entre v_i et v_j, avec $v_i \in MST$ et $v_j \notin MST$;
        Ajoût de la cellule v_j au MST ;
Enregistrement du MST.

3.2.3 Le scan MST

Le scan ULS est un cas particulier d'une des deux méthodes proposées depuis par Assunção et al. [2006]. Ces méthodes, encore une fois basée sur la statistique de scan spatial, s'appuient sur une nouvelle définition de l'ensemble des zones candidates Ω_0 par le biais d'arbres de recouvrement minimum (MST : minimum spanning tree). La détermination du MST nécessite d'attribuer un coût à la liaison entre deux cellules. Ce coût mesure la dissimilarité entre les cellules et dépend des taux d'incidence des deux cellules. L'algorithme de détermination du MST est présenté dans la Table A.1.

La première méthode est dite statique (sMST). L'ensemble Ω_0 est obtenu est scindant l'arbre MST en deux. Si le domaine est constitué de k cellules, il existe $k-1$ façons de scinder l'arbre en supprimant la liaison entre deux cellules. L'ensemble des zones candidates est donc suffisamment petit ce qui permet de trouver la meilleure estimation du cluster rapidement. Le calcul du coût associé à deux cellules utilise la distance de Kullback-Leibler. Cette méthode, dont le scan ULS est un cas particulier, a cependant une puissance faible et son utilisation n'est pas recommandée par les auteurs.

La seconde approche dite dynamique (dMST) construit le MST itérativement en mettant à jour les coûts de façon dynamique à chaque itération. Cette méthode nécessite d'initialiser l'algorithme en choisissant une zone initiale arbitraire. Le dMST dépend donc de ce choix. Ce problème est contourné en choisissant successivement toutes les cellules comme zone initiale. Cette version du MST est plus puissante que la première mais tend à détecter des clusters plus grands que le vrai cluster.

Dans les deux cas, le test de l'hypothèse nulle s'effectue là encore par le biais de simulations de Monte Carlo. Des études de simulation menées par les auteurs ont montrées que le scan dMST est plus puissant que le scan circulaire lorsque le vrai cluster a une forme très allongée. Par contre, si ce dernier a la forme d'une étoile ou d'une couronne, le scan circulaire est plus puissant.

3.2.4 Le scan SA

Duczmal et Assunção [2004] ont proposé une autre version non paramétrique de la statistique de scan. L'ensemble des zones candidates est constitué de toutes les zones regroupant des cellules adjacentes. Ils élargissent donc le choix du cluster par rapport au scan ULS. Cependant, le cardinal de cet ensemble ne permet pas une exploration exhaustive de toutes les zones candidates. Afin de réduire les temps de calcul et analyser seulement les zones les plus prometteuses, une stratégie de recuit simulé est utilisée.

Le balayage de tous les zones constituées de cellules adjacentes se fait par un algorithme qui nécessite la définition de règles permettant de choisir le meilleur voisin (cellule adjacente) à chaque étape. Ce choix doit se faire de façon à minimiser le nombre de zones examinées.

Le choix le plus naturel est de sélectionner la cellule adjacente ayant la vraisemblance la plus élevée. Cependant cette règle ne fonctionne généralement pas car elle conduit la plupart du temps à des zones dont la vraisemblance est un maxima local. La stratégie du recuit simulé consiste à choisir, à des moments judicieux, le voisin de façon aléatoire au lieu de systématiquement sélectionner le voisin ayant la vraisemblance la plus élevée. Cette nouvelle règle donne plus de liberté à l'algorithme et permet de sélectionner des voisinages potentiellement plus intéressant et qui n'auraient jamais été visités sinon.

Trois stratégies sont possibles dans le choix du meilleur voisin :
- Choix aléatoire uniforme parmi les voisins,
- Choix aléatoire parmi les voisins, mais proportionnellement à la valeur de leur vraisemblance,
- Choix du voisin ayant la vraisemblance la plus élevée.

A chaque étape, le choix de la stratégie à appliquer se fait en fonction de 4 paramètres :
- Une variable binaire dénotant la présence d'un voisin avec une vraisemblance plus élevée que les cellules de la zone actuelle,
- Le nombre d'étapes consécutives ou la variable précédente vaut 0,
- Le nombre de fois où la zone actuelle à déjà été visitée plus tôt dans la surveillance,
- Le nombre de cellules de la zone actuelle qui sont adjacentes à la cellule déjà visitée ayant la vraisemblance la plus élevée.

Des seuils sont préalablement fixés pour chacun de ces paramètres. A chaque étape, l'algorithme de sélection vérifie si les seuils sont atteints et adopte la stratégie la plus adaptée de sélection du voisin suivant. La routine de surveillance est finalement abandonnée lorsque l'un des paramètres dépasse le seuil qui lui a été attribué.

L'initialisation de la zone de départ de l'algorithme se fait en utilisant tout d'abord la zone localisée par le scan circulaire, puis une seule cellule choisie aléatoirement. L'algorithme avec une cellule comme zone de départ est répété jusqu'à ce que 99% de toutes les cellules

du domaine étudié aient été visitées au moins une fois. La routine de surveillance est donc appelée des centaines de fois. Le cluster le plus probable est, parmi les zones déterminées par chaque routine, celle qui est associée à la vraisemblance la plus élevée.

La significativité du cluster le plus probable est finalement testée par la procédure de Monte Carlo.

Le scan SA a été appliqué dans l'analyse de la répartition des pharmacies à Montpellier (section 4.8.2).

3.2.5 Le scan flexible

Dernièrement, Tango et Takahashi [2005] ont mis au point une méthode qu'ils ont appelée la statistique de scan flexible. Le principe de cette approche est semblable à celle du scan SA, à savoir que l'ensemble des zones candidates Ω_0 est là encore constitué d'ensembles de cellules adjacentes. L'ensemble Ω_0 est cependant restreint dans ce cas à des zones regroupant au maximum K cellule adjacentes.

Les zones candidates sont déterminées en considérant successivement pour chaque cellule les sous-ensembles de ses $K-1$ plus proches voisins. Contrairement au scan SA, ces limitations permettent d'obtenir un ensemble de zones candidates de taille raisonnable, bien que beaucoup plus élevée que dans le cas du scan circulaire. Cette taille permet ainsi de pouvoir effectuer une recherche exhaustive de la zone maximisant la vraisemblance et permet de se passer de l'algorithme de recuit simulé.

La significativité du cluster le plus probable est, comme pour le scan circulaire, testée en simulant un grand nombre de fois la distribution de la statistique de test sous l'hypothèse nulle (réplicats de Monte Carlo).

Les auteurs ont montré que la statistique de scan flexible a une puissance légèrement inférieure à celle du scan circulaire lorsque le vrai cluster est de forme circulaire, et une puissance légèrement supérieure lorsque le vrai cluster a une forme allongée. Par ailleurs, la limitation de la taille des zones candidates permet au scan flexible d'éviter l'inconvénient du scan SA de détecter des clusters plus grands que les vrais cluster. En contrepartie, le scan flexible ne fonctionne que pour des petites tailles de clusters, inférieures à 30 cellules, essentiellement à cause des temps de calculs qui deviennent très longs au delà[1].

Dans ce même article, Tango et Takahashi [2005] ont introduit la notion de puissance bivariée, notion approfondie depuis dans Takahashi et Tango [2006]. Cette puissance bivariée fait apparaître la puissance usuelle, plus les puissances jointe et marginale. La puissance jointe $P(l, s)$ représente la probabilité de détecter un cluster de taille l incluant s cellules

[1] Les calculs prennent plus d'une semaine si $K > 30$, pour un nombre total de cellules de l'ordre de 200.

du vrai cluster. La puissance marginale représente la probabilité de mettre en évidence un cluster incluant s cellules du vrai cluster, sans tenir compte de la taille du cluster détecté. Cette puissance étendue permet d'obtenir des renseignement très utiles sur les tests de détection. Elle permet notamment de comparer les tests sur le nombre de mal classés, à savoir les faux positifs et les faux négatifs. Cependant cette puissance étendue ne peut être appliquée que dans le cas de données groupées, sa définition étant basée sur les cellules.

L'un des points communs à ces statistiques de scan non-paramétriques est qu'elles ne s'appliquent qu'à des données groupées, contrairement aux versions paramétriques qui peuvent s'appliquer indifféremment à des données groupées ou individuelles. En effet, ces méthodes non-paramétriques nécessitent de définir un critère d'adjacence entre cellules, critère qui n'est pas applicable aux données individuelles.

3.3 Analyses de clusters de cancers et rayonnements ionisants

Nous présentons ici la partie spatiale de l'étude **"Analyses de clusters de cancers et rayonnements ionisants (en vue de disposer d'hypothèses de travail de type étiologique)"**. Les motivations de l'étude et les résultats des analyses temporelles ont été présentés dans la section 1.2.2. Pour les mêmes raisons qu'en temporel, seuls les résultats graphiques concernant le Tarn et La Manche ont été reproduits en Annexe B. Tous les résultats présentés ici ont été obtenus par l'application de la méthode du scan circulaire de Kulldorff [1997] avec un modèle de Poisson.

3.3.1 Données spatiales et population des communes

La location exacte et la population des communes sont nécessaires à l'analyse de la répartition spatiale des cas de cancer. Ces données sont nécessaires même pour les communes n'ayant pas présenté de cas de cancer pendant la période étudiée car l'ensemble de toutes les communes d'un département constitue la population de fond du département.

La géolocalisation des communes a été effectuée sur le site internet "Maporama". Les latitudes et longitudes ont été recueillies en degré décimal. La population des communes de 1990 et 1999 ont été recueillies sur le site de l'INSEE. Pour une étude spatiale, la population ne peut être spécifiée au cours du temps. Nous avons considéré la population de 1990 pour tous les départements sauf pour le département de la Manche. La période d'étude de ce dernier étant $[1994 - 1999]$, nous avons considéré la population de 1999.

La répartition de la population par commune est représentée pour chaque département en annexe B.3. La légende de ces graphiques figure en annexe B.2.

3.3.2 Résultats des analyses spatiales de détection de clusters

Pour chacun des trois types de cancer, un tableau récapitule le nombre de cas dans l'ensemble du département pour l'ensemble de la période étudiée, ainsi que les informations concernant le cluster détecté dans chaque département (s'il existe). Dans ces tableaux, C désigne le nombre total de cas dans le département. n, c et c_e désignent respectivement la population, le nombre de cas et le nombre de cas attendus sous H_0 pour le cluster le plus probable. Par ailleurs, r désigne le rayon du cluster en kilomètres, i l'incidence annuelle moyenne du nombre de cas pour 100000 habitants pendant la période d'étude (calculée pour le département et pour le cluster) et p la p-value du cluster.

Les graphiques illustrant les résultats de ces analyses sont présentés en annexes B.4.

3.3.2.1 Cancer du SNC

	Total Dept		Cluster					
Dépt	C	i	r (km)	n	c	i	$E[c]$	p
14	524	4.0						
25	667	5.7						
34	1001	7.4						
38	1992	9.3	33	495895	1159	11.1	972.05	0.0001
50	188	6.5						
67	1337	5.6						
68	550	6.8						
80	542	5.8						
81	393	6.4						

TAB. 3.2 – Cancers du SNC : informations sur les clusters localisés dans chaque département.

Un cluster de cancer du SNC d'un rayon de 33 km, représenté sur la Figure 3.1, a été détecté dans le sud de l'Isère. Grenoble, qui se situe au nord de ce cluster, est la ville ayant la vraisemblance la plus élevée du département. Son incidence est de 12.2 cas de cancer du SNC par an pour 100000 habitants. La centrale de St Alban, non incluses dans le cluster, se situe à 83 km au nord-ouest du centre du cluster et à 77 kms de Grenoble.

Aucun cluster de cancer du SNC n'a été détecté dans les autres départements.

3.3.2.2 Cancers hématologiques

Les données de l'Hérault (34) n'ont pu être exploitées pour les cancers hématologiques (adresses inconnues pour un grand nombre de patients).

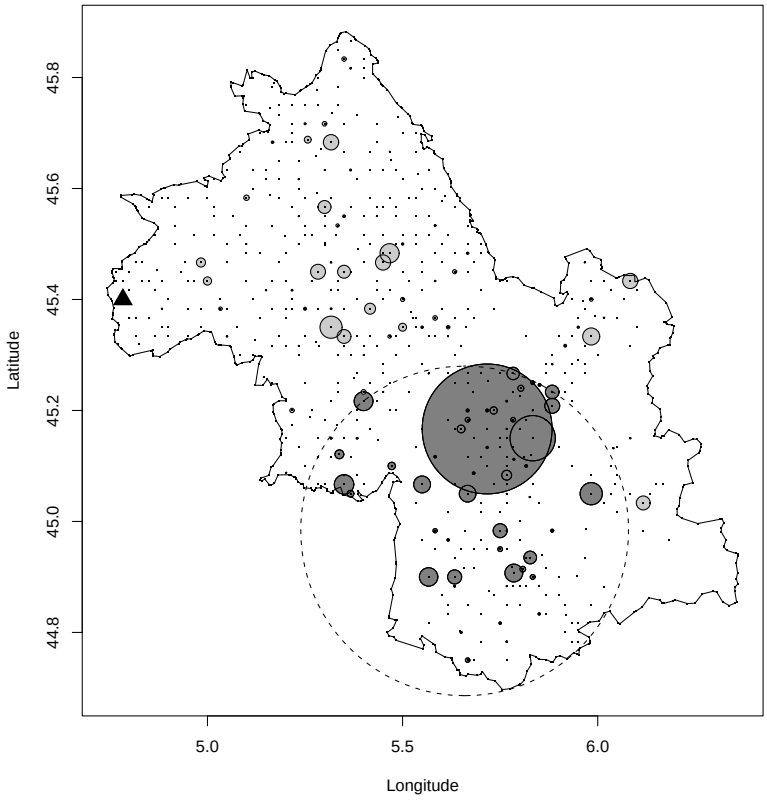


FIG. 3.1 – Vraisemblance des cas de cancer du SNC par commune dans le département de l'Isère (38).

Un cluster de cancers hématologiques a été détecté en bordure de mer dans le nord-est du Calvados (nord-est de Caen) regroupant 71 communes. Les villes ayant les vraisemblances les plus élevées sont Ouistreham ($i = 37.9$) et Dives-sur-Mer ($i = 40.8$).

Le cluster détecté dans le Doubs englobe une grande partie de l'ouest du département (334 communes). Besançon est la ville avec la vraisemblance la plus élevée ($i = 36.4$).

Le cluster détecté dans l'Isère regroupe 4 communes dont Grenoble qui a la vraisemblance la plus élevée ($i = 34.0$). La centrale de St Alban ne fait pas partie du cluster.

La Manche présente un cluster d'un rayon de 27 kms situé dans le sud du département et constitué de 182 communes. Celle qui a la vraisemblance la plus élevée est Granville ($i = 79.2$). Les installations de Flamanville et de la Hague, non incluses dans le cluster détecté, sont respectivement situées à 88 et 104 kms au nord du centre du cluster.

Le Bas-Rhin présente un cluster d'un rayon de 4.5 kms regroupant 14 communes. La commune du département présentant la vraisemblance la plus élevée est Stundwiller ($i =$

	Total Dept		Cluster					
Dépt	C	i	r (km)	n	c	i	c_e	p
14	2962	21.8	16.49	58021	359	28.1	277.87	0.0032
25	3697	31.8	40.82	237043	2018	35.5	1807.76	0.0001
38	5805	27.2	3.93	165707	1193	34.3	946.57	0.0001
50	1303	45.1	26.56	119407	408	56.9	323.15	0.0008
67	6822	28.6	4.47	77288	665	34.4	553.23	0.0029
68	2869	35.6	1.55	1146	18	130.9	4.9	0.0088
80	2474	26.6	6.24	5119	50	57.5	23.12	0.0039
81	2191	35.5	24.31	112188	914	45.3	717.21	0.0001

TAB. 3.3 – Cancers hématologiques : informations sur les clusters localisés dans chaque département.

167.3, 263 habitants) mais elle n'est pas incluse dans le cluster. Immédiatement après, vient Schiltigheim ($i = 37.9$, 29155 habitants) qui est incluse dans le cluster. Cette dernière est proche de Strasbourg (qui ne fait pas partie du cluster).

Le cluster détecté dans le Haut-Rhin regroupe les communes de Steinbrunn-le-Bas (618 habitants, 6 cas en 12 ans) et Steinbrunn-le-Haut (528 habitants, 12 cas en 12 ans). Cette dernière est la commune du département présentant la vraisemblance la plus élevée et a une incidence pour 100000 habitants $i = 189.4$. La centrale de Fessenheim, située à 31 kms au nord des deux communes, ne fait pas partie du cluster.

La Somme présente un cluster de 6 kms de rayon composé de 21 communes, dont La Neuville-lès-Bray, commune de 215 habitants, présentant 7 cas en 17 ans ($i = 191.5$) et qui a la vraisemblance la plus élevée du département.

Enfin le Tarn présente un cluster qui englobe une grande partie du nord du département (87 communes). Albi est la commune du département et du cluster qui présente la vraisemblance la plus élevée avec 377 cas en 18 ans pour 46579 habitants ($i = 45.0$).

3.3.2.3 Cancer de la thyroïde

Le cluster de cancer de la thyroïde détecté dans le Doubs englobe une grande partie du sud-ouest du département (185 communes). Rouhe, avec 2 cas pour 52 habitants en 24 ans est la ville qui présente la vraisemblance la plus élevée ($i = 160.3$).

Le Tarn également présente un cluster dans le nord-est du département. Comme pour les cancers hématologiques, Albi est la commune du département et du cluster qui présente la vraisemblance la plus élevée avec 85 cas en 18 ans ($i = 10.1$).

Aucun cluster de cancer de la thyroïde n'a été détecté dans les autres départements.

	Total Dept		Cluster					
Dépt	C	i	r (km)	n	c	i	c_e	p
14	591	4.6						
25	541	4.6	31.59	209903	287	5.7	234.25	0.0108
34	698	2.5						
38	993	4.7						
50	200	6.9						
67	776	3.3						
68	267	3.3						
80	289	3.1						
81	434	7.0	24.02	92401	157	9.4	117.01	0.0242

TAB. 3.4 – Cancers de la thyroïde : informations sur les clusters localisés dans chaque département.

3.3.3 Analyses spatiales ajustées sur l'âge

L'âge est une covariable qu'il est important de prendre en compte dans les études sur le cancer. Afin d'ajuster l'analyse spatiale suivant cette variable, il est nécessaire de disposer de la répartition de la population par âge dans chaque commune. Cette information est disponible sur le site du recensement de 1999 de l'INSEE (bases de données téléchargeables pour chaque département).

Nous avons utilisé un découpage de l'âge en 5 classes ($k = 5$) : moins de 19 ans, entre 20 et 39 ans, entre 40 et 59 ans, entre 60 et 74 ans, et plus de 75 ans.

Pour les départements autres que La Manche, nous considérons, comme précédemment, la population de 1990 comme population de fond de l'analyse spatiale. Ne disposant pas de la répartition par âge en 1990, nous avons fait l'hypothèse qu'elle était la même qu'en 1999, à populations égales. Pour une commune donnée, notons n_j^{99} le nombre de personne résidant dans la commune en 1999 et appartenant à la classe d'âge j. Notons par ailleurs $n_\cdot^{90}$ et $n_\cdot^{99}$ la population totale de la commune respectivement en 1990 et 1999. La répartition de la population par classe d'âge en 1990 se calcule alors comme suit :

$$n_j^{90} = n_j^{99} \times \frac{n_\cdot^{90}}{n_\cdot^{99}}.$$

Les graphiques illustrant les résultats de ces analyses ajustées sur l'âge sont présentés en annexes B.5.

Les résultats concernant le cancer SNC ont été peu affectés par l'ajustement. Ce dernier a fait apparaître un cluster au nord de l'Hérault, d'un rayon de 40.1 km et regroupant 118 communes dont Montpellier. La prise en compte de l'âge n'a eu aucun impact sur le cluster

détecté dans le sud de l'Isère. Aucun cluster de cancer du SNC n'a été détecté dans les autres départements.

La prise en compte de l'âge a quelque peu modifié la localisation des agrégats de cancers hématologiques. Un cluster, regroupant 129 communes, a été détecté dans le nord du Calvados, à l'ouest de celui détecté sans ajustement. Le Doubs présente un cluster de 255 communes, dont Besançon, situé dans l'ouest du département, au sud du cluster non ajusté. Les clusters détectés dans l'Isère, le Bas-Rhin et le Haut-Rhin se situent au même endroit que sans ajustement, avec un nombre de communes légèrement différent. Le cluster précédemment détecté dans la Manche a disparu, alors que celui de la Somme regroupe les mêmes communes. Enfin, le cluster détecté dans le Tarn se situe légèrement à l'est de celui précédemment détecté.

La prise en compte de l'âge n'a pas eu d'impact sur la localisation et la détection des clusters de la thyroïde.

3.3.4 Discussion

Les analyses temporelles ont mis en évidence des clusters de cancers qui semblent n'être dus qu'à une amélioration du recueil des données au cours du temps.

Les analyses spatiales ont en premier lieu été effectuées sans ajustement sur l'âge.

Les données sur le cancer du SNC ont permis de mettre en évidence un cluster spatial dans l'Isère, tandis que les autres départements ne présentent pas de cluster. L'Isère est l'un des trois départements présentant une installation nucléaire (centrale de St Alban). Cette dernière est située à 80 kms du centre du cluster détecté.

Un cluster de cancers hématologiques a été détecté dans chacun des 9 départements étudiés. Là encore, les 4 installations nucléaires (réparties dans 3 départements) ne font pas partie des clusters détectés.

Concernant le cancer de la thyroïde, les départements du Doubs et du Tarn présentent un cluster. Aucune installation nucléaire n'est présente dans ces deux départements.

La prise en compte de l'âge a permis de mettre en évidence un cluster de cancers du SNC dans l'Hérault et fait disparaître le cluster de cancers hématologiques dans la Manche. Dans le Calvados, le Doubs et le Tarn, les clusters de cancers hématologiques détectés ne sont pas localisés au même endroit selon que l'on ajuste ou pas sur l'âge. Ces changements n'ont cependant pas mis en évidence de clusters autour d'une installation nucléaire.

Une analyse spatio-temporelle ne nous a pas semblé pertinente pour deux raisons. La première est qu'aucun cluster temporel n'a été mis en évidence (les clusters détectés sont uniquement attribuables a une tendance linéaire). La seconde raison concerne le faible nombre de cas par an et par commune.

Les clusters spatiaux qui ont été détectés se situent aussi bien dans les départements accueillant un site nucléaire que dans ceux n'en accueillant pas. Par ailleurs, lorsqu'une installation nucléaire est présente, et si un cluster est détecté dans le département, la distance les séparant est généralement de plusieurs dizaines de kilomètres. Ces résultats ne permettent pas d'établir un quelconque lien entre la répartition (temporelle ou spatiale) des cas de cancers et les rayonnements ionisants provenant des sites nucléaires.

Chapitre 4

Méthode de régression sur données transformées

Nous proposons dans cette section une méthode originale pour la détection de clusters spatiaux de données ponctuelles (Dematteï et al. [2006a]). Cette méthode est une adaptation de la méthode temporelle de Molinari et al. [2001] décrite dans le chapitre précédent. Nous attribuons à chaque point un ordre de sélection ainsi que la distance de ce point à son plus proche voisin une fois les points déjà sélectionnés pris en compte. Cette distance est ensuite pondérée par l'espérance de la distance sous l'hypothèse de répartition uniforme. Les clusters potentiels sont localisés par une modélisation à changement structurel multiple des distances sur l'ordre de sélection et le meilleur modèle (contenant un ou plusieurs agrégats potentiels) est sélectionné en utilisant le test du double maximum de Bai et Perron [1998]. Finalement, une p-valeur est obtenue pour chaque cluster potentiel. Avec cette méthode, plusieurs clusters de formes quelconques peuvent être détectés. Elle offre ainsi l'avantage, par rapport aux méthodes précédemment décrites, de pouvoir détecter des clusters de données ponctuelles ayant une forme arbitraire. En effet, les formes paramétriques de la statistique de scan s'appliquent aussi bien sur des données groupées qu'individuelles, mais la forme des agrégats est prédéfinie. D'un autre côté, les variantes non-paramétriques permettant de détecter des clusters de forme quelconque ne s'appliquent qu'à des données groupées.

4.1 Transformation des données

Soit $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées qui représentent les coordonnées spatiales de n occurrences d'évènements dans une région A, un ensemble borné de $\mathbb{R}^2$. Les données initiales sont constituées par les coordonnées des n points. L'ensemble de ces n points est inclus dans la population sous-jacente de taille N.

Nous introduisons deux variables construites à partir des données initiales et labellisées "distance" et "ordre". La première représente la distance d'un point à son plus proche voisin, une fois les points déjà sélectionnés pris en compte. La variable "ordre"peut être

vue comme un ordre de sélection des points. Elle définit une trajectoire à travers A et nous permettra d'effectuer la régression des distances sur l'ordre. L'hypothèse nulle est la répartition uniforme des points dans A. Sous H_0, aucun cluster n'est détecté. L'idée générale de la méthode est basée sur l'hypothèse que les points inclus dans un cluster ont des ordres de sélection consécutifs et que les distances qui leur sont associées sont plus petites que celles associées aux points hors du cluster (car la densité de points est plus élevée dans le cluster).

Un exemple de données simulées est présenté en Figure 4.1(a). Dans $A = [0, 100]^2$, un échantillon de 70 points est simulé suivant un mélange de trois processus de points uniformes $\frac{5}{7} \times \mathcal{U}([0, 100]^2) + \frac{1}{7} \times \mathcal{U}([10, 30] \times [60, 75]) + \frac{1}{7} \times \mathcal{U}([65, 80] \times [20, 40])$. Dans les zones de simulation des clusters, délimitées par des pointillés, la densité est environ cinq fois plus élevée que la densité dans l'ensemble du domaine d'étude.

4.2 Trajectoire

Pour $k = 1, \ldots, n$, soit x_k une réalisation de X_k et $x_{(k)}$ une réalisation de $X_{(k)}$ (la statistique d'ordre de X_k). $x_{(k)}$ est le point d'ordre de sélection k. La variable ordre est construite par un algorithme récursif initialisé par le choix du point d'ordre 1, $x_{(1)}$. Le point $x_{(k+1)}$ est déterminé par la connaissance des points $x_{(1)}, \ldots, x_{(k)}$.

Le choix de $x_{(1)}$ est arbitraire : nous avons décidé de prendre le point le plus proche du bord du domaine d'étude. Ce choix sera débattu en section 4.7. Puis, étant donnés $x_{(1)}, \ldots, x_{(k)}$, le point $x_{(k+1)}$ est le point le plus proche de $x_{(k)}$ parmi les $n-k$ points encore non sélectionnés.

Une trajectoire est ainsi définie qui relie successivement chaque point au point d'ordre suivant comme le montre la Figure 4.1(b).

4.3 Pondération de la distance

Le processus d'élimination des points déjà sélectionnés diminue le nombre de candidats potentiels dans la recherche du plus proche voisin. Ainsi, la distance observée au plus proche voisin est plus élevée pour les points sélectionnés tardivement. Une pondération de la distance est donc nécessaire. Comme nous allons le voir, cette pondération permet également d'ajuster les calculs sur une inhomogénéïté de la densité de la population sous-jacente. L'espérance de la distance d'un point à son plus proche voisin sous l'hypothèse de répartition uniforme est donnée par Bickel et Breiman [1983]. Ce qui suit est l'adaptation de leur raisonnement à notre cas particulier.

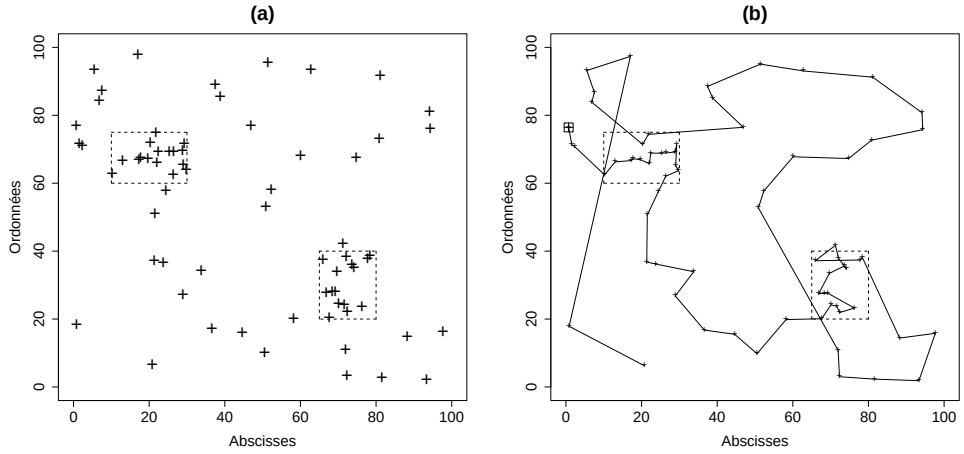


FIG. 4.1 – **(a)** Données simulées ($n = 70$) **(b)** Trajectoire suivie en fonction de l'ordre de sélection des points. Le point carré est le premier point sélectionné. Les aires rectangulaires en pointillés représentent les zones de simulation des clusters.

Soit $x_{(1)}, \ldots, x_{(n)}$ une réalisation de $X_{(1)}, \ldots, X_{(n)}$, n points échantillonnés indépendamment suivant une densité sous-jacente $h(x)$. Comme précédemment défini, $x_{(k)}$ est le point d'ordre k. Pour $k = 1, \ldots, n-1$, définissons $D_k = d(X_{(k)}, X_{(k+1)})$ la distance de $X_{(k)}$ à $X_{(k+1)}$, avec une fonction de densité g_k et une fonction de distribution G_k. Ainsi, $d_k = d(x_{(k)}, x_{(k+1)})$, une réalisation de D_k, est la distance observée de $x_{(k)}$ à $x_{(k+1)}$. La distance pondérée est ensuite définie comme le ratio entre la distance observée et son espérance sous l'hypothèse de répartition uniforme :

$$d_k^w = d_k \times E_{H_0}\left[D_k | X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right]^{-1}. \tag{4.0}$$

Une distance pondérée plus grande (respectivement plus petite) que 1 signifie que la distance observée est plus grande (respectivement plus petite) que son espérance. Ou encore, l'hypothèse de répartition uniforme ne sera pas rejetée si la distance pondérée est statistiquement proche de 1.

Etant donné que D_k est positive et bornée (puisque A est bornée), une intégration par partie nous permet d'écrire l'espérance ci dessus comme suit

$$\int_0^a r g_k(r) dr = \int_0^a \left(1 - G_k(r)\right) dr,$$

où

$$a = \max_{(u,v)\in A^2} d(u, v)$$

est le diamètre de A. Ainsi

$$\begin{aligned} &E\left[D_k | X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right] \\ &= \int_0^a P\left(D_k > r | X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right) dr. \end{aligned}$$

Soit $S(x,r)$ la sphère de centre x et de rayon r. L'ensemble $\{D_k > r | X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\}$ est égal à l'évènement qu'aucun des $X_{(k+1)}, \ldots, X_{(n)}$ ne tombe dans $S(x_{(k)}, r)$. Ainsi

$$\begin{aligned} & P\left(D_k > r | X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right) \\ &= \prod_{i=k+1}^{n} \left[1 - P\left(x_{(i)} \in A_{k-1} \bigcap S(x_{(k)}, r) \mid x_{(i)} \in A_{k-1}\right)\right] \\ &= \left[1 - \frac{\int_{A_{k-1} \bigcap S(x_{(k)}, r)} h(x)dx}{\int_{A_{k-1}} f(x)dx}\right]^{n-k}, \end{aligned}$$

où

$$A_k = A \setminus \left\{\bigcup_{i=1}^{k} S(x_{(i)}, d_i)\right\}$$

est le domaine d'étude total privé de la trajectoire déjà effectuée jusqu'au point d'ordre k. Par convention, $A_0 = A$.

d_k^w est ainsi donné par (4.3), (4.3) et (4.3). Son calcul effectif implique des approximations numériques. Nous avons utilisé la règle du trapèze pour l'intégration numérique dans (4.3). Les intégrales de densité dans (4.3) peuvent être estimées en utilisant la population sous-jacente. Soit W cette population constituée de N individus $\{w_i : i = 1, \ldots, N\}$ avec $N \gg n$. Pour tout ensemble $B \subset A$, $\int_B f(x)dx$ est approximée par $\#\{i/w_i \in B\}$. Ainsi, une potentielle inhomogénéïté de la densité de la population sous-jacente pourra être prise en compte dans le calcul de d_k^w. Pour fixer les choses, prenons l'exemple de l'étude de la répartition spatiale des cas d'une maladie dont la population à risque est la population générale du domaine étudié A. Bien souvent, A est divisée en cellules (administratives par exemple) et la répartition de la population à risque n'est connue qu'à travers sa taille pour chaque cellule (le nombre d'individus de la cellule). La population sous-jacente est alors obtenue en simulant un processus de point uniforme dans chaque cellule avec une taille proportionnelle à la population de la cellule. Les points obtenus pour toutes les cellules représentent les individus w_i et l'ensemble de ces N individus constitue la population sous-jacente W.

La Figure 4.2 illustre le calcul de la pondération pour $k = 21$. La trajectoire déjà effectuée jusqu'à $x_{(20)}$, $\{\cup_{i=1}^{20} S(x_{(i)}, d_i)\} \cap A$, est représentée en blanc. L'aire grise (y compris la portion de disque grisée) est A_{20} et la portion de disque grisée est $A_{20} \cap S(x_{(21)}, r)$. Ainsi,

$$P\left(D_{21} > r | X_{(1)} = x_{(1)}, \ldots, X_{(21)} = x_{(21)}\right) = \left[1 - \frac{\#\{i/w_i \in \text{aire grisée}\}}{\#\{i/w_i \in \text{aire grise}\}}\right]^{n-21},$$

et $E\left[D_{(21)} | X_{(1)} = x_{(1)}, \ldots, X_{(21)} = x_{(21)}\right]$ peut être obtenue par le calcul de la quantité précédente pour un ensemble discret de valeurs de r allant de 0 à a et en utilisant la règle du trapèze.

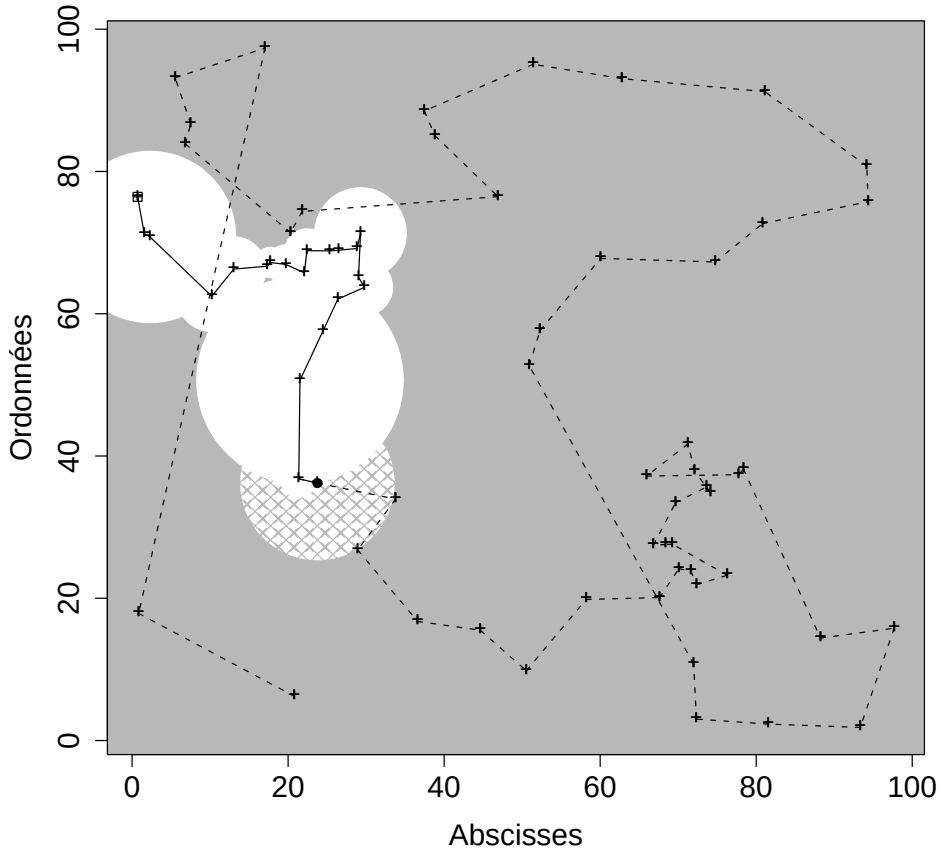


FIG. 4.2 – Illustration du calcul de la pondération. Le point rond et noir est $x_{(21)}$. La portion de disque grisée, $A_{20} \cap S(x_{(21)}, r)$, est représentée dans le cas particulier de $r = d_{21}$.

La série ordonnée des distances pondérées $\{d_k^w : k = 1, \ldots, n-1\}$ obtenue rend maintenant possible la localisation des clusters potentiels.

4.4 Localisation des clusters potentiels

Considérons l'ensemble de données $(k, d_k^w)_{k=1,\ldots,T}$ avec $T = n - 1$. Dans le but de déterminer les bornes des clusters potentiels, nous effectuons la régression de la distance pondérée sur l'ordre de sélection. Sous l'hypothèse d'absence de cluster, une fonction de régression appropriée est la fonction constante

$$f(t) = \overline{d} = \frac{1}{T} \sum_{k=1}^{T} d_k^w \text{ pour } t = 1, \ldots, T \ .$$

Dans ce qui suit, les valeurs de la variable t sont $1, \ldots, T$. Si un cluster est présent, une fonction de régression constante par morceaux avec 2 points de cassure est plus appropriée et

$$f(t) = \overline{d}_{[1;T_1]} \times I_{[1;T_1]}(t) + \overline{d}_{[T_1+1;T_2]} \times I_{[T_1+1;T_2]}(t) + \overline{d}_{[T_2+1;T]} \times I_{[T_2+1;T]}(t) \ ,$$

où T_1 et T_2 sont les points de cassure, la notation $\overline{d}_{[i;j]}$ $(1 \leqslant i < j \leqslant T)$ désigne la moyenne de d_t^w pour t dans $[i; j]$, et $I_{[i;j]}(t) = 1$ si $t \in [i; j]$ et 0 sinon. Le cluster potentiel est la portion entre 2 points de cassure avec la distance moyenne la plus faible. Généralement elle correspond à la portion $[T_1 + 1; T_2]$. Cependant, il peut également être $[1; T_1]$ si le cluster est au début de la trajectoire ou $[T_2 + 1; T]$ s'il est à la fin. Dans ces cas, un modèle à 1 point de cassure est préférable :

$$f(t) = \overline{d}_{[1;T_1]} \times I_{[1;T_1]}(t) + \overline{d}_{[T_1+1;T]} \times I_{[T_1+1;T]}(t) \ ,$$

et le cluster potentiel est $[1; T_1]$ ou $[T_1 + 1; T]$ selon que la distance moyenne la plus faible se situe au début ou à la fin de la trajectoire.

Plus généralement, pour déterminer la présence de m points de cassure ($m + 1$ régimes), la fonction de régression considérée est :

$$f(t) = \sum_{j=1}^{m+1} \overline{d}_{[T_{j-1}+1;T_j]} \times I_{[T_{j-1}+1;T_j]}(t) \tag{4.-3}$$

avec la convention $T_0 = 0$ et $T_{m+1} = T$.

Etant donné la méthode de choix de x_{k+1} parmi le sous-groupe des points encore non sélectionnés, la fonction $f(t)$ sera voisine de 1 tant que la répartition est uniforme, et s'éloignera de 1 dès que la répartition n'est plus uniforme, c'est à dire lorsqu'un cluster apparaît. Donc, graphiquement (voir Figure 4.3), la fonction en escalier $f(t)$ sera proche de 1 sur les marches où il n'y a pas de cluster.

Pour effectuer l'analyse asymptotique dans l'étape de détection, il est nécessaire d'imposer quelques restrictions sur les valeurs possibles des points de cassure. En effet, lors de cette étape, nous avons besoin de tester une hypothèse nulle en présence de paramètres (localisation des points de cassure) qui entrent dans le modèle seulement sous l'hypothèse alternative. Ce problème a été largement traité dans la littérature (Davies [1987], Andrews [1993] et Owen [1991]). Il est à noter que, dans la statistique du scan, le modèle sans cluster (même taux de cas pour toutes les cellules) est la limite de l'hypothèse alternative (des taux plus élevés à l'intérieur d'une zone Z qu'à l'extérieur) lorsque le taux de Z tend vers le taux à l'extérieur de Z, et le paramètre Z n'est pas identifiable dans le situation limite. Cependant le problème est évité dans ce cas particulier puisque la distribution statistique est approchée en utilisant des simulations. Récemment la méthode de Bai et Perron [1998] a pris ce problème en compte dans la modélisation des changements structurels multiples. En particulier, chaque point de cassure doit être asymptotiquement distinct et suffisamment éloigné des bornes de l'échantillon. L'ensemble des partitions possibles est définie comme suit : pour un nombre positif arbitraire $\epsilon \in [0;1]$, $\Delta_\epsilon = \{(T_1, \ldots, T_m)\ ;\ \forall i = 1, \ldots, m+1,\ card\,([T_{i-1}+1; T_i]) \geqslant |T\epsilon|\}$. Par exemple, un ϵ de 0.2 signifie que le nombre de points entre 2 points de cassure doit être au moins de 20% du nombre total de points.

Les points de cassure (bornes du cluster) sont estimés par la résolution du problème des moindres carrés sous contrainte

$$\min_{(T_1,\ldots,T_m)\in\Delta_\epsilon} \sum_{t=1}^{T} \left(d_t^w - f(t)\right)^2 \ . \tag{4.-3}$$

Nous notons $(\hat{T}_1, \ldots, \hat{T}_m)$ la solution de ce problème.

Une méthode, basée sur un algorithme de programmation dynamique, pour efficacement calculer ces estimations est présenté dans Bai et Perron [2003a].

Afin de délimiter la zone couverte par le ou les clusters potentiels, nous déterminons une enveloppe comme suit : nous entourons chaque point localisé dans le cluster par un disque contenant une population égale à la population totale divisée par la taille de l'échantillon (le nombre de cas). Ce disque peut être vu comme la zone d'influence d'un point dans le cas de répartition uniforme. Si la densité de population n'est pas homogène, le rayon n'est pas le même pour tous les disques puisqu'il dépend de la densité de population autour de chaque cas. L'enveloppe est ensuite définie comme l'union de tous les disques. Ainsi, tous les points inclus dans l'enveloppe peuvent être considérés comme faisant parti du cluster.

Un autre façon de construire une enveloppe serait d'utiliser la tessélation de Voronoï, dont une description est donnée par Allard et Fraley [1997]. Cette division naturelle de l'espace en aires disjointes permet de définir une aire d'influence autour de chaque point. Ces aires sont dépendantes des données et ne sont pas obtenues sous l'hypothèse de répartition uniforme. Ceci peut conduire à de grandes aires pour les points situés en bordure du cluster.

Ces deux définitions d'enveloppes peuvent être vues comme complémentaires. On peut choisir entre les deux méthodes ou bien prendre leur intersection. Pour l'exemple illustrant la méthode et pour la simulation de l'influence du premier point, la définition basée sur les disques a été utilisée.

La Figure 4.3 présente les résultats obtenus par les modèles à un et deux clusters sur les même données que celles présentées dans l'exemple de la Figure 4.1. L'union des disques gris représente l'enveloppe du cluster. Les rayons des disques sont pratiquement tous les mêmes car la population sous-jacente est homogène.

4.5 Sélection de modèle

Nous considérons tout d'abord le test de l'absence de points de cassure versus un nombre fixé m de cassures. La statistique de test proposée par Bai et Perron [1998] est

$$F_T(\hat{T}_1,\ldots,\hat{T}_m) = \left(\frac{T-(m+1)}{m}\right)\hat{\delta}' R'(R\hat{V}(\hat{\delta})R')^{-1}R\hat{\delta} = \sup_{(T_1,\ldots,T_m)\in\Delta_\epsilon} F_T(T_1,\ldots,T_m)$$

où $\hat{\delta} = (\hat{\delta}_1,\ldots,\hat{\delta}_{m+1})' = (\overline{d}_{[1;\hat{T}_1]},\ldots,\overline{d}_{[\hat{T}_m+1;T]})'$ et R est une matrice $m\times(m+1)$ telle que $R\hat{\delta} = (\hat{\delta}_1-\hat{\delta}_2,\ldots,\hat{\delta}_m-\hat{\delta}_{m+1})'$. $\hat{V}(\hat{\delta})$ est une estimation de la matrice $(m+1)\times(m+1)$ de variance covariance de $\hat{\delta}$:

$$\hat{V}(\hat{\delta})[i,i] = \hat{\sigma}_i^2\frac{T}{\hat{T}_{i+1}-\hat{T}_i} \text{ pour } i=1,\ldots,m+1 \text{ et } \hat{V}(\hat{\delta})[i,j]=0 \text{ pour } i\neq j\ .$$

$$\hat{\sigma}_i^2 = \frac{1}{\hat{T}_{i+1}-\hat{T}_i}\sum_{j=\hat{T}_i+1}^{\hat{T}_{i+1}}\left(d_j^w - \overline{d}_{[\hat{T}_i+1;\hat{T}_{i+1}]}\right)^2 \text{ pour } i=1,\ldots,m+1.$$

F_T est la statistique qui permet de tester $\hat{\delta}_1 = \ldots = \hat{\delta}_{m+1}$ contre $\hat{\delta}_i \neq \hat{\delta}_{i+1}$ pour un certain i. Une grande valeur de F_T signifie l'éloignement de l'hypothèse d'absence de points de cassure. Les valeurs critiques de cette statistique de test sont données dans Bai et Perron [2003b] pour des valeurs de ϵ comprises entre 0.05 et 0.25.

Dans l'exemple de la Figure 4.3, les valeurs de F_T sont 6.39, 9.71 et 5.6 respectivement pour les modèles à 1, 4 et 8 points de cassure. Seul le premier n'est pas significatif étant donné que les valeurs critiques sont respectivement 9.1, 6.84 et 3.58. Les modèles avec 3, 5, 6 et 7 points de cassure sont également significatifs (valeurs non présentées). Ainsi le modèle à 1 point de cassure représenté sur les Figures 4.3(a) et 4.3(b) n'est pas significatif alors que les modèles à 4 et 8 points de cassure sur les Figures 4.3(c) et 4.3(d) sont significatifs.

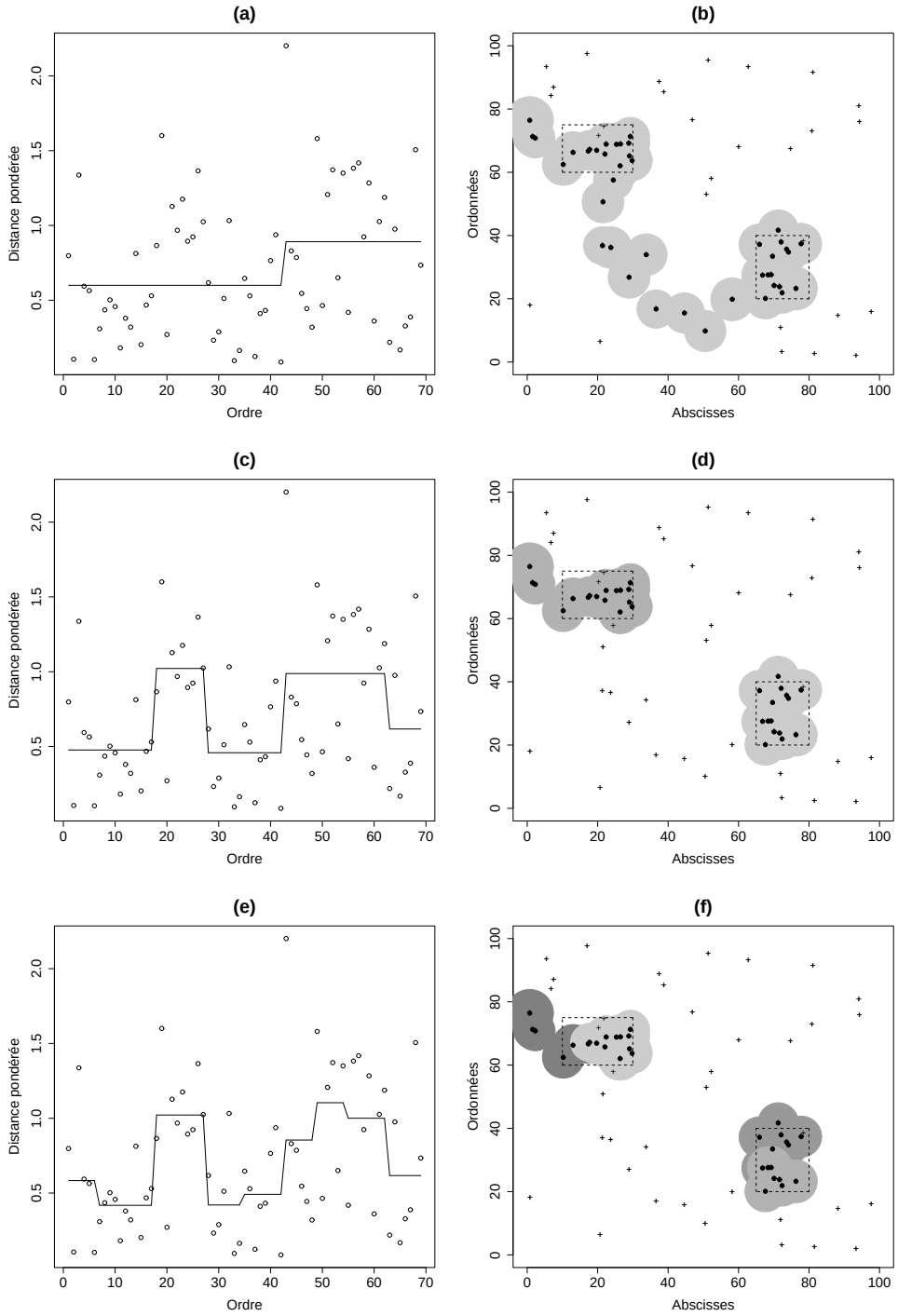


FIG. 4.3 – Résultats pour trois modèles sur les données simulées de la figure 4.1 : régression de la distance sur l'ordre et représentation du ou des clusters localisés par **(a)** et **(b)** le modèle à un cluster avec un point de cassure, **(c)** et **(d)** le modèle à deux clusters avec 4 points de cassure et **(e)** et **(f)** le modèle à deux clusters avec 8 points de cassure. Les points localisés dans le ou les clusters sont les points ronds, entourés par un disque gris. Des niveaux de gris différents sont utilisés pour différencier les portions, lorsque c'est nécessaire.

Le meilleur modèle doit maintenant être sélectionné et le nombre de points de cassure déterminé, tout en prenant en compte le problème des tests multiples. Le test du double maximum, défini par Bai et Perron [1998], permet de tester l'hypothèse nulle d'absence de cassure contre un nombre inconnu de cassures étant donnée une certaine borne supérieure M. Soit $c(\alpha, m)$ la valeur critique asymptotique du test $F_T(\hat{T}_1, \ldots, \hat{T}_m)$ pour un risque de première espèce α. Le test est noté :

$$WD\ max\ F_T(M) = \max_{1 \leqslant m \leqslant M} \frac{c(\alpha, 1)}{c(\alpha, m)} F_T(\hat{T}_1, \ldots, \hat{T}_m)\ .$$

Les valeurs critiques de cette statistique de test corrigée sont données dans Bai et Perron [2003b] pour des valeurs de ϵ comprises entre 0.05 et 0.25 et pour $M \leqslant 9$. Etant donné que ϵ représente la taille minimum des clusters potentiels, le choix de sa valeur dépend de considérations cliniques. Les spécialistes devraient aider les statisticiens à choisir ϵ parce la taille minimale du cluster a une interprétation spécifique. De plus, nous recommandons de ne pas choisir une valeur plus élevée que 0.2 afin d'obtenir un risque de première espèce proche de 0.05. Dans ce qui suit, nous avons choisi $\epsilon = 0.1$. Avec cette valeur, nos simulations ont donné une erreur de première espèce de 0.062.

Le nombre de points de cassure est ensuite choisi comme l'*argmax* de la statistique $WD\ max$.

Dans l'exemple de la Figure 4.3, l'*argmax* est 8 points de cassure. La valeur de la statistique est 16.60 qui est plus grande que la valeur critique pour $M = 8$ (10.39). Ainsi le modèle à 8 points de cassure, correspondant à 2 clusters qui regroupent chacun 2 portions, est significatif.

4.6 Détection des clusters

Le meilleur modèle sélectionné contient une ou plusieurs portions (clusters potentiels). Si le meilleur modèle a une statistique $WD\ max$ significative, l'étape de détection consiste à calculer la p-valeur associée à chaque portion. Dans cette optique, nous calculons la densité dans l'enveloppe de chaque portion. Cette densité est le ratio du nombre de cas dans l'enveloppe sur le nombre d'individus de la population dans l'enveloppe. Nous simulons également 9999 valeurs de la densité sous l'hypothèse d'absence de cluster (pour chacun des 9999 échantillons simulés sous H_0, nous retenons la densité la plus élevée parmi les densités des portions du modèle sélectionné) et nous déterminons, pour chaque portion, le rang de la densité dans le vecteur ordonné des 9999 densités. La p-valeur est ensuite obtenue en divisant ce rang par 10000. Si deux (ou plusieurs) portions ont une intersection de leurs enveloppes qui est non vide, la densité de cas est calculée pour l'union des deux (ou plusieurs) enveloppes.

Dans l'exemple présenté en Figure 4.3, le meilleur modèle contient 4 portions qui doivent être regroupées en deux enveloppes distinctes (une dans le coin supérieur gauche et l'autre dans le coin inférieur droit). L'enveloppe de la portion représentée dans le coin supérieur gauche (deux niveaux de gris) regroupe 20 cas et 50 individus (densité de 0.40) et $p = 0.0003$. L'enveloppe dans le coin inférieur droit (deux niveaux de gris également) regroupe 16 cas et 46 individus (densité de 0.35) et $p = 0.0032$. Dans cet exemple, les deux clusters simulés sont significatifs.

4.7 Influence du premier point

Dans le but d'étudier l'influence du premier point sur la localisation des clusters, un jeu de données de 70 points avec un cluster de 20 points a été simulé. La population sous-jacente est une grille régulière 32×32. Pour chacun des 70 points, la méthode de localisation des clusters a été appliquée, avec chaque point pris comme premier point de la trajectoire. Seul les clusters significatifs au test $WD\ max$ ont été retenus. Le test $WD\ max$ n'était pas significatif dans seulement 2 cas sur 70. Ainsi, nous comptons le nombre de fois où les points de la grille tombent dans l'une des 68 enveloppes de cluster significatif. La Figure 4.4 présente les résultats de cette étude et confirme la robustesse de cette méthode concernant le choix du premier point de la trajectoire.

La méthode peut ainsi être appliquée est choisissant arbitrairement le premier point de la trajectoire. Une règle de détermination d'un tel point peut être par exemple le point le plus proche du bord du domaine d'étude, ou le point le plus proche de l'un des coins (dans le cas d'un domaine rectangulaire).

4.8 Résultats

4.8.1 Etude de puissance et simulations

Nous avons simulé des échantillons de données de 100 points avec différentes situations d'agrégation. Deux zones de simulation de clusters sont définies $C_1 = [20; 60] \times [75; 85]$ et $C_2 = [70; 80] \times [20; 60]$. C_0 désigne l'ensemble du domaine d'étude ($A = [0; 100] \times [0; 100]$) privé des deux zones des clusters simulés : $C_0 = A \setminus (C_1 \cup C_2)$. Pour $i = 0, 1, 2$, γ_i désigne la densité de cas dans C_i, soit encore le ratio entre le nombre de cas C_i et le nombre d'individus dans C_i. Les différentes situations d'agrégation sont : pas de cluster ($\gamma_0 = \gamma_1 = \gamma_2$), un cluster simulé dans C_1 avec une densité k fois plus élevée à l'intérieur de C_1 et avec k successivement égal à 3, 6 et 10 ($\gamma_1 = k \times \gamma_0$ et $\gamma_2 = \gamma_0$), et deux clusters simulés dans C_1 et C_2 avec une densité de cas 6 fois plus élevée à l'intérieur de C_1 et C_2. La population de fond W est une grille régulière de taille 32×32. 1000 échantillons ont été simulés pour chaque situation d'agrégation. Pour chaque échantillon, la méthode a été appliquée et la statistique du test du double maximum, la densité et la p-valeur ont été calculées pour $M = 8$ et $\epsilon = 0.1$. Ensuite, pour chacune des cinq stratégies d'agrégation,

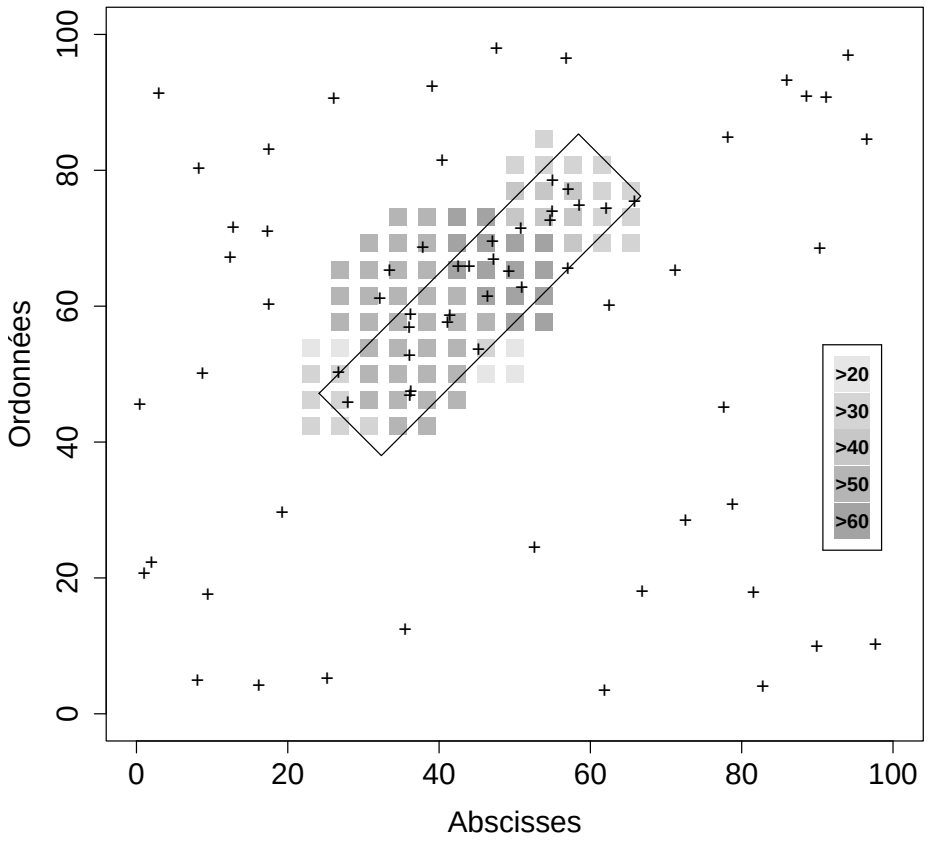


FIG. 4.4 – Illustration de l'influence du choix du premier point sur des données simulées ($n = 70$) suivant un mélange de deux processus de points uniformes $\frac{5}{7} \times \mathcal{U}([0,100]^2) + \frac{2}{7} \times \mathcal{U}(C)$ où C est le rectangle. Chacun des n points a été pris successivement comme premier point de la trajectoire. Le nombre de fois où le point est localisé dans un cluster significatif est représenté par le niveau de gris des points carrés. Par exemple, "> 40" signifie que les points de cette couleur ont été localisés entre 41 et 50 fois sur les 70 modélisations.

nous avons calculé le nombre d'erreurs de classification (faux positifs et faux négatifs) et la proportion d'échantillons présentant un ou plusieurs clusters significatifs ($p < 0.05$). Le nombre d'erreurs a été compté à la fois sur le nombre total de réplicats et sur les échantillons présentant un cluster significatif. Nous avons également appliqué la statistique de scan spatiale dans le but de déterminer les mêmes quantités et de comparer les deux méthodes.

Dans la situation d'absence de cluster, les résultats ont montré une erreur de première espèce de 6.2% pour notre méthode et de 6% pour la statistique de scan spatial. Le nombre d'erreur de classification ne pouvait être calculé dans ce cas puisqu'aucune zone de simulation de cluster n'est définie. Les résultats pour les autres situations d'agrégation sont présentés dans la Table 4.1. La statistique de scan spatial est plus puissante que notre méthode. Cependant, la statistique de scan spatial localise le cluster simulé avec moins de précision que notre méthode puisque la zone de simulation du cluster est allongée. La statistique circulaire ne peut pas ajuster la zone rectangulaire allongée. En effet, globalement, le nombre d'erreurs est plus élevé avec la statistique de scan. Ceci est dû à un nombre de faux positifs deux fois plus élevé. Le nombre de faux négatifs est quant à lui légèrement plus faible avec la statistique de scan. Ces résultats indiquent que la fenêtre circulaire maximisant la vraisemblance englobe la totalité de la zone de simulation, ce qui implique un nombre de faux positive élevé et un nombre de faux négatifs faible. Par ailleurs, la différence du nombre d'erreurs entre les deux méthodes est accentué lorsque le comptage est effectué uniquement sur les clusters significatifs. Ceci signifie que lorsque notre méthode détecte un cluster significatif, elle le localise précisément.

Pour illustrer la flexibilité de notre méthode, nous avons simulé un échantillon de 70 points avec un cluster simulé de 30 points en forme de "L". La zone de simulation du cluster représente 6% de l'aire totale. Nous avons défini une grille régulière de taille 27×27 comme population de fond W. Le résultat de la localisation des clusters est présenté sur la Figure 4.5 avec les deux types d'enveloppes. La valeur de la statistique $WD\,max$ était 25.98, valeur plus élevée que la valeur critique pour $M = 9$ et $\epsilon = 0.1$. L'hypothèse d'absence de points de cassure a été rejetée et le modèle avec 8 points de cassure a été sélectionné. Les quatre portions sont significatives ($p < 0.05$) et leur réunion forme un cluster en forme de "L" qui est lui aussi significatif ($p = 0.0002$). Le cercle de Kulldorff représenté sur la figure est significatif ($p < 0.001$). Il faut noter qu'un large quart de ce cercle est vide de points ce qui illustre que la statistique de scan n'est pas adaptée à la localisation de clusters ayant une forme très différente du cercle.

4.8.2 Agrégats de pharmacies à Montpellier

Une première application pratique de la méthode a été de vérifier l'uniformité de la répartition des pharmacies à Montpellier. En effet, la localisation géographique des pharmacies est supposée dépendre de la population environnante.

Densité de cas relative $\frac{\gamma_1}{\gamma_0}$ dans C_1	1	3	6	10	6	
Densité de cas relative $\frac{\gamma_2}{\gamma_0}$ dans C_2	1	1	1	1	6	
Puissance						
Nous	0.062	0.376	0.908	0.987	0.936	
Kulldorff	0.060	0.485	0.995	1	/	
Erreurs						
Nous	/	11.671	10.479	9.131	11.391	11.166
Kulldorff	/	15.377	11.321	10.377	/	
Faux positifs						
Nous	/	1.206	3.297	3.112	2.926	3.093
Kulldorff	/	6.626	7.142	6.541	/	
Faux négatifs						
Nous	/	10.466	7.182	6.019	8.465	8.073
Kulldorff	/	8.751	4.419	3.836	/	
Erreurs sur les clusters significatifs						
Nous	/	7.785	8.774	8.834	10.479	10.359
Kulldorff	/	17.035	11.229	10.370	/	
Faux positifs sur les clusters significatifs						
Nous	/	3.205	3.631	3.153	3.126	3.304
Kulldorff	/	13.662	7.178	6.541	/	
Faux négatifs sur les clusters significatifs						
Nous	/	4.580	5.143	5.681	7.353	7.054
Kulldorff	/	3.373	4.051	3.836	/	

TAB. 4.1 – Résultats pour l'étude de puissance.

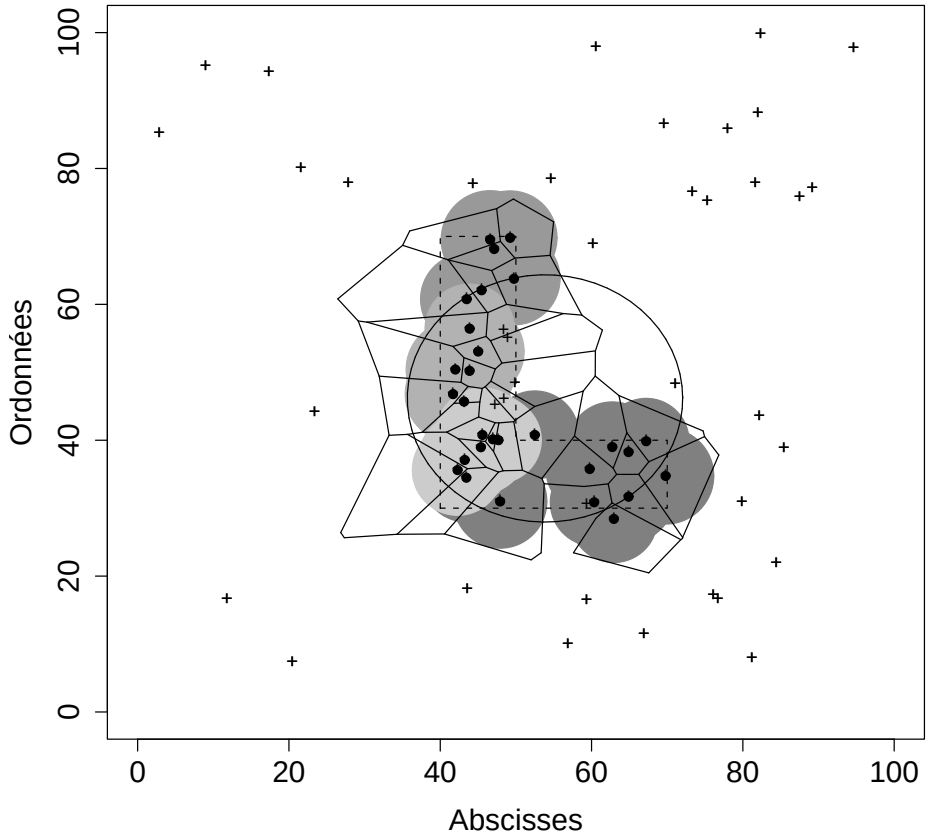


FIG. 4.5 – Echantillon de 70 points avec un cluster de 30 points en forme de "L". La zone de simulation du cluster est représentée par des pointillés. Les points localisés dans le cluster (modèle avec 8 points de cassure) sont représentés par un point noir, et entourés par un disque gris et un polygone de Voronoï. L'union de ces disques gris représente l'enveloppe du cluster basée sur les disques. L'union des polygones représente l'enveloppe du cluster basée sur Voronoï. Le cercle représente le cluster le plus probable localisé par la statistique de scan spatial.

Les 99 pharmacies de Montpellier ont été localisées par GPS (Global Positioning System). Dans le but de prendre en compte les variations de la densité de population sous-jacente, nous avons utilisé le découpage de Montpellier en 30 IRIS. L'IRIS a été défini par l'INSEE en 2000. La population de Montpellier utilisée dans cette définition est celle du recensement de 1999. La densité de population dans chaque IRIS est représentée sur la Figure 4.6 (a). Dans un souci de clarté pour la présentation des résultats, nous avons affiché le numéro de chaque IRIS au centre de chacun d'entre eux. La répartition des pharmacies est représentée sur la Figure 4.6 (b). Le taux de pharmacies sur la totalité de l'aire d'étude est de 0.44 pour 1000 habitants.

Les cas (pharmacies) et les données utilisées pour l'ajustement (population) n'ont pas le même niveau d'agrégation : on connaît la localisation exacte des pharmacies (données ponctuelles) alors que seule la taille de la population est connue pour chaque IRIS (données groupées). L'approche classique est d'agréger les deux types de données au même niveau comme nous l'avons fait pour les statistiques de scan SA et ULS. Dans le but d'appliquer notre méthode, nous avons construit la population sous-jacente en simulant un processus de point uniforme dans chaque IRIS avec une taille proportionnelle à la population de l'IRIS. Notre méthode a détecté un cluster significatif de pharmacies dans le centre ville. Il est représenté sur la Figure 4.6 (c) en gris foncé. Le modèle avec 2 points de cassure (un cluster) a été sélectionné et la valeur de la statistique du test $WD\ max$ était 40.5 et $p = 0.0002$. L'enveloppe de ce cluster regroupe 14 pharmacies pour une population de 7000 individus (taux de 2 pharmacies pour 1000 individus). Ce résultat peut être expliqué par le grand nombre de personnes qui se trouvent dans le centre ville pendant la journée, ce qui augmente considérablement la population à cet endroit et nécessite la présence de plus de pharmacies qu'à la périphérie de la ville. Il faut également noter que le rayon des disques de l'enveloppe sont pratiquement égaux puisque la population de fond est homogène dans les deux IRIS (25 et 30) qui contiennent les pharmacies localisées dans le cluster.

Nous avons également appliqué la statistique de scan ULS (Patil et Taillie [2004]) et la méthode SA (Duczmal et Assunção [2004]). Comme ces méthodes ne sont pas adaptées aux données individuelles, nous avons calculé le nombre de pharmacies dans chaque cellule (IRIS) dans le but d'obtenir des données groupées. Pour ces deux méthodes, le critère d'adjacence entre deux cellules était "leur frontière commune a une longueur positive".

Pour la méthode ULS, puisqu'aucun logiciel fiable n'est disponible pour cette méthode de calcul intensif, nous n'avons appliqué qu'aux données la procédure basée sur les arbres permettant de localiser le cluster le plus probable, et non aux réplicats (Ω_{ULS} doit être recalculé pour chaque réplicat et des algorithmes performants sont nécessaires pour ce calcul). Ainsi nous ne pouvons pas fournir de valeur pour la probabilité p exprimant la significativité du cluster le plus probable trouvé à partir des données. Ce dernier est représenté sur la Figure 4.6 (c) en gris clair (il est composé des IRIS numérotés 12, 20, 25, 26 et 30). Il regroupe 33 pharmacies pour une population de 38800 individus (taux de 0.85/1000). Il est

à noter que ce cluster le plus probable est le même que celui localisé par la statistique du scan spatial pour données groupées. Ce dernier est significatif ($p = 0.007$).

Pour appliquer la méthode SA, nous avons utilisé l'algorithme implémenté en code C++ que les auteurs nous ont transmis. Le cluster le plus probable est représenté sur la Figure 4.6 (c) par deux nuances de gris clair (il est composé des IRIS numérotés 12, 20, 25, 26, 30, 16 et 18 - ce dernier IRIS n'est représenté que partiellement). Il correspond à celui localisé par la méthode ULS plus deux cellules sur la droite, et regroupe 40 pharmacies pour une population de 51600 individus (taux de 0.78/1000). Ce cluster le plus probable n'est pas significatif ($p = 0.2$, obtenue avec 999 réplicats de Monte Carlo). Ce résultat non significatif est une illustration de la puissance relative de la méthode SA et de la statistique de scan de Kulldorff lorsque le vrai cluster a une forme circulaire, ce qui semble être le cas ici. Ce problème est mentionné par Duczmal et Assunção [2004] dans leur conclusion.

Finalement nous avons choisi d'appliquer la statistique de scan spatial circulaire aux données individuelles. Le modèle de Poisson a été utilisé avec la population sous-jacente précédemment simulées. Le cluster localisé par notre méthode semble avoir une forme circulaire ce qui justifie l'utilisation de cette méthode. Le cluster le plus probable localisé par la méthode de Kulldorff est représenté par un disque sur la Figure 4.6 (c). Ce cluster est significatif ($p = 0.002$) et regroupe 16 pharmacies pour une population de 5400 individus (taux de 2.97/1000).

Cet exemple nous permet de comparer les méthodes adaptées aux données individuelles à celles adaptées aux données groupées. Les clusters localisés par la méthode de Kulldorff et la nôtre sont inclus dans les clusters les plus probables localisés par les méthodes ULS et SA, et ceci est cohérent puisque, par définition, ces méthodes ne peuvent détecter un cluster ayant une résolution plus faible que celle des cellules. De plus, parmi les 5 ou 7 cellules qui forment respectivement les clusters ULS et SA, les deux méthodes pour données individuelles ont localisé les zones (ne correspondant pas à des cellules dans leur totalité) avec les taux de pharmacies les plus élevés. Par exemple, notre cluster n'inclut que la partie de l'IRIS 25 concentrant la majorité des pharmacies de cet IRIS. Cet exemple illustre également, comme on pouvait s'y attendre, l'avantage de la méthode de Kulldorff lorsque le vrai cluster a une forme circulaire. Le taux de la zone circulaire est plus élevé que celui de notre cluster (2.97/1000 versus 2/1000). Cependant, cette remarque doit être nuancée puisque le contour de la zone circulaire passe par construction par la localisation exacte d'un cas - la pharmacie la plus éloignée du centre du disque - alors qu'avec notre méthode il est peu probable que le contour de l'enveloppe du cluster passe par un cas.

4.8.3 Leucémie et lymphôme chez les enfants dans le comté de North Humberside, Angleterre.

Nous avons également appliqué notre méthode sur un jeu de données précédemment analysé par Cuzick et Edwards [1990]. 62 cas d'enfants atteints de leucémie ou de lymphôme

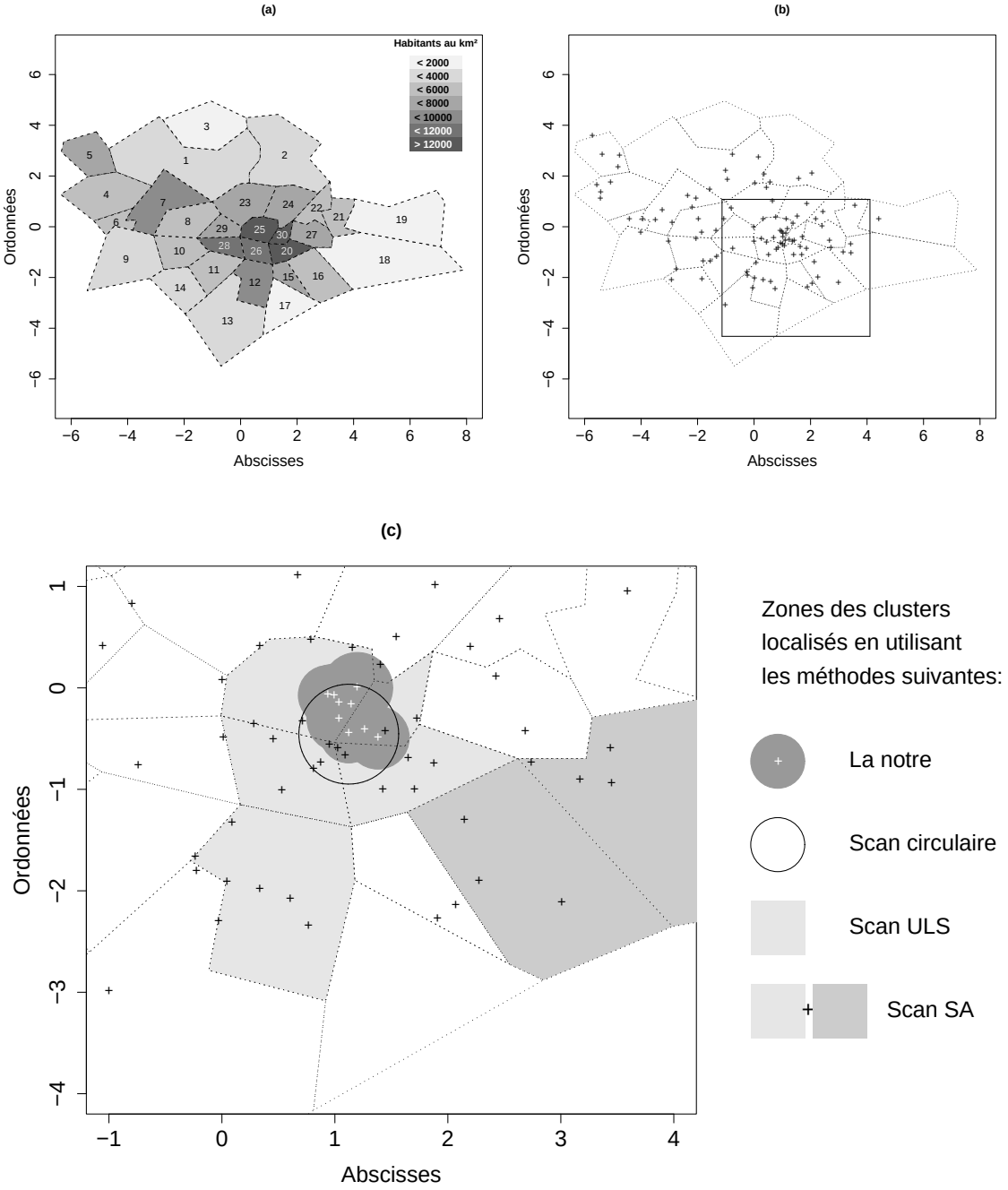


FIG. 4.6 – **(a)** Carte de la densité de population à Montpellier. La densité dans chaque IRIS est représentée par des niveaux de gris. Pour chaque IRIS, le numéro d'identification est écrit en son centre. **(b)** Répartition des pharmacies à Montpellier. Chaque croix représente une pharmacie. **(c)** Agrandissement de la zone carrée de la figure (b) avec les zones des clusters localisés en utilisant les différentes méthodes. L'unité des axes est le kilomètre.

ont été diagnostiqués entre 1974 et 1986 dans le comté de North Humberside et 141 contrôles ont été sélectionnés au hasard dans les registres des naissances. Leur répartition spatiale est représentée sur la Figure 4.7.

Les clusters potentiels localisés par la statistique de scan spatial et notre méthode ne sont pas significatifs. La valeur de la statistique du test $WD\ max$ est 2.47, plus faible que la valeur critique pour $M = 8$ et $\epsilon = 0.1$ (9.42). Le ratio de la log vraisemblance est de 4.84 pour la statistique de scan ($p = 0.676$). Ces clusters les plus probables sont représentés par des aires grises sur la Figure 4.7. Pour ces deux méthodes, les 203 cas et contrôles ont été utilisés pour la population sous-jacente. La statistique de scan spatial a été appliquée en utilisant le modèle de Bernouilli.

4.9 Programmation

Les programmes de la méthode de régression sur données transformées ont été écrits en langage *R* (R Development Core Team [2006]) et ont fait l'objet d'un package de contribution au CRAN, le réseau complet d'archives de *R*. Ce package est nommé SPATCLUS (Dematteï [2006]). Sa description, décrite dans Dematteï et al. [2006b], est présentée en annexe A.

4.10 Discussion

La méthode présentée ici a l'avantage d'être très flexible. Premièrement, elle peut être utilisée pour détecter et localiser plusieurs clusters, sans besoin de faire appel à un ajustement pour tests multiples. Deuxièmement, puisque la méthode ne nécessite pas la définition d'une forme prédéfinie pour les clusters potentiels, les agrégats détectés peuvent être de n'importe quelle forme.

Cette méthode a été conçue uniquement pour les données ponctuelles. Comme précisé dans l'introduction de ce chapitre, avec les données ponctuelles, lorsque celles ci sont disponibles, l'information spatiale détaillée dont on dispose n'est pas perdue. Par ailleurs, notre méthode est libre de toute partition du domaine d'étude. Comme expliqué par Duczmal et Assunção [2004], le choix de la taille des cellules de comptage peut affecter les résultats de la localisation et de la détection de clusters. Ceci est illustré dans l'exemple des pharmacies. En effet, les cas sont souvent localisés près d'une frontière entre deux cellules puisque ces frontière correspondent souvent à des grands axes routiers dans une ville ou à une frontière naturelle, telle qu'une rivière, dans les études à une échelle plus grande. Ceci peut mener à des ambiguïté dans l'attribution d'un cas à une cellule particulière, ce qui est remarqué par Turnbull et al. [1990] dans l'exemple des données d'incidence de leucémie à New York. Ainsi, un léger décalage ou une ambiguïté dans la localisation des cas peut grandement affecter le comptage dans chaque cellule et changer la localisation et la détection des clusters. De

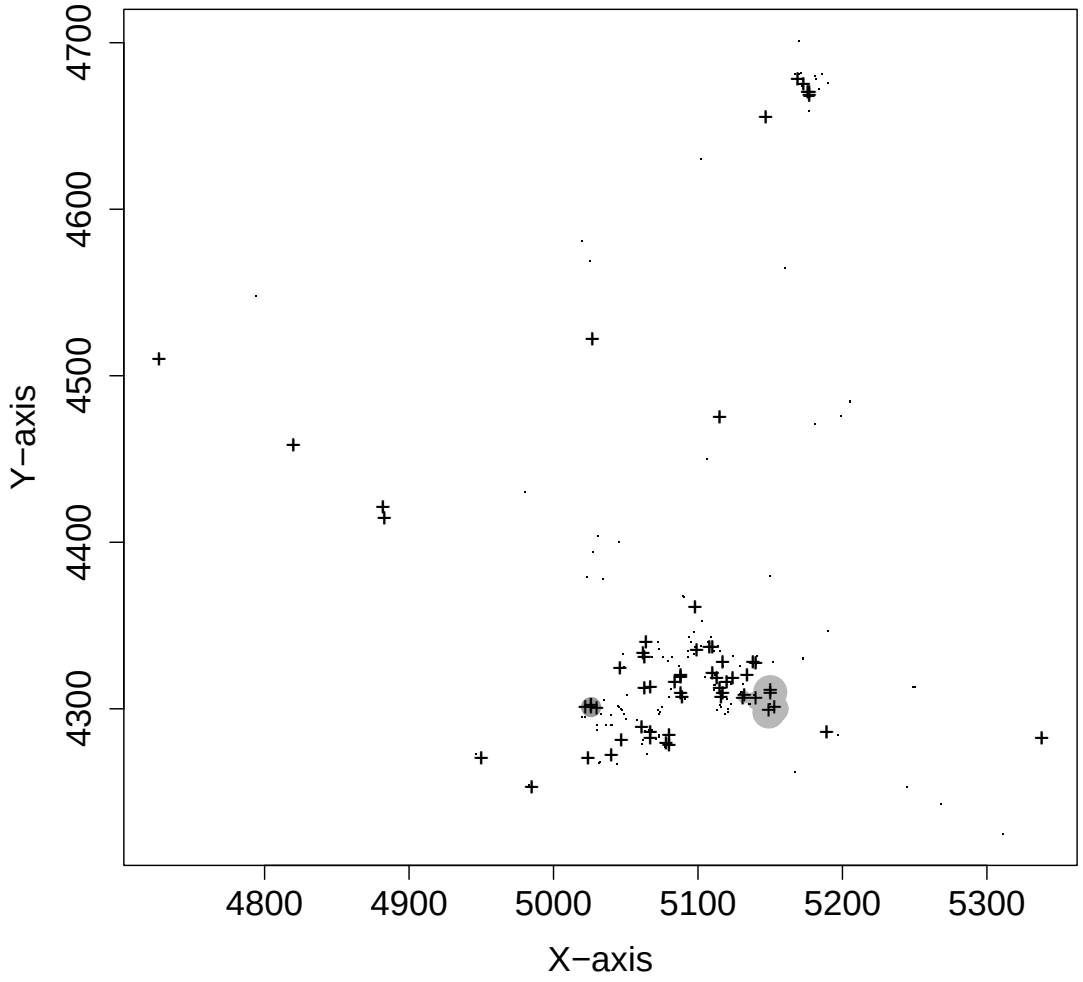


FIG. 4.7 – Répartition des cas de leucémie et de lymphôme chez les enfants dans le comté de North Humberside. Les cas sont représentés par une croix et les contrôles par un point. L'aire grise représente l'enveloppe basée sur les disques du cluster le plus probable localisé par notre méthode. Le petit disque gris centré sur les coordonnées (5020,4300) représente le cluster le plus probable localisé par la statistique de scan spatial.

plus, nous pensons que la méthode présentée ici est préférable aux méthodes pour données groupées lorsque les données ponctuelles sont disponibles.

Bien que notre méthode se rattache à la catégorie des tests de la répartition spatiale avec ajustement sur la non-homogéneïté de la population sous-jacente, elle ne rentre pas dans le cadre général proposé par Kulldorff [2002]. L'écart le plus évident de cette méthode à ce cadre est la transformation des données conditionnellement à la trajectoire (la valeur attribuée à un point, dans notre cas un ratio de distances, dépend des points déjà sélectionnés). Nous sommes conscients que de nombreux tests existent déjà (Kulldorff [2002]). Cependant, cette méthode est innovante et peut être vue comme une façon d'analyser les données spatiales complémentaire aux méthodes habituellement utilisées.

La première limite de cette méthode est la possibilité laissée à la trajectoire de quitter le cluster avant d'être passer par tous les points du cluster. Ceci est cependant en règle générale un faux problème. En effet, les points restant dans le cluster seront détectés comme un second cluster et une analyse visuelle de la proximité des deux clusters détectés permettra de les regrouper en un nouveau cluster plus important.

La seconde limite concerne la disponibilité des données ponctuelles. En effet, ce type de données n'est pas toujours facile à recueillir et les données groupées sont souvent préférées. Dans un tel cas, notre méthode ne peut être appliquée. Cependant, comme cela a été remarqué par Bailey [2001], les systèmes d'information de santé améliorent continuellement le recueil des données ponctuelles. Il s'opère de ce fait actuellement un accroissement de la demande de méthodes pouvant être utilisée sur des données ponctuelles.

L'extension de cette méthode à un processus spatial de dimension $n > 2$ est immédiate, en remplaçant la distance utilisée ici par la distance euclidienne dans $\mathbb{R}^n$. Une fois que les données ont été transformées, la méthode est strictement la même. Une autre extension de cette méthode, moins simple, est son adaptation à la détection de clusters spatio-temporels. Dans ce problème, la principale difficulté est le rôle différent que joue la dimension temporelle par rapport aux dimensions spatiales. Les applications possibles dans $\mathbb{R}^{3+1}$ (volume + temps) sont la détection de cluster en Imagerie par Résonance Magnétique fonctionnelle (IRMf) ou pour les données météorologiques. Le traitement des données IRMf est abordé dans la partie suivante.

Une autre amélioration qui devrait être étudiée est l'ajustement sur des covariables. La méthode proposée ici ajuste uniquement sur une non-homogéneïté de la population de fond. L'ajustement sur des covariables telles que l'âge ou le sexe n'est pas encore possible.

Tous les calculs ainsi que toutes les figures ont été effectués avec le logiciel R. Pour les utilisateurs qui seraient intéressés par l'application de cette méthode, nous pouvons fournir sur simple demande une implémentation des programmes au sein d'un package R, nommé SPATCLUS.

Troisième partie

Application à l'Imagerie par Résonance Magnétique fonctionnelle

Introduction

L'Imagerie par Résonance Magnétique fonctionnelle (IRMf) est une méthode de mesure indirecte de l'activité cérébrale. Lorsqu'une région du cerveau est activée, on observe une augmentation du flux sanguin local de cette région, ainsi qu'une augmentation du taux d'oxygène local. Cette technique mesure un contraste BOLD (Blood Oxygen Level Dependant). La méthode est non invasive et présente une bonne résolution spatiale (de l'ordre du mm). Elle permet d'obtenir des images anatomiques et fonctionnelles.

Depuis l'apparition de la méthode d'acquisition d'image cérébrales par PET (Positron Emission Tomography) au début des années 80, puis avec celle par IRMf au début des années 90, des méthodes et des logiciels d'analyse statistique propres aux données IRMf ont été mis en place du fait de la grande quantité de données et de leurs dépendances spatiale et temporelle.

Cette partie dédiée à la détection de clusters en IRMf est divisée en deux chapitres. Le premier décrit la méthode la plus couramment utilisée pour traiter les données obtenues par acquisition IRMf, appelée Statistical Parametric Mapping (SPM). Puis nous présenterons une approche originale qui se propose d'appliquer la méthode de régression sur données transformées, décrite dans le chapitre précédente, aux cartes d'activation SPM.

Chapitre 5

Méthode standard de détection de clusters en IRMf

L'objet de ce chapitre est de présenter le cadre méthodologique spécifique de l'analyse de données IRMf. Ceci permettra au lecteur de se familiariser avec ce type de données bien particulier. La lecture de ce chapitre n'est cependant pas indispensable à la bonne compréhension des parties suivantes. Quelques rappels essentiels sur les données IRMf seront effectués au début du chapitre 6.

L'objectif de l'analyse standard de données IRMf est d'aboutir à une carte d'activation statistique en 3 dimensions (3D) à partir de centaines d'images acquises sous différentes conditions expérimentales. De nombreuses étapes sont nécessaires afin de passer des données brutes produites par l'appareil d'acquisition aux "SPMs", cartes 3D de statistiques paramétriques, permettant de localiser les zones cérébrales activées.

Les données initiales font tout d'abord l'objet de pré-traitements permettant de corriger le décalage temporel d'acquisition des différentes coupes d'un scan, de corriger les effets des mouvements de la tête du sujet pendant la session IRMf, de normaliser spatialement les images afin de les placer dans un espace anatomique standard, et enfin de lisser spatialement les données pour minimiser le bruit par rapport au signal et assurer la validité des inférences basées sur des tests paramétriques.

Une fois les données pré-traitées, la phase d'estimation des paramètres qui peuvent expliquer ces données s'effectue en utilisant le modèle linéaire généralisé. A partir de ces estimations les statistiques de test sont calculées pour chaque voxel dont on fait la carte statistique - paramétrique - d'où le nom de SPM (Statistical Parametric Map). A partir de

ces cartes (il y en a une par contraste testé), il faut ensuite déterminer le seuil au delà duquel il sera justifié d'accorder la significativité de l'activation, de façon à maintenir l'erreur de première espèce à un niveau raisonnable (minimiser la proportion de faux positifs) tout en maximisant la puissance de l'analyse (minimiser la proportion de faux négatifs).

5.1 Les données IRMf : terminologie et particularités

Les données utilisées en IRMf sont des images 3D du cerveau appelées scans. Ces images sont découpées en voxels (équivalents 3D du pixel). Les différents scans sont acquis au cours du temps alors que le sujet est soumis à différentes conditions expérimentales. En général, une condition expérimentale recouvre plusieurs scans successifs. La durée d'acquisition d'un scan est appelée TR (Temps de Répétition). La donnée recueillie dans chaque voxel du cerveau et chaque scan est le signal BOLD qui mesure le niveau neuronal d'activité du voxel. Pour chaque voxel, la variable réponse est ainsi un vecteur $N \times 1$, où N est le nombre de scans de la série d'acquisition temporelle. La présence de chaque condition codée en 0 ou 1 constitue les variables explicatives[1]. Celles-ci peuvent donc être considérées comme des fonctions temporelles ayant la forme de box-car (fonctions en créneaux). D'autres variables dites de confusion, telles que la session (série d'acquisition) ou le sujet, sont prises en comptes en tant que covariables.

Les données IRMf présentent des caractéristiques particulières. Tout d'abord, la quantité de données traitées est immense. Considérons à titre d'exemple une série d'acquisition fonctionnelle d'une durée de 20 minutes, un TR de 3 secondes et une résolution de $2 \times 2 \times 2\ mm^3$ (tailles des voxels). On obtient pour chacun des 200000 à 300000 voxels, une série temporelle de 400 points, ce qui donne finalement un nombre de valeurs du BOLD compris entre 80 et 120 millions. Le tout pour un seul sujet et une seule session d'acquisition. Par ailleurs, ces données ne sont bien évidemment pas indépendantes les unes des autres, que ce soit spatialement ou temporellement (la réponse neuronale suite à un stimuli a une étendue spatiale de l'ordre de 2 à 5 mm et s'étend temporellement sur plus de 30 secondes après l'induction). Il faut bien évidemment estimer et tenir compte dans la modélisation du signal de ces corrélations spatiales et temporelles. L'inférence est elle aussi affectée par ces corrélations, mais aussi par la grande quantité de données amenant à des comparaisons multiples. Enfin, dans le cas d'études portant sur plusieurs sujets, il faut prendre en considération les différences (en taille et en localisation) des aires anatomiques et fonctionnelles entre les cerveaux de chaque sujet.

[1] Nous verrons dans la partie 5.3 que ce ne sont pas directement ces variables binaires qui sont utilisées comme régresseurs.

5.2 Les pré-traitements et leurs justifications

Les données brutes issues de l'appareil d'acquisition IRMf ne sont pas directement exploitables. Certains pré-traitements (avant tout traitement statistique) sont nécessaires afin que les étapes de modélisation, d'inférence et d'interprétation soient par la suite effectuées dans de bonnes conditions. Dans ce qui suit, chacun des traitements nécessaires à l'obtention de données statistiquement exploitables sera justifié et détaillé d'un point de vue méthodologique.

5.2.1 Suppression des 4 premières images

La stabilisation du signal IRMf n'est pas atteinte immédiatement lors du lancement de l'acquisition. On considère généralement que les 4 premiers scans ne sont pas fiables (Anton et al. [2001]) (le nombre dépend en fait du TR). Ces premières images ne sont donc pas prises en compte dans la suite des traitements.

5.2.2 Correction des décalages temporels

Chaque scan est constitué de coupes (slices). Les différentes coupes d'un scan sont généralement acquises en mode entrelacé, l'acquisition des coupes impaires s'effectuant avant celle des coupes paires, le tout du bas vers le haut. Dans certains cas, le mode d'acquisition des différentes coupes est séquentiel, les coupes étant acquises successivement du bas vers le haut. Dans les deux cas, les différentes coupes d'un scan ne sont pas acquises au même instant. La correction des décalages temporels, également appelée "slice timing", consiste à ramener, par interpolation temporelle, l'instant d'acquisition de toutes ces coupes à un instant commun (celui de la coupe choisie comme référence). En général la coupe de référence est celle qui se situe au milieu du cerveau.

Lorsque les coupes sont acquises en mode entrelacé, il faut effectuer le slice timing avant la correction du mouvement (Anton et al. [2001]). En effet, dans l'ordre inverse, on risque de déplacer le contenu de certains voxels d'une coupe à une coupe adjacente acquise à un instant très différent (pouvant aller jusqu'à TR/2). Lorsque l'acquisition est séquentielle, l'étape de Slice timing doit être effectuée après la correction du mouvement.

Une fois le slice timing effectué, il faut supprimer les 2 dernières images de la série, leur interpolation temporelle n'étant pas fiable du fait de l'absence de scans postérieurs (Anton et al. [2001]).

5.2.3 Correction du mouvement

L'objectif est ici de corriger les artefacts dus aux mouvements de la tête du sujet. Si cette étape est omise, des faux positifs (voxels considérés comme activités alors qu'ils ne le sont pas en réalité) risquent d'apparaître à la périphérie du cerveau.

On choisit un scan de référence[2] au sein de la série temporelle et on corrige le déplacement des autres scans par rapport à ce scan de référence. Le déplacement est considéré comme rigide, c'est à dire composé uniquement de translations et de rotations. Il y a donc 6 paramètres à estimer (une rotation et une translation pour chacun des 3 axes) pour chaque scan. Ces paramètres sont estimés par un algorithme de minimisation de la distance entre le scan initial et le scan transformé.

Notons t_x, θ_x, t_y, θ_y, t_z et θ_z les paramètres de translation et de rotation respectivement autour des axes X, Y et Z. La matrice T de translation selon le vecteur $t = (t_x, t_y, t_z)$ et les matrices R_x, R_y et R_z de rotation respectivement de θ_x, θ_y et θ_z autour des axes X, Y et Z s'écrivent alors :

$$T = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad R_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & \cos\theta_x & \sin\theta_x & 0 \\ 0 & -\sin\theta_x & \cos\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$R_y = \begin{pmatrix} \cos\theta_y & 0 & \sin\theta_y & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{et} \quad R_z = \begin{pmatrix} \cos\theta_z & \sin\theta_z & 0 & 0 \\ -\sin\theta_z & \cos\theta_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

La matrice de transformation correspondant à une translation de vecteur t, et des rotations de θ_x, θ_y et θ_z autour des axes X, Y et Z est le produit $T \times R_x \times R_y \times R_z$.

Pour chaque scan, les paramètres de translation et de rotation minimisant la distance par rapport au scan de référence sont estimés. On obtient ainsi une matrice de transformation par scan. Cette transformation est appliquée au scan correspondant afin de corriger les déplacements de la tête du sujet au cours de l'acquisition. On dit que les scans ont été recalés par rapport au scan de référence.

5.2.4 Normalisation spatiale

Son objectif est de plonger toutes les images (les scans) dans un espace commun (le template), en général le repère du MNI construit à partir de 152 cerveaux du Montreal Neurological Institute. Cette étape est indispensable si on veut pouvoir comparer les résultats obtenus sur des sujets différents.

La normalisation spatiale déforme les images de telle façon que les régions fonctionnelles homologues des différents sujets soient aussi proches que possible. Des problèmes calculatoires peuvent cependant apparaître (minima locaux, pas assez d'information dans les images, calculatoirement cher). Il est donc nécessaire de faire un compromis en corrigeant les grosses différences et en lissant les images normalisées.

[2] En général on choisit le premier de la série (Anton et al. [2001])

La méthode consiste à déterminer la transformation spatiale qui minimise la somme des carrés des différences entre l'image et le template et qui maximise le caractère lisse de la déformation.

La technique de normalisation la plus utilisée est la transformation affine s'appuyant sur l'estimation de 12 paramètres. Aux 6 paramètres déjà vus lors de la correction du mouvement s'ajoutent 3 paramètres de zooms (z_x, z_y, z_z) et 3 de recadrage (r_{xy}, r_{yz}, r_{xz}). La matrice de transformation s'écrit alors :

$$T \times R_x \times R_y \times R_z \times \begin{pmatrix} Z_x & 0 & 0 & 0 \\ 0 & Z_y & 0 & 0 \\ 0 & 0 & Z_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & r_{xy} & r_{xz} & 0 \\ 0 & 1 & r_{yz} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Cette matrice est estimée pour chaque scan. Elle définit la transformation géométrique à appliquer à chaque scan afin que ceux-ci soient définis dans un repère commun. Les images obtenues sont dites normalisées.

5.2.5 Lissage spatial

Avant l'analyse statistique des données, il est conseillé de lisser spatialement les données. En effet, le signal d'intérêt est généralement étendu sur plusieurs voxels. Ceci est dû à la fois à la nature des sources neuronales et à l'étendue de la réponse hémodynamique comme expliqué par Brett et al. [2003]. Des expériences menées en imagerie avec une haute résolution, notamment par Friston [2003], ont montrées que l'échelle spatiale de ces réponses se situe entre 2 et 5 mm. Ainsi, le lissage augmentera le ratio signal/bruit et ceci d'autant mieux que la taille du filtre de lissage se rapprochera de celle de l'étendue spatiale de la réponse hémodynamique. Par ailleurs, le lissage spatial est rendu nécessaire par la phase d'inférence statistique. Cette dernière utilise la théorie GRF (Gaussian Random Fields) pour laquelle il faut que les champs d'erreur soient une version quadrillée d'un champ aléatoire continu sous jacent avec une distribution multi gaussienne (Friston [2003]). Ceci n'est possible que si la taille des voxels est suffisamment petite par rapport à celle du lissage (Worsley et Friston [1995]). Enfin, dans le cas d'analyses multi-sujets, le lissage spatial va permettre de résoudre les problèmes causés par la variabilité neuroanatomique entre les différentes sujets soumis à la même tâche expérimentale. Ainsi, comme expliqué par Friston [2003] et Worsley et al. [1996], plus la taille du filtre sera élevée et plus le regroupement des données de plusieurs sujets sera interprétable.

Pour ces raisons, le choix de la taille du noyau (ou filtre) gaussien utilisé est primordiale. Il faut qu'elle soit suffisamment importante pour que les propriétés précédentes soient vérifiées, mais qu'elle reste raisonnable afin de ne pas dénaturer totalement les données. Worsley et Friston [1995] recommandent d'utiliser un filtre dont la taille est au moins 2 fois celle des voxels avant d'appliquer des résultats issus de la théorie GRF.

Lisser une image par un noyau gaussien consiste à remplacer la valeur de chaque voxel par une moyenne pondérée à partir d'elle même et de la valeur des voxels voisins, ou en d'autres termes à effectuer la convolution de la réponse (la valeur des voxels) avec ce noyau sur chacun des 3 axes successivement. La taille du filtre est appelée FWHM (Full Width at Half Maximum). Un FWHM de 10 voxels signifie que, à 5 voxels du centre (le voxel dont on veut lisser la valeur), la valeur du noyau est la moitié de son pic (au centre).

La convoluée d'une fonction f par un noyau gaussien de taille FWHM (en mm) en un point i de l'axe est :

$$t(i) = \sum_{j=-d}^{d} f(i-j)g(j),$$

$$\text{où} \qquad g(j) = \frac{1}{\sqrt{2\pi s^2}} exp\left(-\frac{j}{2s^2}\right) \qquad \text{et} \qquad s = \frac{FWHM}{\sqrt{8\ln 2}}.$$

$g(j)$ est l'amplitude de la gaussienne à j unités du centre et s est la variance de la gaussienne. La valeur de d est approximativement de 3 FWHMs. Au delà, la valeur du coefficient $g(j)$ peut être considérée comme négligeable.

La valeur de la fonction lissée au point i de l'axe est donc une moyenne pondérée (la somme des coefficients pondérateurs est 1) de la valeur du signal f aux points j distants de i d'au plus d unités, et dont les poids $g(j)$ diminuent avec l'éloignement de i.

5.3 La modélisation

Les méthodes utilisée en IRMf pour ajuster les données à chaque voxel sont jusqu'à présent essentiellement paramétriques. Un modèle linéaire est utilisé. Cependant il n'entre pas dans le cadre du modèle linéaire général pour lequel les erreurs sont supposées être indépendantes et identiquement distribuées ($\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$). La sphéricité des termes d'erreurs ne peut pas être supposée (indépendance violée par les corrélations temporelles) ce qui nous amène à une classe plus large de modèles linéaires, le modèle linéaire généralisé (GLM). L'hypothèse de normalité des erreurs étant toutefois conservée, nous nous limiterons, conformément aux préconisations de Kiebel et Holmes [2003], à la sous-classe de GLMs utilisant la loi normale comme fonction de lien.

Nous allons commencer par écrire le modèle et nous rappellerons le cadre du modèle linéaire général. Puis la génération des régresseurs et le lissage temporel seront détaillés avant de donner la modélisation finale utilisant le modèle GLM.

5.3.1 Un modèle linéaire temporel

Pour chaque voxel, nous disposons d'une série temporelle de N observations acquises aux temps $t_1, \cdots, t_i, \cdots, t_N$. L'intervalle séparant 2 temps d'acquisition successifs est égal à 1 TR. L'objectif est de modéliser pour chaque voxel la série temporelle observée par une

combinaison linéaire de fonctions explicatives plus un terme d'erreur. Les fonctions explicatives sont un ensemble approprié de régresseurs définis de façon à ce que leurs combinaisons linéaires couvrent l'ensemble des réponses IRMf possibles. Nous allons dans un premier temps expliciter la forme du modèle.

Pour un voxel j fixé la modélisation de la réponse sur les K régresseurs s'écrit comme combinaison linéaire des régresseurs plus un terme d'erreur :

$$y_i^j = \sum_{k=1}^{K} \beta_k^j x_{ik} + \varepsilon_i^j,$$

où $i = 1, \cdots, N$ est l'indice du scan considéré. Les N scans successifs constituent les N observations de la série temporelle à modéliser. Toujours pour le voxel j, l'ensemble de ces N équations se regroupent sous la forme matricielle suivante :

$$\begin{pmatrix} y_1^j \\ \vdots \\ y_i^j \\ \vdots \\ y_N^j \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{iK} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nk} & \cdots & x_{NK} \end{pmatrix} \begin{pmatrix} \beta_1^j \\ \vdots \\ \beta_k^j \\ \vdots \\ \beta_K^j \end{pmatrix} + \begin{pmatrix} \varepsilon_1^j \\ \vdots \\ \varepsilon_i^j \\ \vdots \\ \varepsilon_N^j \end{pmatrix}.$$

Ainsi le modèle s'écrit en notation matricielle pour le voxel j :

$$Y^j = X\beta^j + \varepsilon^j$$

où : Y^j est est le vecteur $N \times 1$ des réponses (valeurs du BOLD dans le voxel j pour les N scans), X est la matrice $N \times K$ de design (une ligne par scan, une colonne par régresseur), β^j est le vecteur $K \times 1$ des coefficients du voxel j (un coefficient par condition) et ε^j est le vecteur $N \times 1$ des résidus de la modélisation dans le voxel j pour les N scans.

Dans un souci de clarté, l'indice j désignant le voxel considéré sera dorénavant omis.

5.3.2 Le modèle linéaire général

Les hypothèses gaussiennes et de sphéricité se traduisent par :

$$Y \sim \mathcal{N}\left(X\beta, \sigma^2 Id_N\right) \text{ soit encore } \varepsilon \sim \mathcal{N}\left(0, \sigma^2 Id_N\right).$$

L'estimateur des moindres carrés (qui est ici également l'estimateur de variance minimale parmi les estimateurs sans biais) du couple (β, σ) s'écrit :

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{N-K}\hat{\varepsilon}'\hat{\varepsilon},$$

où $\hat{\varepsilon} = Y - \hat{Y} = RY = \left[Id_N - X(X'X)^{-1}X'\right] Y$. $\hat{Y} = X\hat{\beta}$ est la projection orthogonale de Y sur l'espace de dimension $K = rang(X)$ engendré par les colonnes de X.

$\hat{\beta}$ et $\hat{\sigma}$ ont les propriétés suivantes :

$$\begin{cases} \hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right), \\ \\ \left(N-K\right) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-K}, \\ \\ \hat{\beta} \text{ et } \hat{\sigma} \text{ sont indépendants .} \end{cases}$$

Il est également possible de regrouper les écritures matricielles de la modélisation de tous les voxels en une seule écriture matricielle, $Y = X\beta$ en notant $Y = \left(Y^1|Y^2|\cdots|Y^J\right)$, $\beta = \left(\beta^1|\beta^2|\cdots|\beta^J\right)$ et $\varepsilon = \left(\varepsilon^1|\varepsilon^2|\cdots|\varepsilon^J\right)$, où J est le nombre de voxels par scan. Cette équation est appelée "image regression". En pratique, on travaille voxel par voxel. Le même modèle est appliqué pour chaque voxel.

5.3.3 Le test T avant lissage temporel

Une fois que les paramètres sont estimés, on calcule pour chaque voxel la statistique permettant de tester la nullité d'un contraste de paramètres.

La statistique de test relative au contraste c s'écrit pour un voxel :

$$T = \frac{c'\hat{\beta}}{\sqrt{\hat{\sigma}^2 c' \left(X'X\right)^{-1} c}}.$$

L'hypothèse testée est $H_0 : c'\beta = 0$. Sous H_0, $T \sim St_{N-K}$.

Supposons que le protocole consiste à soumettre les sujets à une condition "activation" et une condition "contrôle". On peut alors se demander quelles sont les régions cérébrales activées plus fortement par la condition "activation" que par la condition "contrôle". Dans ce cas, l'hypothèse nulle testée est $H_0 : \beta(activation) - \beta(controle) = 0$ et l'hypothèse alternative est $H_1 : \beta(activation) - \beta(controle) > 0$. La statistique T est calculée pour chaque voxel, et l'ensemble des valeurs de tous les voxels forment une carte statistique. Les voxels pour lesquels la valeur de T est élevée sont donc des voxels activés plus fortement par la condition "activation" que par la condition "contrôle". Les régions cérébrales significativement activées par la condition "activation" sont obtenues en seuillant la carte statistique.

5.3.4 Génération des régresseurs

Les colonnes de X, appelée "design matrix", contiennent en fait les valeurs des régresseurs (fonctions continues) aux différents temps d'acquisition des scans. Ces régresseurs sont définis à partir des fonctions de stimuli (une par condition expérimentale) et de fonctions de bases choisies afin qu'elles permettent de modéliser la réponse attendue du BOLD. La matrice X est bien évidemment la même pour tous les voxels (tous les voxels ont été acquis suivant le même protocole expérimental).

La construction de la fonction de stimuli S_k ($k = 1, \cdots, K$) se fait en spécifiant les instants de début et la durée de tous les évènements de la condition k. L'ensemble des vecteurs des instants de début et de durée de toutes les conditions est appelé SOA (Stimulus Onset Asynchrony). A partir de ces SOA, le logiciel SPM99 génère une représentation interne de la série d'acquisition et de son design en compartimentant chaque TR en 16 intervalles de temps (16 est le nombre d'intervalles par TR sélectionné par défaut par SPM99). L'occurrence d'un stimulus est ensuite représentée de façon binaire dans les fonctions de stimulus en attribuant la valeur 0 ou 1 à chaque intervalle de temps de longueur $TR/16$. On obtient ainsi des fonctions de stimuli S_k binaires de résolution temporelle $TR/16$ supérieure à celle de la réponse qui est TR. On dit que les fonctions de stimuli sont sur-échantillonnées.

Les fonctions de stimuli étant définies, il faut leur injecter l'information concernant la forme de la réponse BOLD attendue pour obtenir les régresseurs. Ceci se fait en utilisant un ensemble de fonctions de base. SPM99 utilise la HRF (Hemodynamic Response Function) et ses dérivées. Dans le cas d'un design par bloc de conditions (par opposition à un design de stimuli brefs), il est suffisant de n'utiliser que la HRF comme fonction de base. On effectue la convolution de la HRF avec chacune des fonctions de stimuli, et on discrétise (on sous-échantillonne) la convoluée aux temps t_i de mesure de la réponse BOLD (tous les TR). La colonne k de la matrice de design X correspond donc à la convolution discrétisée de la fonction de stimulus S_k avec la HRF soit encore :

$$x_{ik} = f_k(t_i) \text{ où } f_k = HRF \otimes S_k.$$

La convolution avec la HRF est donc une transformation des prédicteurs afin d'optimiser la part de la réponse expliquée. En effet, puisque le signal en réponse à un stimuli instantané à la forme de la HRF, le signal est optimalement restitué si la série temporelle des prédicteurs est lissée par convolution avec la HRF, ce qui permettra d'obtenir une estimation des paramètres optimale. La figure 5.1 illustre le passage des régresseurs initiaux (ou plus exactement des fonctions de stimuli sur-échantillonnées) à ceux utilisés dans le modèle après convolution avec la HRF.

5.3.5 Lissage temporel

5.3.5.1 Le filtrage temporel aux basses fréquences

Les bandes de basse fréquence dans les données contiennent plus de bruit que les autres. Ceci peut être dû à des sources physiques (ex : lents changements de température ambiante), des sources physiologiques (ex : biorythme et cycles cardiaques) et/ou des effets de mouvement résiduel et de leur interaction avec le champ magnétique statique (Henson [2003], Kiebel et Holmes [2003]). Ainsi, en filtrant les données avec un filtre passe-haut, on peut supprimer une grande partie du bruit et ainsi accroître le ratio signal/bruit. Il faut prendre garde à ne pas choisir une fréquence de coupure trop élevée afin de minimiser la perte du signal d'intérêt ni trop faible afin de ne pas laisser passer trop de bruit. En pratique, on choisit une période de coupure égale à deux fois l'intervalle de temps maximum séparant 2

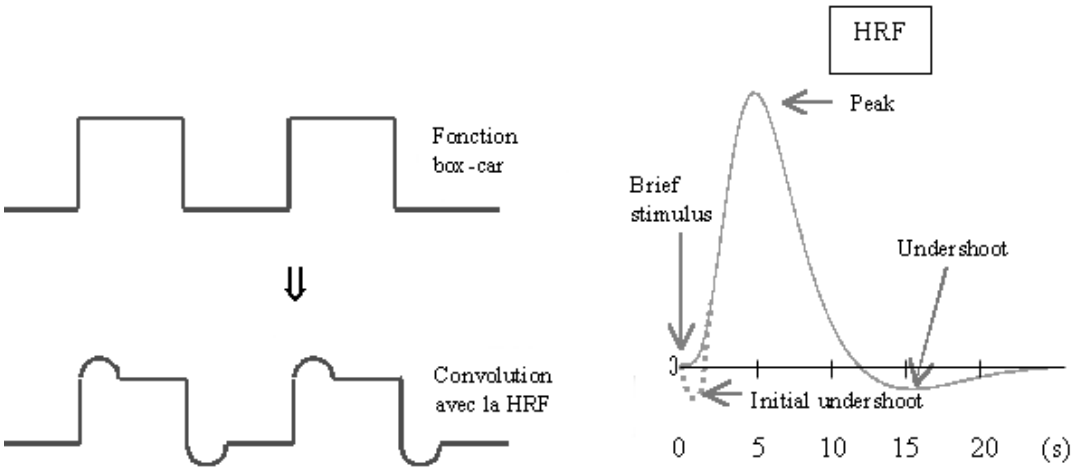


FIG. 5.1 – Convolution des prédicteurs (box-car) avec la fonction de réponse hémodynamique (HRF). On obtient la forme des prédicteurs utilisés dans le modèle.

occurrences de la condition expérimentale la plus fréquente si elle n'excède pas 120 s. Sinon on choisit une fréquence de coupure de 1/120 Hz.

Le filtre passe-haut utilise un ensemble de fonction de base $f_r(t)$ qui sont des fonctions cosinus discrètes appelée DCT (Discrete Cosine Transform). Le nombre de fonction R dépend du choix de la fréquence de coupure selon l'expression $R = 2N.TR/f_c + 1$ où f_c est la fréquence de coupure exprimée en secondes. Mathématiquement :

$$f_r(t_i) = \sqrt{\frac{2}{N}} \cos\left(r\pi \frac{t_i}{N}\right) \text{ pour } i = 1, \cdots, N \text{ et } r = 1, \cdots, R.$$

Ces fonctions DCT sont intégrées dans la matrice de design X comme facteurs de confusion. Puis la matrice des résidus formées à partir de ces facteurs de confusion est appliquée aux données. Une autre façon de voir les choses est de considérer qu'elle font partie d'une matrice de lissage temporel S qui est appliquée à la fois aux données et au modèle.

5.3.5.2 Les corrélations temporelles hautes fréquences

Les données IRMf sont temporellement corrélées à petite échelle (indépendamment des corrélations induites par le protocole expérimental) ce qui signifie que les termes d'erreur sont corrélés avec leurs voisins temporels. Ces autocorrélations temporelles doivent être prisent en compte afin de pas biaiser les estimations des tests statistiques. Il faut cependant garder à l'esprit que si on veut appliquer le modèle GLM, il faut que les erreurs aient une distribution multi-gaussienne. Une solution permettant de résoudre ce problème est de lisser temporellement les données comme expliqué par Worsley et Friston [1995]. C'est la méthode utilisée par SPM2. Ce lissage temporel peut également être considéré comme un

filtrage temporel aux hautes fréquences. Cette méthode revient donc à ajouter à la matrice S une composante de filtre passe-bas. Cette matrice est appliquée à la fois aux données et au modèle de façon à ce que, si le noyau de lissage est suffisamment large (en comparaison à l'étendue des corrélations des réponse), l'autocorrélation induite par le lissage recouvrira l'autocorrélation intrinsèque (des données non lissées) de telle façon que :

$$V = S\Sigma S' \sim SS',$$

où V est la matrice d'autocorrélation des erreurs après lissage et Σ est la matrice d'autocorrélation intrinsèque.

5.3.6 Le modèle linéaire généralisé

On applique donc aux réponses, aux régresseurs et aux résidus la matrice S de lissage temporel pour obtenir le modèle :

$$SY = SX\beta + S\varepsilon \quad \text{soit encore} \quad SY = \widetilde{X}\beta + S\varepsilon \quad \text{avec} \quad \widetilde{X} = SX.$$

Les hypothèses gaussiennes sont désormais $S\varepsilon \sim \mathcal{N}\left(0, \sigma^2 SS'\right)$ et les estimateurs de β et σ^2 s'écrivent

$$\hat{\beta} = \left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}'SY \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{trace(RV)}\hat{\varepsilon}'\hat{\varepsilon},$$

où $V = SS'$, $R = Id_N - \widetilde{X}\left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}'$ et $\hat{\varepsilon} = RY$.

Sous les hypothèses gaussiennes, $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma\left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}'V\widetilde{X}\left(\widetilde{X}'\widetilde{X}\right)^{-1}\right)$.

5.3.7 Le test T après lissage temporel

La statistique de test relative au contraste c s'écrit maintenant :

$$T = \frac{c'\hat{\beta}}{\sqrt{\hat{\sigma}^2 c'\left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}'V\widetilde{X}\left(\widetilde{X}'\widetilde{X}\right)^{-1}c}}. \tag{5.0}$$

L'hypothèse testée est $H_0 : c'\beta = 0$. Sous H_0, $T \sim St_\nu$ où

$$\nu = trace(RV)^2/trace(RVRV)$$

est le nombre de degrés de liberté corrigés. SPM99 applique ainsi une approximation de Satterthwaite décrite par Glaser et Friston [2003] qui utilise implicitement une mesure de violation de la sphéricité.

La statistique de test a donc, sous l'hypothèse nulle, une distribution de Student avec un degré de liberté corrigé ν[3]. La valeur de cette statistique résume, pour un voxel donné, l'information contenue dans la série temporelle concernant le contraste de paramètres considéré. L'objectif de la partie suivante est de déterminer à partir de quel seuil la valeur de la statistique signale la présence d'une activation dans le voxel considéré.

[3]Ce degré de liberté ne sera généralement pas un entier.

5.4 Inférence statistique

Une correction pour comparaisons multiples dépendantes est nécessaire. La théorie des champs aléatoire gaussiens (GRF) permet de corriger le seuil de significativité en prenant en compte le fait que les voxels voisins ne sont pas indépendants en vertu de la continuité des données initiales. La correction GRF est moins sévère que la correction de Bonferroni pour le nombre de voxels. La théorie GRF traite du problème de la comparaison multiple dans le contexte de champs statistiques continus spatialement étendus, d'une façon analogue à la correction de Bonferroni pour des familles de tests statistiques discrets. Appliquer la correction de Bonferroni reviendrait à contrôler le nombre attendu de voxels faux positifs, alors que la théorie GRF contrôle le nombre attendu de régions activées à tort, régions pouvant englober des centaines de voxels faux positifs.

D'après Poline et al. [1997a], les hypothèses nécessaires à l'application de la correction GRF sont :

1. Les champs d'erreur sont une version quadrillée d'un champ aléatoire continu sous jacent avec une distribution multi gaussienne.
2. Les composants de ce champs ont une distribution multi gaussienne et une stationnarité au sens large (la distribution de probabilité multivariée d'un voisinage de points est indépendante de la position de ce voisinage dans le champs, ce qui est vérifié si la taille des voxels est au moins la moitié de la taille du lissage FWHM).
3. Les seuils utilisés sont élevés (> 1.64).

5.4.1 Les différents niveaux d'inférence

Cette correction passe par le calcul de la probabilité qu'un nombre c de clusters d'au moins k voxels soit activé à un seuil u. Cette probabilité est notée $P_W(u, k, c)$ où W est la taille du lissage. Les différents paramètres sont choisis pour que $P_W(u, k, c) = \alpha$ sous l'hypothèse nulle. Plusieurs niveaux d'inférence peuvent en fait être envisagés, ce qui donne lieu à des tests différents. Le tableau 5.1 récapitule les hypothèses nulles testées ainsi que les valeurs des paramètres pour chacun des 3 niveaux d'inférence. Il apparaît clairement dans ce tableau que les 3 niveaux d'inférences sont emboîtés. Le niveau voxel est un cas particulier du niveau cluster avec $k = 0$ et le niveau cluster est lui même un cas particulier du niveau global avec $c = 1$.

L'intérêt du niveau global d'inférence réside dans le fait que le nombre de cluster observé présentant un profil d'activation est fortement peu probablement dû au hasard. Le niveau global d'inférence est ainsi plus puissant que le niveau cluster, lui même plus puissant que le niveau voxel.

A des seuils élevés, le nombre de clusters C dans un volume V (de dimension D) est une approximation du nombre de maxima et il a été montré par Adler [1981] qu'il a une

Niveau d'inférence	Hypothèse nulle testée	Seuil u	Nombre de clusters c	Taille des clusters k
Voxel	Aucun voxel n'est au dessus du seuil u	u	1	0
Cluster	Aucun cluster de taille $\geqslant$ k (en nombre de voxels) n'est au dessus du seuil u	u	1	k
Global	Il y a moins de c clusters de taille $\geqslant k$ au dessus du seuil u	u	c	k

TAB. 5.1 – Niveaux d'inférence

distribution de Poisson : $P(C = c) = \frac{1}{c!}E(C)^c e^{-E(C)}$ où $E(C)$ est l'espérance du nombre de maxima (c'est à dire de clusters), $E(C) = V(2\pi)^{-\frac{D+1}{2}} W^{-D} u^{D-1} e^{-\frac{u^2}{2}}$. Le nombre de voxels K compris dans un cluster est distribué suivant $P(K \geqslant k) = exp\left(-\beta k^{2/D}\right)$ où

$$\beta = \left[\frac{\Gamma\left(\frac{D}{2}+1\right) E(C)}{V\Phi(-u)}\right]^{2/D}.$$

Au niveau global, Friston et al. [1996] ont montré que la probabilité d'avoir au moins c clusters de taille au moins k dépassant le seuil u est

$$P_W(u,k,c) = 1 - \sum_{i=0}^{c-1}\sum_{j=1}^{\infty} P(C=j)\binom{j}{i} P(K \geqslant k)^i P(K<k)^{j-i}$$

$$= 1 - \sum_{i=0}^{c-1} \frac{1}{i!}\left(E(C)P(K \geqslant k)\right)^i exp - \left(E(C)P(K \geqslant k)\right).$$

Au niveau cluster, la probabilité qu'au moins un cluster de taille au moins k dépasse le seuil u est $P_W(u,k,1) = 1 - exp\left(-E(C)P(K \geqslant k)\right)$.

Au niveau voxel, la probabilité qu'au moins un voxel dépasse le seuil u est $P_W(u,0,1) = 1 - exp\left(-E(C)\right)$ car $P(K > 0) = 1$.

5.4.2 Inférence combinant intensité et étendue spatiale

Une méthode a été proposée par Poline et al. [1997b] permettant de tenir compte à la fois de l'intensité de l'activation et de l'étendue spatiale de l'activation. L'idée de cette méthode est la suivante.

1. On cherche une approximation de la probabilité qu'un cluster donné soit de taille $S > s_0$ et ait une intensité $H > h_0$

2. L'équation $P(S \geqslant s_0, H \geqslant h_0) = \alpha$ a un nombre infini de couples (s_0, h_0) solutions. On parle de courbe isocumulative. Le risque d'erreur est simplement défini comme le minimum entre le risque pour l'étendue et le risque pour le pic maximum : $min\left(P(S \geqslant s_0), P(H \geqslant h_0)\right) = \alpha$.

3. Pour un seuil t élevé, les clusters sont indépendants et le nombre C de clusters au dessus de t suit approximativement une loi de Poisson de moyenne $E(C)$ comme nous l'avons vu dans le paragraphe précédent. Si on note P_{rej} la probabilité que la probabilité combinée de l'étendue et du pic d'un cluster tombe dans la région de rejet, et si c clusters apparaissent dans le volume V, la probabilité qu'au moins un cluster soit rejeté est simplement :

$$P(\text{Au moins un cluster rejeté}/C = c) = 1 - (1 - P_{rej})^c$$

et en faisant la somme sur c pondérée par $P(C = c) = E(C)^c exp(-E(C))/c!$, on obtient :

$$P(\text{Rejet}) = \sum_{c=0}^{\infty} \left(1 - (1 - P_{rej})^c\right) \frac{E(C)^c exp(-E(C))}{c!} = 1 - exp(-E(C)P_{rej})$$

comme p-value finale pour une recherche parmi tous les clusters.

5.5 Analyses multi-sujets

La plupart des analyses multi-sujets menées jusqu'à présent utilisent des analyses à effets fixes qui ne prennent en compte que la variabilité intra-sujet (ou scan-to-scan). Il n'est pas possible alors de faire de l'inférence sur les effets de population. Les analyses à effets aléatoires permettent de prendre en compte une deuxième source de variabilité qui est la variabilité inter-sujets. La variable sujet est alors considérée comme aléatoire, ce qui est nécessaire puisque la réponse d'un sujet varie d'une session à l'autre et, à fortiori, d'un sujet à l'autre. La prise en compte de ces 2 sources de variations rend possible les inférences sur la population à partir de laquelle les sujets de l'étude ont été tirés (Penny et Holmes [2003]).

5.5.1 Analyses à effets fixes

L'analyse se fait toujours en utilisant le modèle linéaire généralisé. Seulement, au lieu d'avoir des données provenant d'un seul sujet, nous avons désormais des données provenant de plusieurs sujets. Ces données sont concaténées en un seul vecteur colonne Y. Aux colonnes désignant les différentes conditions expérimentales pour chaque sujet viennent s'ajouter des colonnes différenciant les sujets entre eux. Ces variables ne sont pas d'intérêt mais leur prise en compte dans le modèle permet d'améliorer l'ajustement du modèle aux données.

5.5.2 Analyses à effets aléatoires

Lorsqu'on a la même matrice de design pour chaque sujet (balanced design), Penny et Holmes [2003] ont montré qu'il est équivalent de travailler avec des estimateurs du maximum de vraisemblance et des "Summary-Statistics".

Les analyses à effets aléatoires s'implémentent alors comme suit :

1. Ajuster le modèle pour chaque sujet en utilisant des modèles GLMs différents pour chaque sujet : 1^{er} niveau de l'analyse.
2. Définir l'effet d'intérêt pour chaque sujet avec un vecteur de contraste, afin d'obtenir une image de contraste contenant la valeur du contraste estimé pour chaque voxel.
3. Alimenter le modèle GLM avec les images de contraste et effectuer un test T pour un échantillon : 2^{eme} niveau de l'analyse.

Un test T de comparaison de 2 échantillons est utilisé si on veut comparer 2 groupes de sujets.

5.6 Conclusion

La méthode décrite ici est une référence qui est utilisée dans la majorité des analyses de données IRMf. Comme toutes les méthodes, elle a ses limites. Citons en quelques unes. Tout d'abord l'utilisation d'une modélisation paramétrique implique de nombreuses modifications des données, notamment lors des étapes de lissage spatial et temporel. Ensuite la prise en compte des corrélations temporelle est contournée en les forçant à avoir la structure voulue par le biais d'un lissage temporel. Enfin, face au grand nombre de voxels, une analyse univariée de masse est utilisée, ce qui est en soi discutable.

L'objectif de cette thèse n'est pas de discuter du bien fondé des méthodes utilisées en IRMf. La méthode standard a été décrite ici afin de définir les données IRMf et de bien comprendre ce que représentent les pics d'activation qui sont utilisés dans le chapitre suivant.

Chapitre 6

Une approche originale pour la détection de clusters de pics d'activation

L'application de méthodes de détection de clusters spatiaux à l'imagerie médicale a déjà été abordée par d'autres auteurs. La statistique de scan spatial a notamment été utilisée par Yoshida et al. [200l] et Neill et al. [2005]. Les premiers se servent du test de Kulldorff pour définir un "hotspot" comme étant une région ayant un pourcentage significatif de neurones activés. Les seconds ont illustré sur des données IRMf la rapidité d'exécution de leur algorithme permettant, selon eux, de trouver un cluster spatial plus de 1400 fois plus vite que la statistique de scan spatial usuelle. La nature des données qu'ils utilisent n'est pas très claire mais leur approche est intéressante. Les données consistent en deux cartes d'activation du cerveau d'un sujet, l'une obtenue sous une condition de contrôle et l'autre sous une condition expérimentale. Ils appliquent alors la statistique de scan en définissant un nombre de cas et une population pour chaque voxel où une activité cérébrale apparaît : le nombre de cas d'un voxel est définit comme l'activation IRMf de ce voxel sous la condition expérimentale et la population de fond par l'activité sous la condition de contrôle. Pour les voxels sans activité cérébrale, ils définissent un nombre de cas et une population égaux à 0.

Dans cette partie, nous présentons une application sur des données IRMf de notre méthode de détection de clusters spatiaux vue au chapitre 4. Notre approche est différente de celle de Neill et al. [2005]. Elle utilise les résultats de la méthode SPM. Son but est de localiser des clusters correspondant aux régions cérébrales simultanément activées par la plupart des sujets.

La procédure consiste tout d'abord en la détermination de pics d'activation pour chaque sujet par la méthode standard (SPM) décrite dans le chapitre précédent et brièvement rappelée ici. Les pics de tous les sujets sont ensuite regroupés pour former un jeu de données $3D$. Finalement, la méthode de détection de clusters est appliquée à ce jeu de données ponctuelles dans le but de localiser et détecter les éventuels clusters de pics d'activation.

Cette approche sera illustrée par une application à des données IRMf obtenues lors d'un protocole visant à déterminer les zones impliquées par une tache de fluidité verbale. Les résultats de cette application sont présentés dans Dematteï et al. [2006b]. Nous comparerons nos résultats avec ceux obtenus par la statistique de scan spatial.

6.1 Résumé de la méthode SPM

Pour un sujet, les données initiales recueillies par l'appareil d'acquisition IRMf sont constituées de plusieurs centaines d'images $3D$ du cerveau, obtenues successivement sous différentes conditions expérimentales (pour fixer les choses, prenons un exemple avec deux conditions expérimentales A et B). Ces images, appelées scans, sont découpées en voxels et le niveau neuronal d'activité est mesuré à l'intérieur de chaque voxel.

Après toute une série de pré-traitements, la modélisation SPM permet d'obtenir, pour un contraste c de conditions expérimentales (par exemple $c = B - A$), une carte statistique d'activation. Cette carte, appelée SPM, est également en 3D et contient la valeur d'une statistique de test pour chaque voxel. Elle résume l'information contenue dans tous les scans acquis sous les conditions A et B. Les voxels ayant une valeur élevée pour la statistique de test sont des voxels présentant une activité cérébrale plus forte sous la condition B que sous la condition A.

L'étape d'inférence consiste à choisir un seuil et à seuiller la carte statistique. Un voxel qui a une valeur de la statistique supérieure au seuil dénote une activité cérébrale dans ce voxel significativement plus forte sous B que sous A. Un ensemble de voxels adjacents activés constitue une région cérébrale d'activation.

6.2 Principe de la détection de clusters de pics d'activation

Pour chaque région cérébrale activée, on définit un pic d'activation. Celui-ci est le voxel qui, à l'intérieur d'une région, contient la valeur maximale de la statistique de test. Ce pic a l'avantage de résumer l'emplacement d'une région activée par une seule coordonnée. Chaque sujet présente généralement plusieurs régions activées et donc plusieurs pics d'activation. Ceci nous permet d'obtenir pour chaque individu un ensemble de données ponctuelles en $3D$.

Dans l'hypothèse où la condition expérimentale induit des activations cérébrales dans les mêmes zones chez les différents individus, on devrait pouvoir retrouver, en superposant les pics des différents sujets, des clusters correspondant à des zones cérébrales simultanément activées chez la plupart des sujets. Nous proposons de déterminer ces zones d'activation grâce à la méthode de régression présentée dans le chapitre 4.

L'approche que nous proposons ici peut être résumée comme suit :
- Détermination des pics d'activation pour chaque sujet par une analyse SPM individuelle,
- Regroupement des pics d'activation de tous les individus : leur localisation définit les données utilisées pour l'analyse de détection de cluster,
- Détermination du ou des clusters de pics d'activation par la méthode de régression sur données transformées.

6.3 Application à un protocole sur la fluidité verbale

6.3.1 Paradigme expérimental

Une tache de fluidité verbale était effectuée par 11 femmes droitières lors d'un protocole IRMf par blocs classique. Cinq conditions de contrôle et cinq conditions d'activité étaient alternées. Chacune de ces conditions était constituée de 10 volumes et durait 50 secondes. Avant le scan, chaque sujet était informé de la tache à effectuer.

Ces deux types de conditions étaient une tache de fluidité verbale (condition d'activation) et une tache de comptage (condition de contrôle). Pendant les cinq conditions d'activation, les sujets devaient produite silencieusement autant de mots que possible commençant par une lettre présentée oralement : "F", "A", "S", "T" et "N" respectivement, comme dans Schlosser et al. [1998]. Pendant les cinq conditions de contrôle, il était demandé aux sujets de compter à partir de 1 à un rythme d'environ 1 par seconde.

6.3.2 Acquisition IRM

Les examens fonctionnels ont été réalisés sur une Vision Siemens de 1.5 Tesla. L'immobilisation de la tête des sujets a été établie par des coussinets. Afin de diminuer le bruit de l'acquisition, des boules Quiès ont été fournies à chaque sujets.

Les acquisitions étaient composées de 15 à 17 coupes transversales contiguës de 8 mm d'épaisseur. Les images fonctionnelles ont été acquises avec un $TR = 64$ ms, une matrice de taille 128×128 et un champs de vue de 256 mm. La taille des voxels étaient ainsi de $2 \times 2 \times 8$ mm. Pendant chaque scan, une série de 100 acquisitions séquentielles était obtenue. La totalité des images du cerveau était collectée toutes les 5 secondes. La tache de fluidité verbale était divisée en 10 époques de 50 secondes chacune, pour une durée totale de la tache de 8 minutes et 20 secondes.

6.3.3 Prétraitements et analyse

La méthode SPM a été appliquée à chaque sujet pour localiser les pics d'activation à un niveau individuel. Les données ont été analysées avec le logiciel SPM99 (SPM99 ; Wellcome Departement of Cognitive Neurology, London, UK).

Sujet	Nombre total de pics	Cluster 1		Cluster 2		Cluster 3	
		n	(%)	n	(%)	n	(%)
1	28	4	(11.1)	5	(12.8)	2	(5.4)
2	24	3	(8.3)	3	(7.7)	4	(10.8)
3	28	5	(13.9)	4	(10.3)	3	(8.1)
4	31	3	(8.3)	5	(12.8)	5	(13.5)
5	34	3	(8.3)	1	(2.6)	3	(8.1)
6	26	3	(8.3)	4	(10.3)	4	(10.8)
7	47	6	(16.7)	4	(10.3)	3	(8.1)
8	25	0	(0)	1	(2.6)	0	(0)
9	44	2	(5.6)	4	(10.3)	5	(13.5)
10	39	4	(11.1)	3	(7.7)	5	(13.5)
11	28	3	(8.3)	5	(12.8)	3	(8.1)
Total	354	36	(10.2)	39	(11)	37	(10.5)

TAB. 6.1 – Répartition des pics d'activations par sujet et par cluster. Le nombre de pics appartenant à chacun des 3 clusters est donné pour chaque sujet. Les pourcentages sont calculés par ligne (par rapport au nombre total de pics par sujet).

Toutes les images ont été réalignées au premier volume, normalisées par rapport au repère du cerveau standard défini par le Montreal Neurological Institute (MNI). Les images fonctionnelles ont ensuite été lissées en utilisant un filtre gaussien avec une FWHM de 8 mm. Finalement, le contraste "activité-contrôle" a été défini.

Un modèle à effets fixes a été appliqué au contraste "activité-contrôle" pour chaque sujet. Les comparaisons multiples ont été prises en compte par l'utilisation d'un seuil non corrigé de $p < 0.0001$. Cette pré-analyse à un niveau individuel a permis de localiser les pics d'activation de chaque sujet.

Ces pics d'activation ont ensuite été regroupés afin de former un jeu de données $3D$. Ce dernier a été analysé avec la méthode de détection de clusters dans le but de déterminer, à un niveau de groupe, quelles zones cérébrales sont activée chez la plupart des sujets.

6.4 Résultats

Pour chaque femme, entre 24 et 47 pics d'activation (maxima locaux) ont été détectés par la méthode standard. La répartition du nombre de pics par sujet est présentée dans la Table 6.1. Les sujets présentent une moyenne de 32 pics, pour un total de 354 pics, tous sujets confondus.

La détection de clusters a été effectuée sur ces 354 pics. Le modèle avec huit points de cassure (quatre clusters potentiels) a été sélectionné. La valeur de la statistique du $WD\ max$

était 25.2, plus élevée que la valeur critique. L'un de ces quatre clusters potentiels n'était pas significatif, tandis que les trois autres était des clusters significatifs.

Les trois clusters de pics d'activation qui ont été détectés sont représentés sur la Figure 6.1. L'un est localisé dans le lobe occipital (cluster 1 dont les pics sont visualisés par des sphères) et les deux autres sont localisés dans le lobe frontal (clusters 2 - sphères - et 3 - cubes). Chacun de ces clusters contient entre 36 et 39 pics d'activation. Comme nous pouvons le constater dans la Table 6.1, la quasi-totalité des sujets présentent entre deux et cinq pics dans chaque cluster. Seul un sujet (le 8) est atypique car il ne présente en tout qu'un seul pic d'activation (un dans le cluster 2 et aucun dans les deux autres). Nous pouvons donc en conclure que ces trois clusters correspondent à des régions cérébrales simultanément activées chez la plupart des sujets. Ces régions sont donc impliquées lors de la réalisation de la tache de fluidité verbale.

Par ailleurs, la statistique de scan spatial de Kulldorff [1997] a été appliquée sur ce jeu de données en $3D$. La taille maximale du cluster spatial est fixée par défaut dans le logiciel Satscan à 50% de la population à risque (ici les voxels). Avec cette valeur, le cluster le plus probable regroupait 261 des 354 pics, soit plus de la moitié de la totalité des pics. Nous avons donc finalement fixé cette valeur à 30%. Le cluster le plus probable est une sphère de centre $(9, -5, -53)$ et de rayon 54.65. Ce cluster est significatif et regroupe 151 pics. Il est représenté sur la Figure 6.1 par une sphère blanche transparente. Nous remarquons qu'ici, la statistique de scan spatial ne fonctionne pas : cette approche détecte un cluster très grand qui n'est pas interprétable.

6.5 Discussion

Sur l'exemple présenté ici, tout s'est bien passé à plusieurs titres. Tout d'abord, les différents individus ont globalement répondus avec la même intensité à l'expérience puisque tous ont un nombre de pics d'activation du même ordre de grandeur. Ensuite, et notamment grâce à cette réponse homogène, les clusters détectés se sont avérés représentatifs de l'ensemble des individus de la population étudiée ce qui nous a permis de conclure.

Lors de l'application de cette approche, on aurait cependant pu être confrontés à des situations plus embarrassantes. Le ou les clusters détectés aurait par exemple pu être composé de pics ne provenant que d'une partie de la population (disons la moitié). Dans ce cas, il serait difficile de conclure de façon nette que la zone cérébrale correspondant au(x) cluster(s) est activée par la tache expérimentale effectuée par les sujets. Un autre cas de figure gênant consisterait à ce que certains sujets se démarquent des autres par une absence de réponse cérébrale ou au contraire par une sur-activation induite par la tache effectuée. On se retrouverait alors avec des individus sous ou sur-représentés au niveau du nombre de pics d'activation.

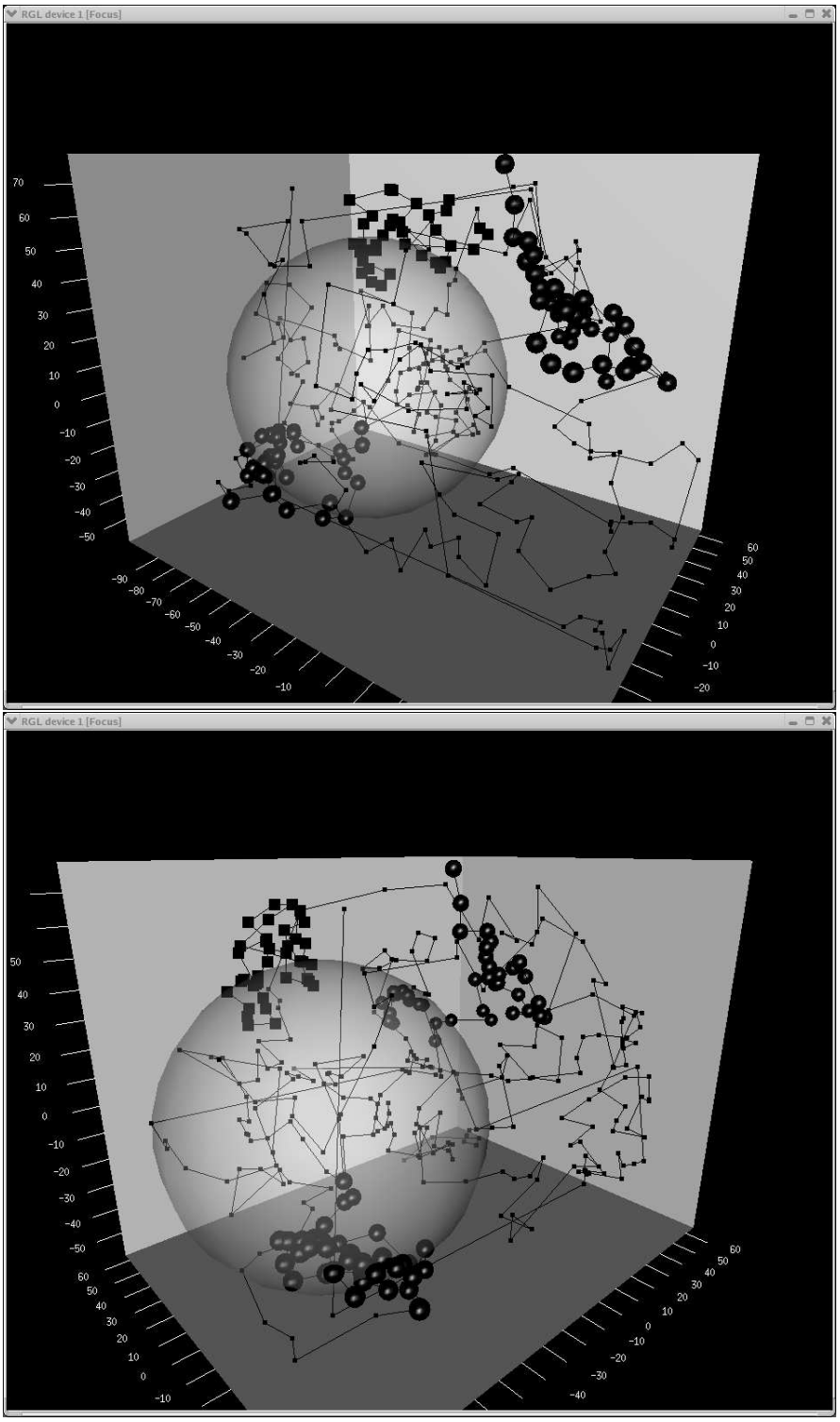


FIG. 6.1 – Représentation en 3D des pics d'activation IRMf. En haut : vue droite du cerveau par l'avant. En bas : vue droite du cerveau par l'arrière. Chaque pic est représenté par un petit cube noir. Un segment relie deux pics lorsqu'ils sont successifs sur la trajectoire. Les points inclus dans un cluster significatif sont représentés par une sphère (clusters 1 et 2) ou un gros cube noir (cluster 3). Le cluster le plus probable détecté par la statistique de scan spatial est représenté par la sphère blanche transparente.

Les problèmes qui seraient posés par ces différents cas de figure proviennent en fait de ce que l'approche proposée ici est l'analogue d'une analyse SPM à effet fixe. Comme nous l'avons vu précédemment, ce type d'analyse regroupe les données d'acquisition de tous les sujets et en fait une carte d'activation SPM globale qui ne prend pas en compte l'effet sujet. Ici, nous avons au préalable déterminé la carte d'activation de chaque sujet, puis regroupé leurs pics d'activation. C'est à ce niveau que notre approche ne prend pas en compte l'aléa de la variable individu. Afin de mesurer cet effet sujet, il serait intéressant de comparer les résultats obtenus par notre approche à ceux d'une analyse SPM à effet aléatoires.

Concernant la visualisation des résultats, nous n'avons pour l'instant pas réussi à récupérer les coordonnées des limites du cerveaux qui sont déterminées par le logiciel SPM. Nous ne sommes donc pas en mesure d'offrir une représentation graphique classique des résultats montrant les contours du cerveau ce qui permettrait de mieux situer anatomiquement les clusters de pics d'activation détectés.

Précisons pour finir que les données utilisées (les pics d'activation) ne sont définies que par la localisation des pics. L'intensité des activations n'est pas prise en compte. On attribue donc implicitement le même poids à tous les pics, quelque soit leur intensité, ce qui est discutable.

Quatrième partie

Perspectives spatio-temporelles

Chapitre 7

Détection de clusters spatio-temporels

Les données obtenues par IRMf que nous avons abordées dans la partie III font intervenir à la fois la dimension temporelle, correspondant à la durée d'acquisitions des images, et les dimensions spatiales, les images acquises étant en 3 dimensions. A propos des analyses de ce type de données, on ne peut cependant pas vraiment parler d'analyses de cluster spatio-temporelles. Les deux types de dimensions sont en effet traitées séparément. Le temps est pris en compte dans une modélisation de la réponse à l'intérieur de chaque voxel. La dimension spatiale n'intervient qu'après cette modélisation temporelle lors du seuillage et de la visualisation en $3D$ des résultats obtenus pour chaque voxel.

La détection de clusters spatio-temporels est une discipline récente qui mérite toute l'attention qui a successivement été portée à la recherche d'agrégats dans le temps, puis à la détection de clusters dans l'espace. Kulldorff a rapidemenent su adapter sa statistique de scan spatial au cas spatio-temporel. Comme en spatial, de nombreuses méthodes, notamment dérivées de la statique de scan spatio-temporelle, voient le jour.

7.1 Approches existantes

Plusieurs procédures ont été développées dans le but d'étudier l'agrégation spatio-temporelle des données de santé géographiques. Le test d'interaction spatio-temporelle de Knox [1964] constitue une approche précoce de ce type de test. Ce test est basé sur l'idée qu'une combinaison d'une proximité spatiale et d'une proximité temporelle des cas de maladie représente un cluster spatio-temporel. On devrait plutôt parler d'interaction spatio-temporelle dans ce cas puisque des agrégats spatio-temporels peuvent exister alors même qu'il n'y a pas d'interaction temporelle, ce qui a été montré par Kulldorff [1999b]. Par ailleurs ce test d'interaction est une méthode globale dans le sens ou les éventuels agrégats ne peuvent être localisés.

La première véritable approche permettant de localiser et détecter un agrégat spatio-temporel est certainement la statistique de scan spatio-temporelle de Kulldorff [1998] qui est une adaptation directe de la statistique de scan spatiale. Au lieu d'utiliser des fenêtres circulaires, la méthode a été utilisée avec des fenêtres cylindriques, ou la base circulaire se déplace dans l'espace et la hauteur du cylindre représente l'intervalle temporel. Ce premier test est aujourd'hui appelé rétrospectif, par opposition à la statistique de scan prospective (Kulldorff [2001]) employée sur des bases de données qui évoluent avec l'apparition de nouveaux cas. Cette dernière permet la détection de clusters émergeants, c'est à dire de clusters qui se finissent au temps présent. Elle donne également un ajustement de la statistique prospective pour tenir compte des analyses répétées sur différentes périodes temporelles. Kulldorff et al. [2005] ont récemment proposé la statistique de scan de permutation spatio-temporelle qui ne nécessite pas la connaissance de la population à risque. Seuls les cas sont donc nécessaires et ce test est une approche prospective.

D'autres approches ont également été proposées. Celle de Iyengar [2004] se propose, à partir de la statistique de scan, de chercher des clusters de forme pyramidale plutôt que cylindrique. Pour chaque temps inclus dans l'intervalle du cluster potentiel, la surface géographique considérée est un carré dont la taille peut varier en fonction du temps. Assunção et al. [2003] ont proposé une statistique de test similaire à la statistique de scan de permutation spatio-temporelle. L'hypothèse nulle considérée est que le processus de point est un processus de Poisson non-homogène ayant une intensité séparable dans le temps et dans l'espace. Par ailleurs, la localisation exacte de chaque point dans le temps et l'espace est supposée (données individuelles plutôt que données groupées). Enfin, Sebastian et al. [2006] utilisent, pour étudier la fusion de cellules, la fonction K de Ripley (Ripley [1977]). Ils définissent successivement les fonctions temporelle, spatiale et spatio-temporelle. L'indépendance de la localisation spatiale et de l'occurence temporelle des évènements se traduit par une séparabilité de la fonction K.

Toutes ces méthodes ont leurs avantages et leurs inconvénients. On peut par exemple reprocher à la statistique de scan spatio-temporelle cylindrique de ne pas être flexible du point de vue de l'évolution de la surface géographique de l'agrégation au cours du temps. D'autres plus flexibles sont couteuses en temps de calcul : les analyses effectuées dans Iyengar [2004] ont pris 34 heures contre 2.5 heures pour la détection cylindrique. La détection de clusters spatio-temporels n'en est, rappelons le, qu'à ses débuts et de nombreuses améliorations sont à apporter comme ce fut le cas en deux dimensions.

7.2 Mise en évidence d'une mini-épidémie dermatologique en Inde

Afin d'illustrer la détection de clusters spatio-temporels, nous avons appliqué la version cylindrique de la statistique de scan spatio-temporelle sur des données réelles. Comme nous

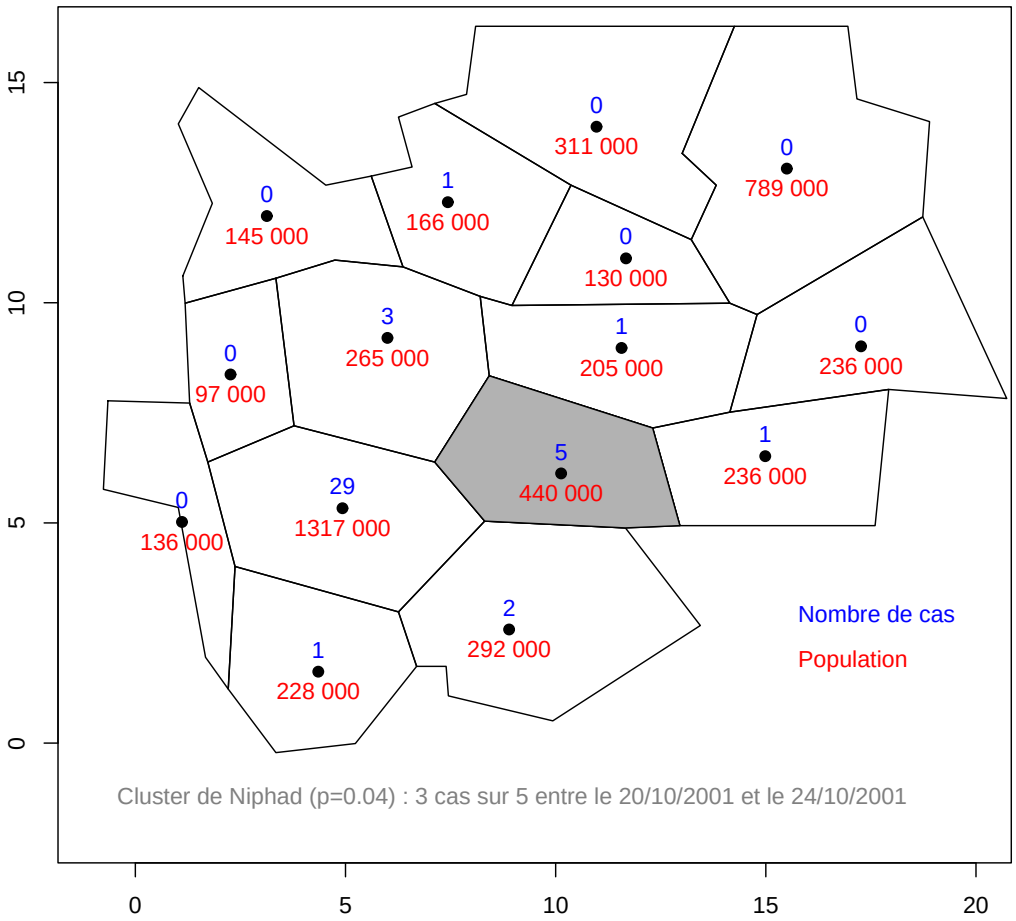


FIG. 7.1 – Représentation du district de Nashik avec ses talukas. Pour chaque taluka, le nombre de cas et la population totale sont affichés.

allons le voir, cette étude a permis de mettre en évidence une mini-épidémie dermatologique en Inde.

Entre le 1^{er} mai 2000 et le 31 octobre 2002 (30 mois), 43 cas d'enfants de moins de 12 ans atteints du syndrome de Gianotti-Crosti (maladie dermatologique dorénavant notée SGC) ont été diagnostiqués dans le district de Nashik en Inde à partir du registre d'une clinique dermatologique située à Nashik. Pour chaque enfant, l'adresse et la date du premier diagnostic ont été relevés. La géolocalisation des adresses n'ayant pas pu être effectuée, les données ont été agrégées par talukas (divisions administratives du district). Le district de Nashik comporte 15 talukas. Le nombre de cas et la population de chaque taluka sont présentés sur la figure 7.1.

Etant donné que nous ne disposions pas de l'exhaustivité des cas de SGC dans le district de Nashik, il a fallut tenir compte du biais spatial. Un grand nombre de cas sont situés dans le taluka de Nashik car la clinique ayant permis d'identifier les cas est située dans ce taluka. Pour corriger ce biais, nous avons donc estimé la probabilité d'observation d'un cas de SGC pour chacun des 15 talukas du district. Ces estimations sont basées sur l'expérience du médecin ayant recueilli les données, et dépendent de la distance géographique séparant la clinique de l'adresse de chaque enfant, de critères sociologiques et de la présence d'une autre clinique dans chaque taluka. Cette correction a été utilisée pour modifier la population à risque de chaque taluka afin d'obtenir des résultats non biaisés. A partir de la population corrigée de chaque taluka, nous avons appliqué la statistique de scan spatio-temporelle aux 47 cas.

Après correction du biais, la valeur de la statistique de test est 11.2 et le cluster le plus probable correspond au taluka de Niphad ($p = 0.044$). Sur les 5 cas de SGC diagnostiqués dans ce taluka, 3 l'ont été entre le 20 et le 24 octobre 2001. Il s'est avéré que les 3 enfants avaient tous participé à la même cérémonie de mariage quelques jours avant l'éruption dermatologique. Les 3 enfants proviennent de communautés différentes et n'ont aucun autre lien entre eux (crèche ou école par exemple). Après enquête auprès des familles, un quatrième enfant atteint de SGC a développé des symptômes compatibles avec le SGC 8 jours après la cérémonie, et a été suivi par une autre clinique. Sachant que seulement 6 enfants étaient présents à la cérémonie, nous avons pu conclure à une mini-épidémie de 4 patients.

7.3 Approche en cours

Une approche que nous proposons de développer consiste à généraliser au cas spatio-temporel la méthode de régression sur données transformées présentées dans le chapitre 4 pour la détection de clusters spatiaux. Cette approche viendrait en complément des méthodes existantes en ce sens qu'elle s'appliquerait sur des données individuelles plutôt que groupées tout en utilisant la population sous-jacente. Cette approche est en cours de développement et seule une brève présentation est donnée ici.

7.3.1 Détermination de la trajectoire

L'extension de la méthode spatiale au cas spatio-temporel repose sur la détermination de la trajectoire. Le principe est le suivant : plutôt que de déterminer l'ordre des points dans la trajectoire à partir de la distance au plus proche voisin spatial, l'inclusion des points dans la trajectoire est définie par l'ordre temporel d'apparition des évènements.

Soient $t_1, \ldots, t_n$ les temps d'occurence de n évènements dans une région A. Notons $x_1, \ldots, x_n$ les coordonnées spatiales de ces évènements. Le processus de point spatio-temporel est défini par $(t_k, x_k)_{k=1,\ldots,n}$.

Notons $t_{(1)}, \ldots, t_{(n)}$ les temps ordonnés par ordre croissant. $x_{(k)}$ désigne alors la localisation spatiale de l'évènement d'ordre k, celui qui est intervenu au temps $t_{(k)}$. Ainsi, le premier point de la trajectoire est l'évènement intervenu en premier dans le temps. Nous pouvons maintenant associer à l'ordre k la distance spatiale séparant deux évènements successifs temporellement : $d_k = d\left(x_{(k)}, x_{(k+1)}\right)$.

L'idée est que les points inclus dans un cluster spatio-temporel se suivent temporellement et sont proches spatialement. Ils seront donc repérés par des points successifs sur la trajectoire avec des distances associées qui sont faibles.

Il nous faut aborder le cas ou plusieurs évènements interviennent en même temps. Supposons que la trajectoire est connue jusqu'au point d'ordre $k : t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ et $x_{(k)}$ est le dernier point connu de la trajectoire. Si le temps de la série $t_1, \ldots, t_n$ immédiatement supérieur à $t_{(k)}$ correspond à deux évènements localisés en x_i et x_j ($t_i = t_j$), le point d'ordre $k+1$ sera le plus proche spatialement de $x_{(k)}$: $x_{(k+1)} = x_i$ si $d\left(x_{(k)}, x_i\right) < d\left(x_{(k)}, x_j\right)$ et $x_{(k+1)} = x_j$ si $d\left(x_{(k)}, x_j\right) < d\left(x_{(k)}, x_i\right)$. Cette procédure se générale aisément au cas où plus de deux points sont temporellement *ex aequo*.

7.3.2 Correction pour populations non-homogènes

Nous allons faire l'hypothèse que la répartition de la population à risque ne varie pas au cours du temps. Cette hypothèse est acceptable lorsque les données sont recueillies sur une période temporelle qui n'est pas trop grande. Nous ne proposons donc pas ici de correction temporelle de l'évolution de la population à risque. Par contre, l'inhomogénéïté de la répartition spatiale doit être prise en compte par le biais d'une correction de la distance spatiale associée à chaque point. Un point situé dans une zone ayant une forte densité de population a plus de chance d'avoir un voisin temporel proche spatialement (comparativement à un point situé dans une zone à faible densité). La distance associée à ce point doit donc être augmentée proportionnellement à la densité locale de population autour de ce point.

Nous reprenons les notations du chapitre 4 pour définir la population sous-jacente W constituée de N individus $\{w_i : i = 1, \ldots, N\}$ avec $N \gg n$. La densité de population d'une région B est $\delta(B) = \#\{i/w_i \in B\}/|B \bigcap A|$ où $|B \bigcap A|$ désigne l'aire de la partie de B qui est incluse dans A. La distance corrigée, notée d_k^c, d'un point $x_{(k)}$ à son plus proche voisin temporel se calcule alors comme suit :

$$d_k^c = d_k \times \frac{\delta\left(S(x_{(k)}, r)\right)}{\delta(A)},$$

où $S(x_{(k)}, r)$ est la sphère de centre $x_{(k)}$ et de rayon r. $\delta\left(S(x_{(k)}, r)\right)$ représente la densité de population locale autour de $x_{(k)}$. Elle dépend du choix du rayon r. Un choix qui semble naturel est de considérer $r = \sqrt{\frac{|A|}{n\pi}}$. Ainsi, en remarquant que $\sum_{k=1}^{n} |S(x_{(k)}, r)| = |A|$, la

densité locale d'un point est calculée à partir d'une zone dont l'aire est égale à l'aire totale de A divisé par le nombre de points n.

7.3.3 Localisation et détection

Nous pouvons considérer la série $(k, d_k^c)_{k=1,\dots,n-1}$ obtenue par transformation des données initiales $(t_k, x_k)_{k=1,\dots,n}$. Rappelons que k désigne l'ordre temporel d'apparition d'un évènement et que d_k^c est la distance spatiale corrigée d'un point à son plus proche voisin temporel. La localisation et la détection des clusters spatio-temporels s'effectue de la même façon que dans le cas purement spatial (chapitre 4). L'hypothèse d'absence de cluster se traduit ici par une répartition uniforme des temps t_k au sein de la période d'étude et par une répartition uniforme des points x_k dans le domaine d'étude. Les échantillons utilisés lors de l'étape de détection sont générés sous cette hypothèse.

7.3.4 Discussion

L'approche présentée ici n'est pas encore implémentée. Lorsque ce sera fait, nous pourrons comparer cette méthode avec la statistique de scan spatio-temporelle. Nous pourrons également étudier son efficacité dans différentes situations d'agrégation des données.

Par construction de la méthode, les clusters qui seront les plus à même d'être détectés seront ceux constitués par des évènements qui se suivent dans le temps tout en étant proches spatialement. Par contre, la méthode ne fonctionnera pas en présence de deux agrégats spatialement éloignés qui interviennent parallèlement dans le temps : la trajectoire a de fortes chances de faire des aller-retours entre les deux clusters, associant des distances élevées à une partie des points inclus dans les deux agrégats. Ce cas de figure met en avant la nécessité d'effectuer au préalable une analyse purement spatiale et une analyse purement temporelle. La première permettra de localiser deux clusters spatiaux, et la seconde mettra en évidence un cluster temporel composé des évènements des deux agrégats. Notons toutefois que cette situation d'agrégation bien particulière fera également échouer la statistique de scan spatio-temporelle : le cylindre représentant le cluster le plus probable englobera les deux clusters ainsi que les points situés spatialement entre eux.

Conclusion générale

Nous avons ici abordé en premier lieu la détection de clusters temporels. C'est un vaste domaine de recherche qui a déjà fait l'objet d'une littérature abondante et notamment d'une thèse à part entière. Concernant les méthodes existantes sur le sujet, nous nous sommes donc essentiellement concentrés sur la méthode de référence, la statistique de scan temporelle. Nous avons également proposé une utilisation pratique du théorème de Lucien Le Cam. Concernant notre contribution dans le cadre temporel, nous avons abordé ce qui nous paraît être deux faiblesses des méthodes existantes et qui concernent l'utilisation de simulations et les données censurées.

Le test de la significativité d'un cluster temporel sans avoir recourt à l'utilisation de simulations est une problématique qui n'est pas aisée à traiter. Nous avons ici proposé une solution qui est intermédiaire dans le sens où nous n'avons pas proposé de p-valeur exacte mais une borne supérieure pour cette dernière. Nous avons également traité le problème des données temporelles incomplètes. Cette situation a été rencontrée dans l'étude du lien entre cancer et rayonnements ionisants où certaines dates de diagnostic n'était pas complètement renseignées. Nous avons en premier lieu traité ce problème en ne considérant dans les analyses que les observations dont la date était complète. Puis nous avons proposé une approche permettant de prendre en compte les observations ayant une information temporelle censurée. Nous avons pu constater l'impact de cette prise en compte sur les résultats. Nous sommes conscients que ces deux approches méritent d'être encore approfondies.

Ce travail a également été l'occasion de mettre en opposition les données groupées et les données ponctuelles, surtout et notamment dans le domaine spatial. Jusqu'à présent, l'utilisateur se trouvant en présence de données ponctuelles avait deux alternatives. La première consiste à garder cette finesse de résolution pour les données et à utiliser la statistique de scan circulaire ou elliptique, autrement dit une version paramétrique : dans ce cas la forme des agrégats potentiels est prédéfinie. La deuxième solution consiste à transformer les données afin de se ramener à des données groupées et ainsi pouvoir utiliser les versions non-paramétriques de la statistique de scan. L'avantage de cette deuxième approche est qu'elle laisse la possibilité au cluster d'avoir une forme arbitraire. Son inconvénient est de perdre de l'information en utilisant des données groupées plutôt que ponctuelles. La méthode de régression sur données transformées que nous avons proposée offre une troisième alternative en permettant d'analyser directement les données ponctuelles, et donc de conserver toute

l'information géographique, sans imposer de contraintes sur la forme des éventuels agrégats. Il nous semblait important de proposer cette nouvelle approche compte tenu du fait que l'information spatiale concernant les évènement de santé tend à devenir de plus en plus précise et fiable.

Cette méthode de régression sur données transformées permet également de prendre en compte la répartition de la population à risque. Nous avons vu que les méthodes d'analyse de clusters ont progressivement eu le souci d'intégrer un ajustement des résultats sur une éventuelle inhomogénéïté de la population sous-jacente. Les méthodes destinées à être appliquées sur des données groupées peuvent utiliser le nombre d'habitant par cellule pour effectuer cet ajustement. Pour les approches sur données ponctuelles, le problème est que bien souvent la taille de la population à risque n'est connue qu'à un niveau agrégé. Dans l'application de notre approche, nous avons contourné cette difficulté en répartissant aléatoirement les individus de la population à risque à l'intérieur de chaque cellule afin d'obtenir une population sous-jacente ayant le même niveau de résolution que les données ponctuelles sur lesquelles porte l'analyse de cluster. Cette transformation de la population nécessite cependant de connaître avec précision les limites géographiques des différentes cellules.

Comme nous avons pu le constater, les données obtenues par IRMf sont très particulières. Leur traitement requiert une méthodologie spécifique qui a été présentée et a servi de base à l'utilisation des méthodes de détection de clusters, telle que la statistique de scan spatiale ou la méthode de régression sur données transformées, sur ces données. L'analyse de l'étude sur la fluidité verbale a permis de mettre en évidence des clusters de pics d'activation. Ces agrégats sont révélateurs de zones cérébrales impliquées chez les plupart des sujets dans le traitement des taches auxquelles ils étaient soumis. L'interprétation de ces régions cérébrales nécessitera une collaboration avec des spécialistes de la neuro-imagerie.

Les perspectives de recherche dans la détection de clusters spatio-temporels sont nombreuses. Les approches proposées dans ce domaine sont prometteuses mais nécessitent d'être approfondies. Nous espérons être en mesure de proposer très prochainement une application pratique de l'approche évoquée dans la dernière partie qui se propose d'étendre la méthode développée dans le cadre spatial.

Cette approche spatiale a déjà fait l'objet d'une implémentation au sein d'un package R qui est disponible sur internet. Une perspective d'évolution de ce package est de le compléter avec l'approche spatio-temporelle, mais également d'y intégrer la méthode de régression temporelle. Nous obtiendrons ainsi un module complet d'analyse de clusters permettant d'analyser des données temporelles, spatiales et spatio-temporelles.

Bibliographie

Adler R. J. (1981). *The geometry of random fields*. Wiley.

Allard D. et Fraley C. (1997). *Non parametric maximum likelihood estimation of features in spatial point processes using voronoïtesselation*. Journal of American Statistical Association, vol. 92. pages 1485–1493.

Andrews D. W. K. (1993). *Tests for parameter instability and structural change with unknown change point*. Econometrica, vol. 61. pages 821–856.

Anton J. L., Dauchot K. et Pelegrini-Issac M. (2001). *SPM99 : Guide pratique pour utilisateur novice*.

Assunção R. M., Tavares A. et Kulldorff M. (2003). *An early warning system for space-time cluster detection*. In *GeoInfo*.

Assunção R., Costa M., Tavares A. et Ferreira S. (2006). *Fast detection of arbitrarily shaped disease clusters*. Statistics in Medicine, vol. 25. pages 723–742.

Bai J. et Perron P. (1998). *Estimating and testing linear models with multiple structural changes*. Econometrica, vol. 66. pages 47–78.

Bai J. et Perron P. (2003a). *Computation and analysis of multiple structural change models*. Journal of Applied Econometrics, vol. 18. pages 1–22.

Bai J. et Perron P. (2003b). *Critical values for multiple structural change tests*. Econometrics Journal, vol. 6. pages 72–78.

Bailey T. C. (2001). *Spatial statistical methods in health*. Cadernos de Saúde Pública, vol. 17. pages 1083–1098.

Bernstein S. (1946). *The theory of probabilities*. Gastehizdat Publishing House.

Besag J. et Newell J. (1991). *The detection of clusters in rare diseases*. Journal of the Royal Statistical Society A, vol. 154. pages 143–155.

Bickel P. J. et Breiman L. (1983). *Sums of functions of nearest neighbour distances, moment bounds, limit theorems and a goodness of fit test*. Annals of Probability, vol. 11. pages 185–214.

Bonaldi C. (2003). *Analyse de clusters sur le temps, application en carcinologie et en épidémiologie*. Thèse de doctorat de l'Université de Montpellier I.

Brett M., Penny W. et Kiebel S. (2003). *An introduction to random field theory. humain brain function II : SPM courses notes, chapter 14*.

Cuzick J. et Edwards R. (1990). *Spatial clustering for inhomogeneous populations*. Journal of the Royal Statistical Society B, vol. 52. pages 73–104.

David H. A. (1980). *Order statistics*. Wiley series in Probability and Mathematical Statistics.

Davies R. B. (1987). *Hypothesis testing when a nuisance parameter is present only under the alternative*. Biometrika, vol. 74. pages 33–43.

Dematteï C. (2006). SPATCLUS, R package for arbitrarily shaped multiple spatial cluster detection for case event data. Contributed package to the Comprehensive R Archive Network (CRAN), 2006. URL `http://www.R-project.org`.

Dematteï C. et Molinari N. (2006a). *Multiple temporal cluster detection test using exponential inequalities*. Far East Journal of Theoretical Statistics., vol. 19. pages 231–244.

Dematteï C. et Molinari N. (2006b). *P-value calculations for multiple temporal cluster detection*. En révision dans les Comptes Rendus de l'Académie des Sciences, Paris, Série I.

Dematteï C., Molinari N. et Daurès J. P. (2005), *Analyse de clusters de cancers et rayonnements ionisants : Rapport d'études temporelle et spatiale*. Rapport technique.

Dematteï C., Molinari N. et Daurès J. P. (2006a). *Arbitrarily shaped multiple spatial cluster detection for case event data*. A paraître dans Computational Statistics and Data Analysis. doi : 10.1016/j.csda.2006.03.011.

Dematteï C., Molinari N. et Daurès J. P. (2006b). *SPATCLUS : an R package for arbitrarily shaped multiple spatial cluster detection for case event data*. Computer Methods and Programs in Biomedicine, vol. 84. pages 42–49.

Dempster A., Laird N. et Rubin D. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society B, vol. 39. pages 1–38.

Diggle P. J., Morris S. et Morton-Jones T. (1999). *Case-control isotonic regression for investigation of elevation in risk around a point source*. Statistics in Medicine, vol. 18. pages 1605–1613.

Duczmal L. et Assunção R. (2004). *A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters*. Computational Statistics and Data Analysis, vol. 45. pages 269–286.

Dwass M. (1957). *Modified randomization tests for nonparametric hypotheses*. Annals of Mathematical Statistics, vol. 28. pages 181–187.

Elderer F., Myers E. et Mantel M. (1964). *A statistical problem in space and time : Do leukemia cases come in clusters ?* Biometrics, vol. 20. pages 623–626.

Friston K. J. (2003). *Introduction : Experimental design and statistical parametric mapping. humain brain function II : SPM courses notes, chapter 1.*

Friston K. J., Holmes A. P., Poline J. B., Price C. J. et Frith C. D. (1996). *Detecting activations in PET and fMRI : levels of inference and power*. NeuroImage, vol. 4. pages 223–235.

Glaser D. E. et Friston K. J. (2003). *Variance components. Humain Brain Function II : SPM courses notes, chapter 9.*

Green P. (1995). *Reversible jump markov chain monte carlo computation and bayesian model determination*. Biometrika, vol. 82(4). pages 711–732.

Grimson R. (1991). *A versatile test for clustering and a proximity analysis of neurons*. Methods of Information in Medicine, vol. 30. pages 299–303.

Henson R. (2003). *Analysis of fMRI time series : Linear time invariant models, event related fMRI and optimal experimental design. humain brain function II : SPM courses notes, chapter 10.*

Iyengar V. S. (2004). *On detecting space-time clusters*. In *KDD '04 : Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA. ACM Press. ISBN 1-58113-888-1. pages 587–592.

Kiebel K. et Holmes A. P. (2003). *The general linear model. humain brain function II : SPM courses notes, chapter 7.*

Knox G. (1959). *Secular pattern of congenital oesophageal atresia*. British Journal of Preventive Social Medicine, vol. 13. pages 222–226.

Knox G. (1964). *The detection of space-time interactions*. Applied Statistics, vol. 13. pages 25–29.

Kulldorff M. (1997). *A spatial scan statistic*. Communications in Statistics - Theory and Methods, vol. 26. pages 1481–1496.

Kulldorff M. (1998). *Evaluating cluster alarms : a space-time scan statistic and brain cancer in Los Alamos, New Mexico*. American Journal of Public Health, vol. 88. pages 1377–1380.

Kulldorff M. (1999a). *An isotonic spatial scan statistic for geographical disease surveillance*. Journal of the National Institute of Public Health, vol. 48. pages 94–101.

Kulldorff M. (1999b). *The Knox method and other tests for space-time interaction*. Biometrics, vol. 55. pages 544–552.

Kulldorff M. (2001). *Prospective time periodic geographical disease surveillance using a scan statistic*. Journal of the Royal Statistical Society A, vol. 164. pages 61–72.

Kulldorff M. (2002), *Tests for spatial randomness adjusted for an inhomogeneity : A general framework*. Rapport technique.

Kulldorff M., Heffernan R., Hartman J., Assunção R. M. et Mostashari F. (2005). *A space-time permutation scan statistic for the early detection of disease outbreaks*. Public Library of Science Medicine, vol. 2. pages 216–224.

Kulldorff M., Huang L., Pickle L. et Duczmal L. (2006). *An elliptic spatial scan statistic*. A paraître dans Statistics in Medicine.

Kulldorff M. et Information Managements Services I. (2004). *Satscan v5.1 : Software for the spatial and space-time scan statistics. http* : *//www.satscan.org*.

Kulldorff M. et Nagarwalla N. (1995). *Spatial disease clusters : Detection and inference*. Statistics in Medicine, vol. 14. pages 799–810.

Larsen R. J., Holmes C. L. et Heath C. W. (1973). *A statistical test for measuring unimodal clustering : A description of the test and of its application to cases of acute leukemia in metropolitan atlanta, georgia*. Biometrics, vol. 29. pages 301–309.

Lawson A. (2001). *Statistical methods in spatial epidemiology*. Wiley.

Le Cam L. (1958). *Un théorème sur la division d'un intervalle par des points pris au hasard*. Publications de l'Institut de Statistique de l'Université de Paris VII. pages 7–16.

Lewis P. et Shedler G. (1979). *Simulation of nonhomogeneous poisson processes by thinning*. Naval Research Logistics Quarterly, vol. 26. pages 403–413.

Molinari N., Bonaldi C. et Daurès J. P. (2001). *Multiple temporal cluster detection*. Biometrics, vol. 57. pages 577–583.

Nagarwalla N. (1996). *A scan statistic with variable window*. Statistics in Medicine, vol. 15. pages 845–850.

Naus J. I. (1965). *The distribution of the size of the maximum cluster of points on a line*. Journal of the American Statistical Association, vol. 60. pages 532–538.

Naus J. I. (1966). *A power comparison of two sets of non-random clustering*. Technometrics, vol. 8. pages 493–517.

Neill D. B., Moore A. W., Pereira F. et Mitchell T. (2005). *Detecting significant multidimensional spatial clusters.* In *Advances in Neural Information Processing Systems*, volume 17. pages 969–976.

Owen A. (1991). *Comment on "Multivariate adaptative regression splines" by Friedman J.H.* The Annals of Statistics, vol. 19. pages 102–112.

Patil G. P. et Taillie C. (2004). *Upper level set scan statistic for detecting arbitrarily shaped hotspots.* Environmental and Ecological Statistics, vol. 11. pages 183–197.

Penny W. D. et Holmes A. J. (2003). *Random-effects analysis. humain brain function II : SPM courses notes, chapter 12.*

Poline J. B., Holmes A. P., Worsley K. J. et Friston K. J. (1997a). *Statistical inference and the theory of random fields. humain brain function : SPM courses notes, chapter 4.*

Poline J. B., Worsley K. J., Evans A. C. et Friston K. J. (1997b). *Combining spatial extend and peak intensity to test for activations in functional imaging.* NeuroImage, vol. 5. pages 83–96.

R Development Core Team . *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2006. URL `http://www.R-project.org`.

Ripley B. D. (1977). *Modelling spatial patterns.* Journal of the Royal Statistical Society B, vol. 39. pages 172–192.

Schlosser R., Hutchinson M., Joseffer S., Rusinek H., Saarimaki A., Stevenson J., Dewey S. L. et Brodie J. D. (1998). *Functional magnetic resonance imaging of the human brain activity in a verbal fluency task.* J. Neurol. Neurosurg. Psychiatry, vol. 64. pages 492–498.

Sebastian R., Diaz M. E., Ayala G., Letinic K., Moncho-Bogani J. et Toomre D. (2006). *Spatio-temporal analysis of constitutive exocytosis in epithelial cells.* IEEE/ACM Transactions on Computational Biology and BioInformatics, vol. 3. pages 17–32.

Szalay A. S., Budavári T., Connolly A., Gray J., Matsubara T., Pope A. et Szapudi I. (2002), *Spatial clustering of galaxies in large datasets.* Rapport technique.

Takahashi K. et Tango T. (2006). *An extended power of cluster detection tests.* Statistics in Medicine, vol. 25. pages 841–852.

Tango T. (1984). *The detection of disease clustering in time.* Biometrics, vol. 40. pages 15–26.

Tango T. (1995). *A class of tests for detecting 'general' and 'focused' clustering of rare diseases.* Statistics in Medicine, vol. 14. pages 2323–2334.

Tango T. (2000). *A test for spatial disease clustering adjusted for multiple testing.* Statistics in Medicine, vol. 19. pages 191–204.

Tango T. et Takahashi K. (2005). *A flexibly shaped spatial scan statistic for detecting clusters.* International Journal of Health Geographics, vol. 4-11.

Turnbull B. W., Iwano E. J., Burnett W. S., Howe H. L. et Clark L. C. (1990). *Monitoring for clusters of disease : Application to leukemia incidence in upstate new york.* American Journal of Epidemiology, vol. 132. pages 136–143.

Vinson T. et Baldry E. (1999). *The spatial clustering of child maltreatment : Are micro-social environments involved ?* Australian Institute of Criminology, Trends and Issues, vol. 119.

Wallenstein S. (1989). *Testing for a pulse in seasonal event data.* Biometrics, vol. 45. pages 817–830.

Weinstock M. A. (1981). *A generalized scan statistic test for the detection of clusters.* International Journal of Epidemiology, vol. 10. pages 289–293.

Whittemore A. S., Friend N., Brown B. W. et Holly E. A. (1987). *A test to detect clusters of disease.* Biometrika, vol. 74. pages 631–635.

Worsley K. J. et Friston K. J. (1995). *Analysis of fMRI time-series revisited - again.* NeuroImage, vol. 2. pages 173–181.

Worsley K. J., Marrett S., Neelin P. et Evans A. C. (1996). *Searching scale for activation in PET images.* Human Brain Mapping, vol. 4. pages 74–90.

Yoshida M., Naya Y. et Miyashita Y. (200l). *Anatomical organization of forward fiber projections from area TE to perirhinal neurons representing visual long-term memory in monkeys.* Proc. Natl. Acad. Sci. USA, vol. 100. pages 4257–4262.

Annexe A

Description du package SPATCLUS

Le contenu du package SPATCLUS est présenté dans cette annexe. Le diagramme fournit par la Figure A.1 permet d'avoir une vue d'ensemble du package. Les principaux algorithmes, ceux concernant la transformation des données et la localisation des points de cassure, sont décrits. Le package implémente essentiellement la méthode de détection de clusters spatiaux décrites dans la Section 4. La fonction principale est la fonction *clus*( ). La statistique de scan spatial de Kulldorff [1997] étant une méthode de référence, le package contient également un module d'exportation au format du logiciels SatScan (Kulldorff et Information Managements Services [2004]).

A.1 Interface utilisateur

Une fois que R a démarré, une fenêtre appelée "R Console" apparaît. A l'intérieur de cette fenêtre, l'utilisateur tape ses commandes et R affiche les résultats des calculs demandés. Chaque commande doit être écrite à droite du symbole ">". Le résultat d'une commande peut être sauvegardé dans un objet R en utilisant l'opérateur d'assignement "$<-$". Toutes les fonctions sont appelées de la même façon. Par exemple la commande

$$resclus <- clus(data = data\_ex, pop = pop\_ex, limx = c(0,1), limy = c(0,1))$$

analysera le jeu de données de coordonnées des cas *data_ex* avec le jeu de données de coordonnées de la population *pop_ex*. Le domaine d'étude est ici définit comme étant le carré unité. Les résultats de cette analyse seront sauvegardés dans *resclus*, un objet R de type liste.

Afin de pouvoir utiliser le package SPATCLUS, l'utilisateur doit au préalable taper la commande

$> library(spatclus)$

qui chargera le package dans R.

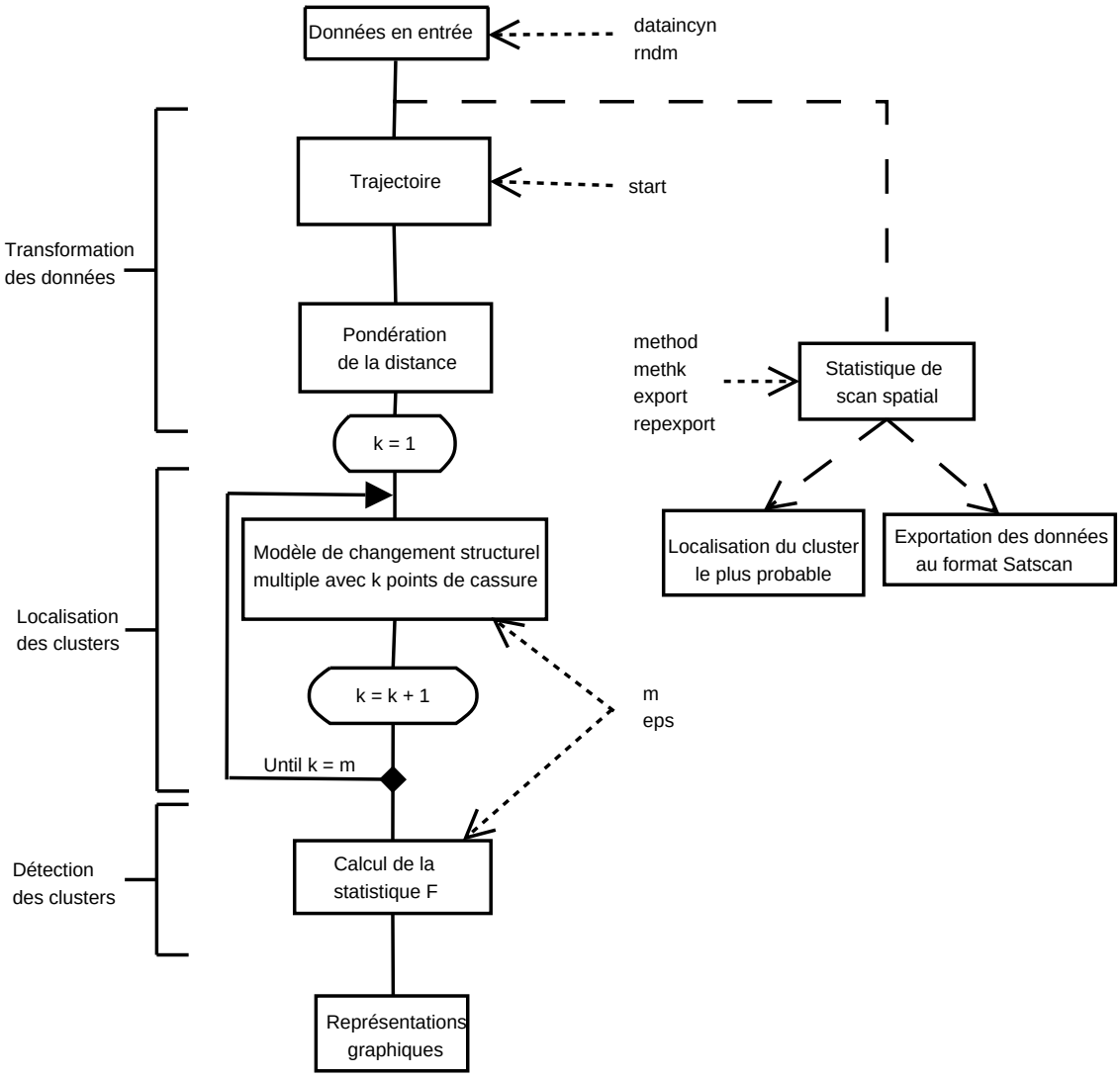


FIG. A.1 – Diagramme décrivant le package SPATCLUS.

A.2 Données en entrée

En $2D$, la fonction *clus*( ) a 4 arguments essentiels qui doivent être spécifiés :

data : Objet R de type Data frame qui contient 2 colonnes donnant les coordonnées des cas.

pop : Objet R de type Matrice qui contient 2 colonnes donnant les coordonnées des individus de la population sous jacente.

limx : Vecteur de 2 éléments contenant les limites du domaine d'étude sur l'axe des abscisses.

limy : Vecteur de 2 éléments contenant les limites du domaine d'étude sur l'axe des ordonnées.

En $3D$, les objets **data** et **pop** contiennent une troisième colonnes. L'utilisateur doit également spécifier le paramètre **limz**, un vecteur de 2 éléments contenant les limites du domaine d'étude sur le troisième axe.

A.3 Paramètres optionnels

La fonction *clus*( ) possède également plusieurs arguments optionnels qui agissent sur les différentes étapes de la méthode. Des valeurs par défaut (VPD) sont attribuées à chacun de ces paramètres.

- Données en entrée :

  dataincyn (VPD="n") : "y" signifie que les cas sont déjà inclus dans la population sous-jacente. "n" signifie que les cas ne sont pas inclus et ajoute les coordonnées de *data* à celles de *pop* .

  rndm (VPD=NaN) : Vecteur identifiant les lignes qui contiennent les coordonnées des cas dans la matrice *pop* (seulement si datainc="y").

- Trajectoire :

  start (VPD=1) : Indique le rang du premier point de la trajectoire en terme de distance par rapport au bord du domaine. 1 signifie que le premier point de la trajectoire est le point le plus proche du bord.

- Localisation et détection des clusters :

  m (VPD=5) : Nombre maximum de points de cassure.

  eps (VPD=0.2) : Taille minimale du cluster (ratio par raport au nombre total de cas).

- Statistique de scan spatial et module d'exportation au format SatScan :

  method (VPD=1) : 1 pour la méthode de détection de clusters multiples, 2 pour la statistique de scan spatial, 3 pour les 2 méthodes.

  methk (VPD=3) : Dans le calcul de la statistique de scan spatial, 1 pour le modèle de Bernoulli, 2 pour le modèle de Poisson, 3 pour les deux modèles.

TAB. A.1 – Algorithme de transformation des données

READ $data$, pop, pas, $x_{(1)}$
FOR $k = 1$ to $n - 1$
  IF $k > 1$ THEN
    $pop \leftarrow pop \setminus \{u/d\left(x_{(k-1)}, u\right) \leqslant d\left(x_{(k-1)}, x_{(k)}\right)\}$
  ENDIF
  $a_k \leftarrow \max_{u \in pop} d\left(x_{(k)}, u\right)$
  SET S to 0
  FOR $r = 0$ to a_k by pas
    SET $rpop$ to pop
    $rpop \leftarrow rpop \setminus \{u/d\left(x_{(k)}, u\right) > r\}$
    $S \leftarrow S + \left(1 - \frac{\#rpop}{\#pop}\right)^{n-k}$
  ENDFOR
  $E[d_k] \leftarrow pas \times \left(S - \frac{1}{2}\right)$
  $x_{(k+1)} \leftarrow \text{argmin}_{x \in data} d\left(x_{(k)}, x\right)$
  $d_k \leftarrow d\left(x_{(k)}, x_{(k+1)}\right)$
  $d_k^w \leftarrow \frac{d_k}{E[d_k]}$
  $data \leftarrow data \setminus \{x_{(k)}\}$
  PRINT $x_{(k)}$, d_k^w
ENDFOR

export (VPD="n") : Si $method = 2$ ou $method = 3$, et si $export =$ "y", les données seront exportées au format SatScan dans le répertoire "repexport".

repexport (no VPD) : Si $export =$ "y", répertoire dans lequel les données seront exportées.

A.4 Algorithme de transformation des données

Dans cette section, l'algorithme utilisé pour la détermination de la trajectoire et la pondération de la distance est détaillé. La méthodologie correspondante est décrite dans les sections 4.1, 4.2 et 4.3.

Dans l'algorithme présenté dans la Table A.1 et écrit en pseudocode, $data = \{x_1, \ldots, x_n\}$ est l'ensemble des localisations des n cas et $pop = \{u_1, \ldots, u_N\}$ est l'ensemble des localisations des N individus appartenant à la population sous-jacente. La trajectoire est initialisée par le choix du premier point de la trajectoire $x_{(1)}$ dans $data$, et nous le considérons comme donné dans l'algorithme. Pour une meilleure compréhension, nous avons utilisé un langage ensembliste plutôt que matriciel.

Quelques explications sont nécessaires à une compréhension complète de la correspondance entre les quantités utilisées dans cet algorithme et celles utilisées dans les équations (4.3) et (4.3). Dans la $k^{\text{ième}}$ itération de la boucle FOR globale :

- après le bloc IF, *pop* représente A_{k-1} et *#pop* est utilisé pour approximer la quantité $N \times \int_{A_{k-1}} f(x)dx$,
- dans la boucle FOR imbriquée, *rpop* représente $A_{k-1} \bigcap S(x_{(k)}, r)$ et *#rpop* est utilisé pour approximer la quantité $N \times \int_{A_{k-1} \bigcap S(x_{(k)},r)} f(x)dx$,
- la boucle FOR imbriquée permet de calculer la quantitée $pas \times \left(S - \frac{1}{2}\right)$ qui représente une estimation de

$$\int_0^a \left[1 - \frac{\int_{A_{k-1} \bigcap S(x_{(k)},r)} f(x)dx}{\int_{A_{k-1}} f(x)dx}\right]^{n-k} dr$$

  par la méthode des trapèzes,
- la dernière étape permet de sauvegarder les coordonnées $x_{(k)}$ du $k^{\text{ième}}$ cas de la trajectoire ainsi que la distance pondérée d_k^w qui lui est associée.

A.5 Algorithme de localisation des points de cassure

Considérons la régression de la série ordonnée des distances pondérées $\{d_k^w : k = 1, \ldots, n-1\}$ sur l'ordre de sélection k. La fonction de régression est donnée par l'équation (4.4). Afin de déterminer les points de cassure pour le modèle à m points de cassure dans l'équation (4.4), nosu avons utilisé l'approche par programmation dynamique proposée par Bai et Perron [2003a] qui permet de réduire considérablement les temps de calcul. L'algorithme, divisé en deux parties, donné dans la Table A.2 est une traduction en langage pseudocode de cette méthode.

Le paramètre ϵ et la partition optimale $(\hat{T}_1, \ldots, \hat{T}_m)$ sont définis dans la section 4.4.

Cet algorithme donne une description complète de la façon dont les points de cassure sont calculés. Dans la première partie, la somme des carrés des résidus notée $ssr_{i,j}$ est calculée seulement pour les segments $[i; j]$ qui sont nécessaires à la détermination des m points de cassure. Dans la seconde partie, la partition optimale est obtenue en résolvant le problème récursif $S_{r,j} = \min_{rh \leqslant i \leqslant j-h}[S_{r-1,i} + ssr_{i+1,j}]$. $S_{r,j}$ désigne la somme des carrés des résidus associée à la partition optimale contenant r points de cassure pour les j premières observations.

A.6 Données en sortie et représentations graphiques

La sortie de la fonction *clus*( ) est une liste d'objets qui contient :

res : Une matrice contenant, pour chaque point ordonné par son rang dans la trajectoire, la distance à son plus proche voisin, l'espérance de cette distance, et la distance pondérée.

pop : Une matrice avec 2 ou 3 colonnes (selon que l'on se trouve en $2D$ ou en $3D$) qui contient les coordonnées des individus de la population sous-jacente.

bc : Une liste de vecteurs de tailles 1 à M. Le $k^{\text{ième}}$ élément de la liste contient l'estimation des points de cassure pour le modèle avec k points de cassure.

TAB. A.2 – Algorithme de localisation des points de cassure

READ m, ϵ, d_1^w, d_2^w, ..., d_{n-1}^w
$T \leftarrow n-1$
$h \leftarrow |T\epsilon|$
FOR $i = 1$ to T
    FOR $j = 1$ to T
        IF $j - i \geqslant h - 1$
            $\overline{d_{i,j}^w} \leftarrow \frac{1}{j-i+1} \sum_{k=i}^{j} d_k^w$
            $ssr_{i,j} \leftarrow \sum_{k=i}^{j} \left(d_k^w - \overline{d_{i,j}^w} \right)^2$
        ENDIF
    ENDFOR
ENDFOR

IF $m = 1$
    $\hat{T}_1 \leftarrow \text{argmin}_{h \leqslant j \leqslant T-h}[ssr_{1,j} + ssr_{j+1,T}]$
ENDIF
FOR $j = h$ to T
    $S_{0,j} \leftarrow ssr_{1,j}$
ENDFOR
IF $m > 1$
    FOR $r = 1$ to $m - 1$
        FOR $j = (r+1)h$ to $T - (m-r)h$
            $S_{r,j} \leftarrow \min_{rh \leqslant i \leqslant j-h}[S_{r-1,i} + ssr_{i+1,j}]$
            $b_{r,j} \leftarrow \text{argmin}_{rh \leqslant i \leqslant j-h}[S_{r-1,i} + ssr_{i+1,j}]$
        ENDFOR
    ENDFOR
    $S_{m,T} \leftarrow \min_{mh \leqslant j \leqslant T-h}[S_{m-1,j}]$
    $\hat{T}_m \leftarrow \text{argmin}_{mh \leqslant j \leqslant T-h}[S_{m-1,j}]$
    FOR $k = m - 1$ to 1
        $\hat{T}_k \leftarrow b_{k,\hat{T}_{k+1}}$
        PRINT $\hat{T}_k$
    ENDFOR
ENDIF

stat : Une liste de valeurs de la statistique non corrigées (F), la valeur de la statistique corrigée (wdm), la valeur du seuil pour la statistique WDM ($wdms$), la significativité ($signif$) et le nombre de points de cassure qui optimise la statistique WDM ($kmax$).

kulld.p : Un vecteur contenant les résultats de la méthode de scan spatial avec le modèle de Poisson. *lambda* est la valeur de la statistique de scan, *loglambda* est son logarithme, cx et cy sont les coordonnées du centre du cercle et $rayon$ est son rayon.

kulld.b : Un vecteur contenant les mêmes résultats que ci-dessus avec le modèle de Bernouilli.

Cette liste d'objets peut être utilisée comme argument pour les deux fonctions de représentation graphique. La fonction $plotreg(\ )$ affiche les points avec l'ordre de sélection en abscisse et la distance pondérée en ordonnée, et trace la fonction de régression de régression avec k points de cassure. La fonction $plotclus(\ )$ représente le nuage de points et les clusters localisés pour le modèle avec k points de cassure. k est généralement égal à la valeur de $stat\$kmax$.

A.7 Module d'exportation au format SatScan

Dans ce module, la localisation du cluster par la statistique de scan spatial de Kulldorff [1997] est implémentée, mais la p-valeur n'est pas fournie. Pour une analyse complète avec cette méthode incluant l'inférence par la méthode de Monte Carlo, on peut utiliser la logiciel Satscan (Kulldorff et Information Managements Services [2004]) disponible gratuitement. Le package SPATCLUS permet aux utilisateurs d'exporter les données dans un format directement utilisable par ce logiciel. Pour cela, il faut spécifier les paramètres suivants :

method $= 3$

methk $= 1$ ou 2 (modèle de Bernouilli ou de Poisson)

export $=$ "y"

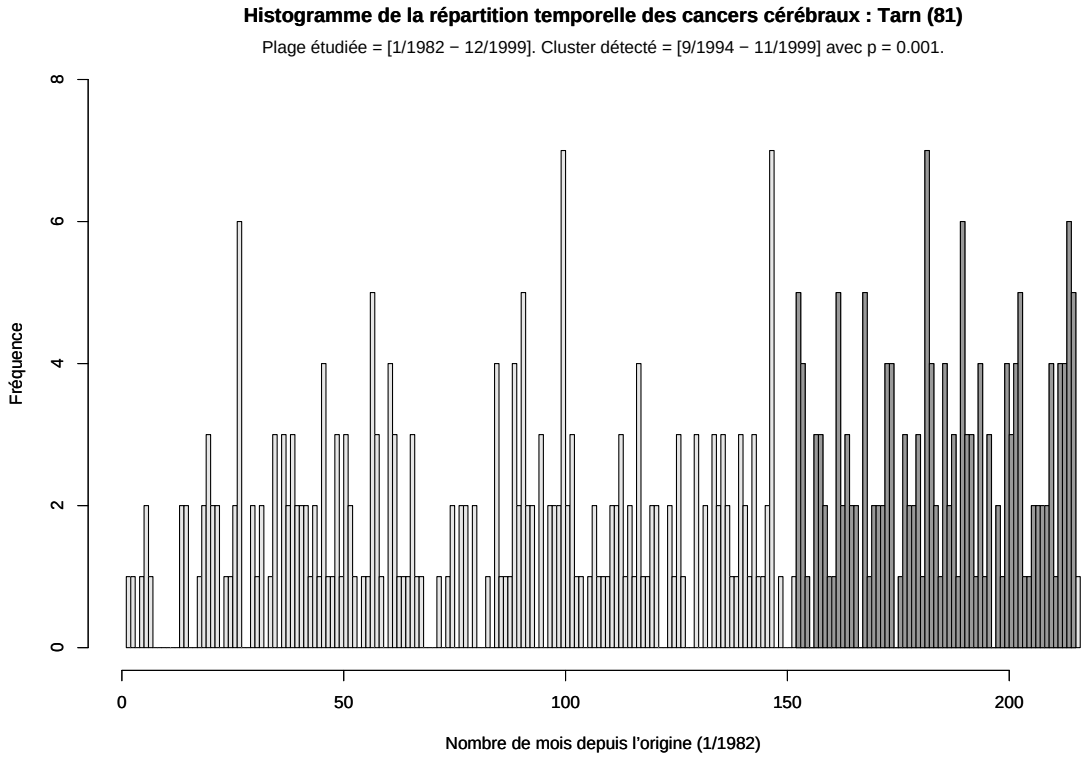
repexport $=$ "dir". dir désigne le répertoire dans lequel les données seront exportées au format Satscan.

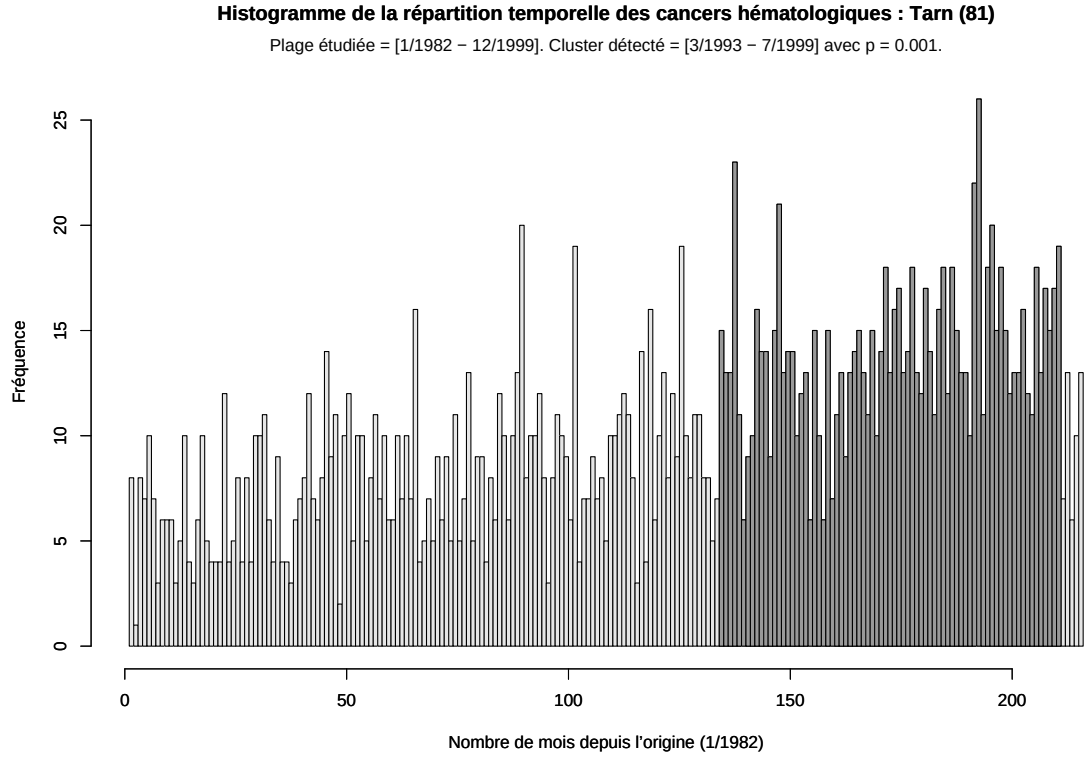
Annexe B

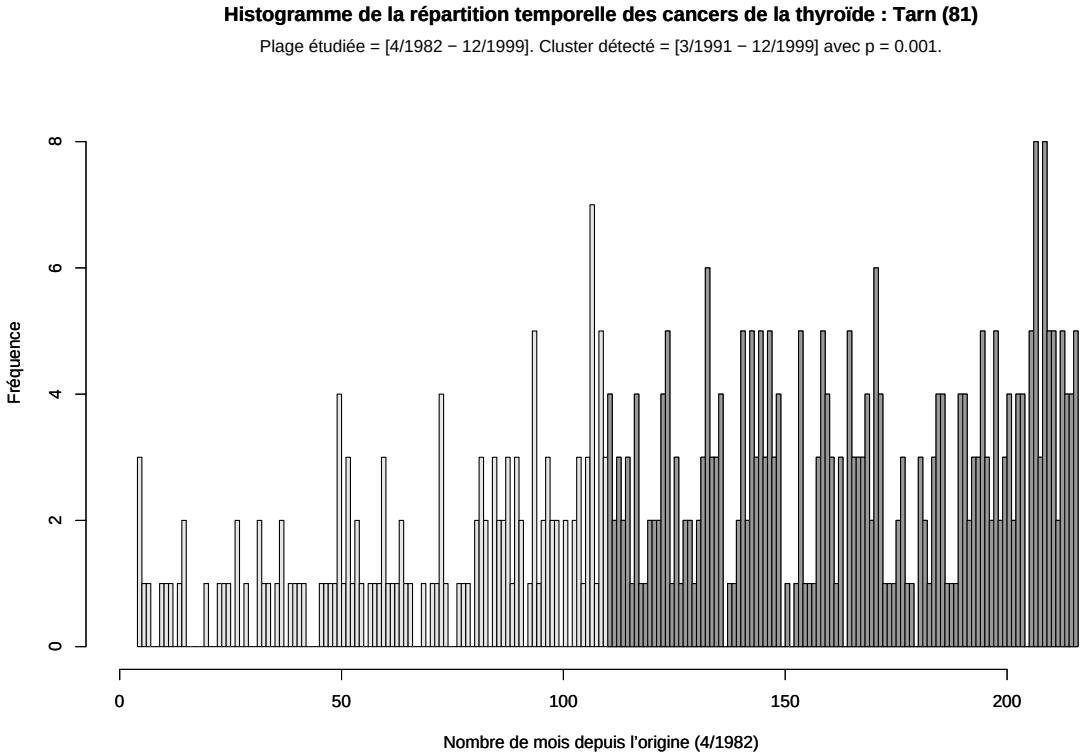
Clusters de cancers : graphiques

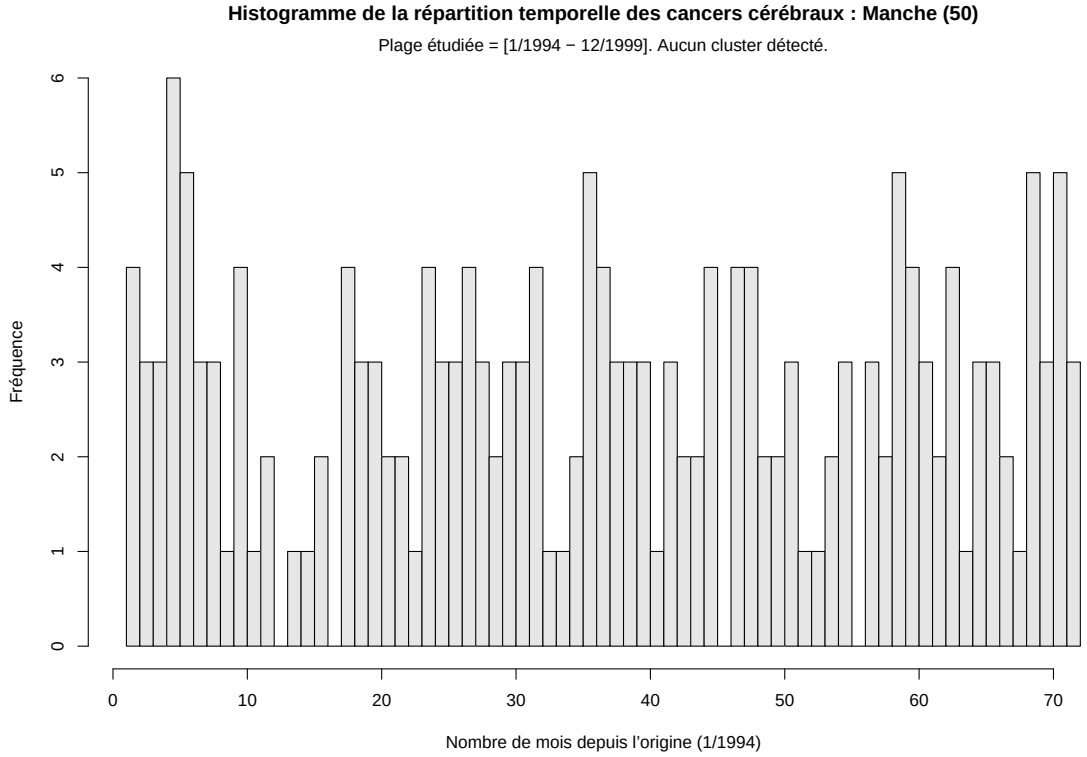
B.1 Histogrammes de la répartition temporelle des cas de cancers

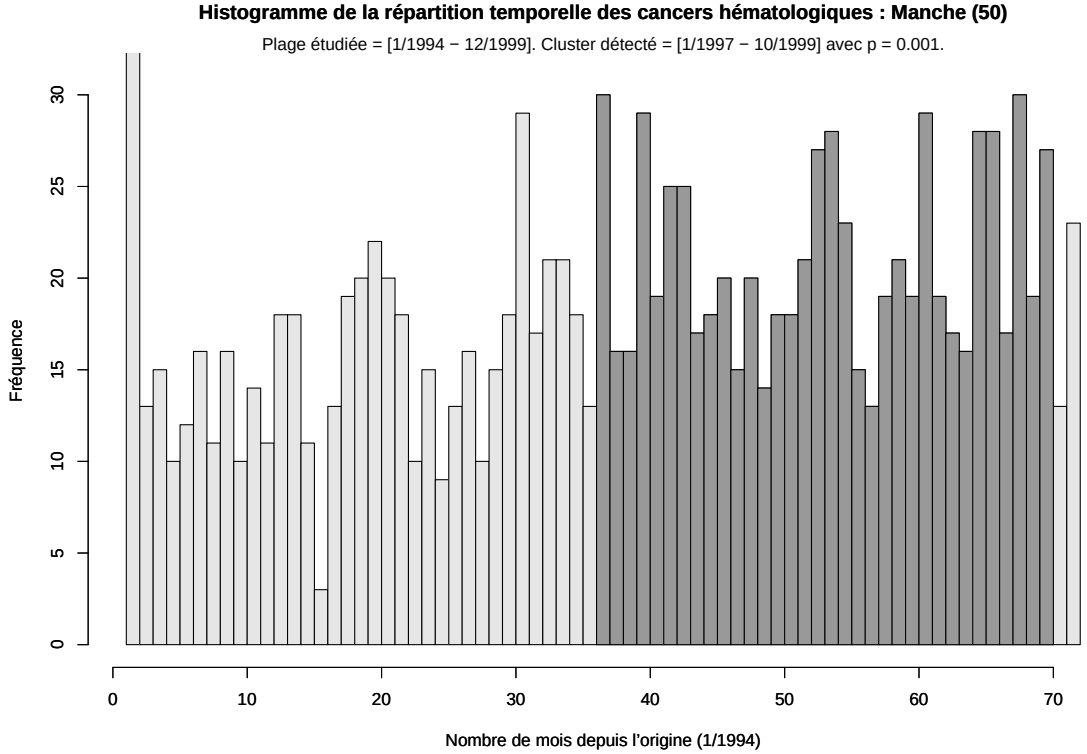
Les clusters significatifs sont repérés par des barres de couleurs plus foncées.

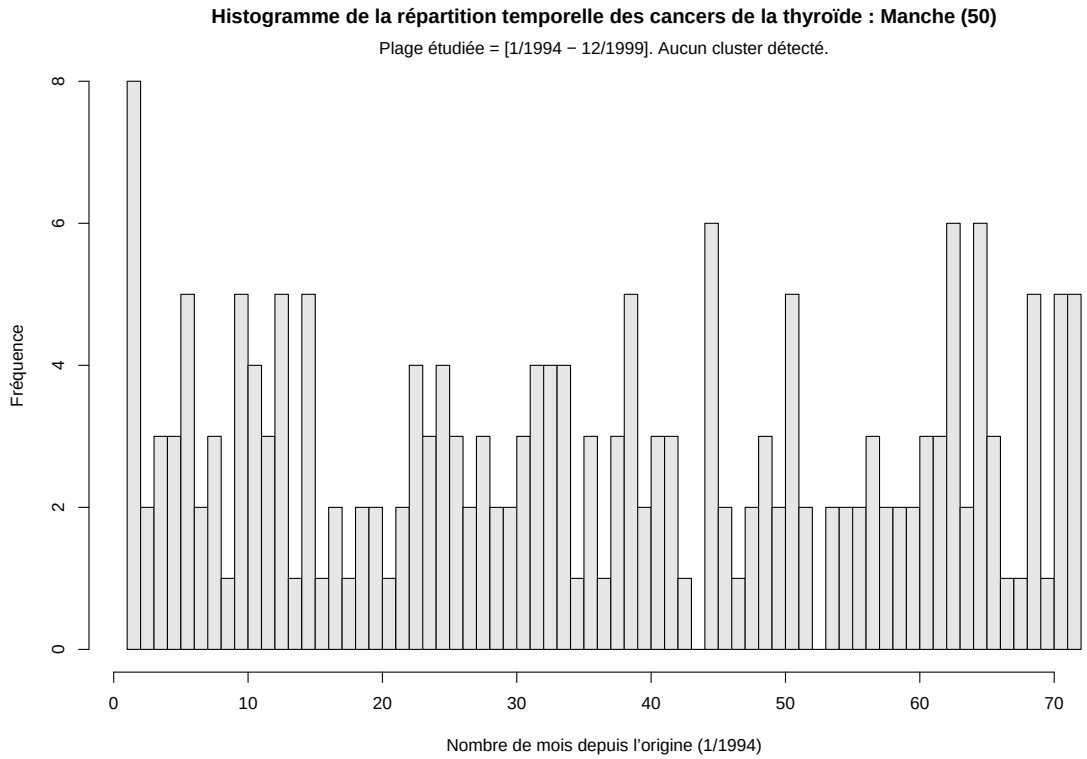



Histogramme de la répartition temporelle des cancers hématologiques : Tarn (81)
Plage étudiée = [1/1982 − 12/1999]. Cluster détecté = [3/1993 − 7/1999] avec p = 0.001.
Fréquence
0
5
10
15
20
25
0
50
100
150
200
Nombre de mois depuis l'origine (1/1982)


Histogramme de la répartition temporelle des cancers de la thyroïde : Tarn (81)
Plage étudiée = [4/1982 − 12/1999]. Cluster détecté = [3/1991 − 12/1999] avec p = 0.001.
Fréquence
0
2
4
6
8
0
50
100
150
200
Nombre de mois depuis l'origine (4/1982)


Histogramme de la répartition temporelle des cancers cérébraux : Manche (50)
Plage étudiée = [1/1994 − 12/1999]. Aucun cluster détecté.
Fréquence
0
1
2
3
4
5
6
0
10
20
30
40
50
60
70
Nombre de mois depuis l'origine (1/1994)


Histogramme de la répartition temporelle des cancers hématologiques : Manche (50)
Plage étudiée = [1/1994 − 12/1999]. Cluster détecté = [1/1997 − 10/1999] avec p = 0.001.
Fréquence
0
5
10
15
20
25
30
0
10
20
30
40
50
60
70
Nombre de mois depuis l'origine (1/1994)


Histogramme de la répartition temporelle des cancers de la thyroïde : Manche (50)
Plage étudiée = [1/1994 − 12/1999]. Aucun cluster détecté.
Fréquence
0
2
4
6
8
0
10
20
30
40
50
60
70
Nombre de mois depuis l'origine (1/1994)

B.2 Légende des graphiques spatiaux

B.2.1 Répartition de la population par commune

Chaque commune est entourée d'un disque gris dont le rayon est proportionnel à la taille de sa population. Les communes représentées par des points ont une population trop petite pour être représentées par un disque.

B.2.2 Visualisation de la vraisemblance par commune

Chaque commune ayant un nombre de cas supérieur au nombre de cas attendu sous H_0, est entourée d'un disque plein dont le rayon est proportionnel à la vraisemblance de la commune (calculée d'après le nombre de cas et la population de la commune). Le cluster détecté est délimité par un cercle en pointillés. Les communes localisées à l'intérieur du cluster détecté sont représentées en gris foncé (disque gris foncé si vraisemblance > 0, point gris foncé si vraisemblance $= 0$). Les communes à l'extérieur du cluster détecté sont en gris clair.

Il a été choisi de représenter la vraisemblance (incidence améliorée) plutôt que l'incidence afin de privilégier les communes importantes par la taille de leur population, et ainsi pouvoir visualiser la quantité qui est utilisée dans la détermination des clusters. Si on considère 2 communes ayant la même incidence de cancer (rapport nombre de cas sur population), la commune ayant le plus grand nombre de cas, et donc la population la plus importante, aura une vraisemblance plus élevée. Cela permet de donner plus de poids à une commune de 100000 habitants ayant eu 100 cas de cancers (incidence de 1/1000) qu'à une commune de 1000 habitants ayant eu 1 cas de cancer (même incidence).

Il peut arriver que des communes (dont l'emplacement est matérialisé par un point) se trouvent à la limite de la zone circulaire délimitant le cluster (représentée en pointillé). Ce cas de figure implique que le disque gris représentant la vraisemblance ne sera pas entièrement inclus dans le cercle en pointillé, voire même que le disque gris englobe entièrement le cercle en pointillé lorsque le rayon du cluster est petit et que la vraisemblance de certaines communes incluses dans le cluster est grande (cas des départements 34, 38 et 68 pour le cancer hématologique).

B.2.3 Visualisation des installations nucléaires

Les centrales nucléaires de St Alban, Flamanville et Fessenheim (départements 38, 50 et 68) ainsi que l'usine de retraitement de La Hague (département 50) sont représentées par un triangle.

B.3 Représentation de la population par commune

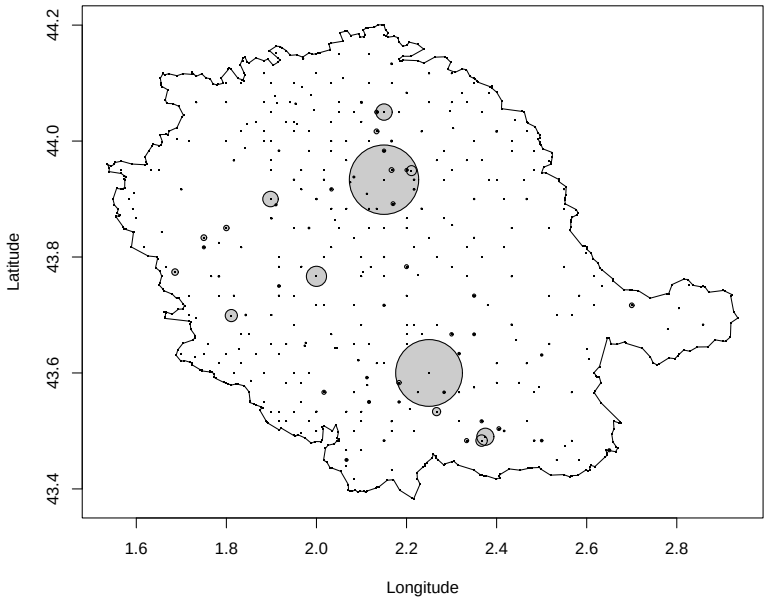


FIG. B.1 – Répartition de la population par commune dans le département du Tarn (81).

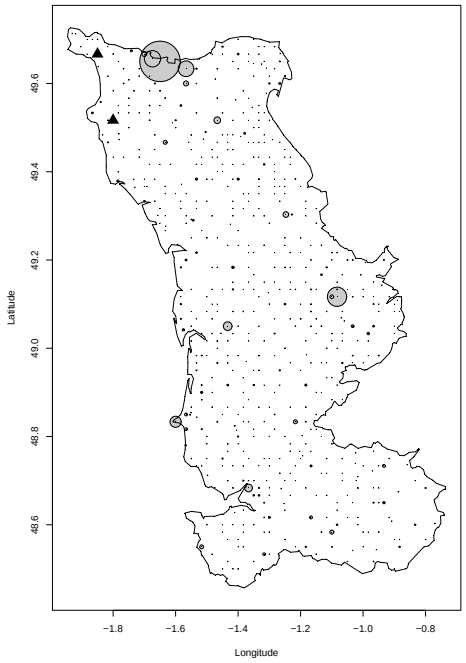


FIG. B.2 – Répartition de la population par commune dans le département de la Manche (50).

B.4 Visualisation de la vraisemblance par commune

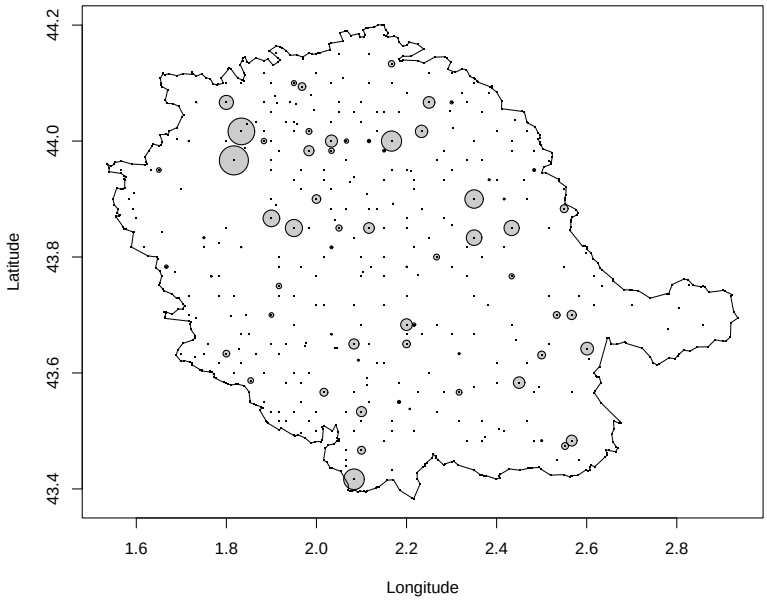


FIG. B.3 – Vraisemblance des cas de cancer du SNC par commune dans le département du Tarn (81).

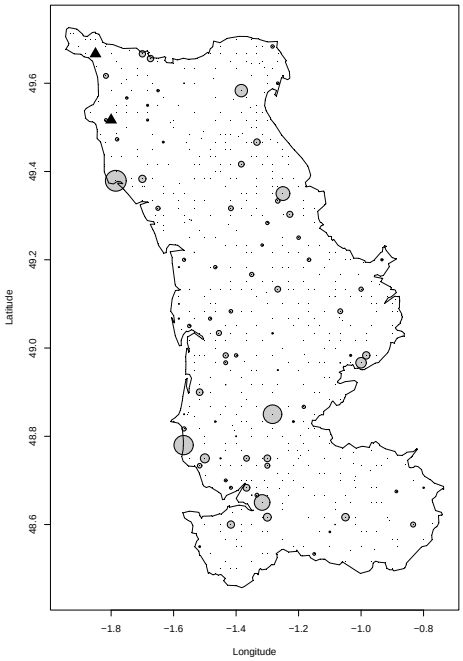


FIG. B.4 – Vraisemblance des cas de cancer du SNC par commune dans le département de la Manche (50).

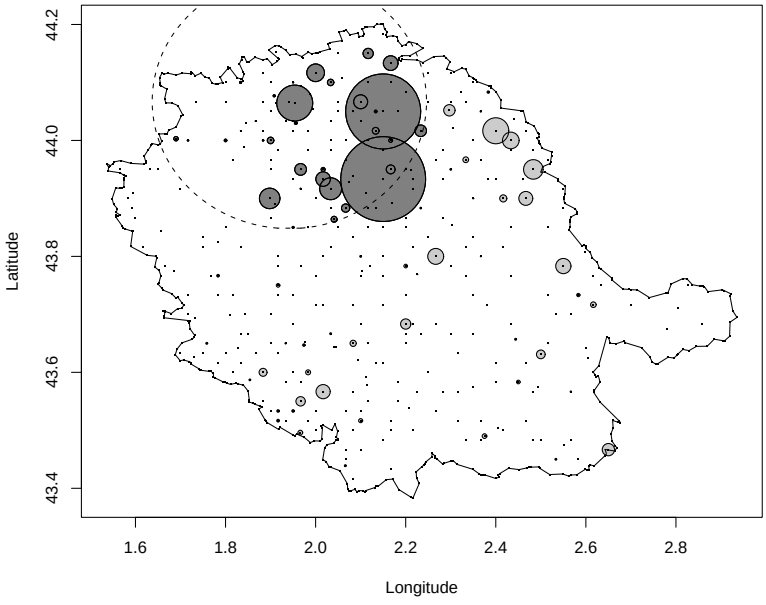


FIG. B.5 – Vraisemblance des cas de cancers hématologiques par commune dans le département du Tarn (81).

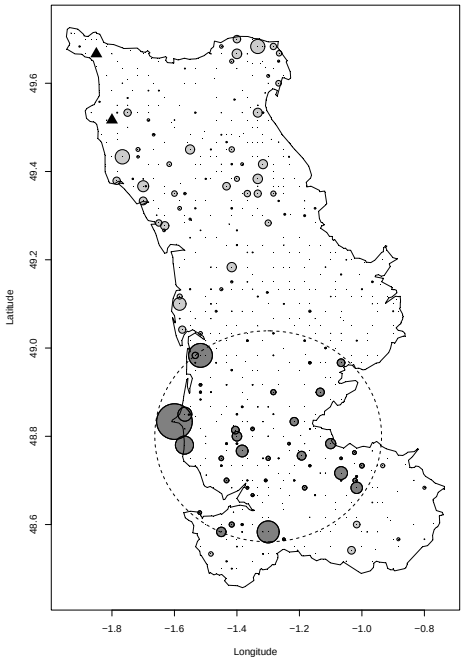


FIG. B.6 – Vraisemblance des cas de cancers hématologiques par commune dans le département de la Manche (50).

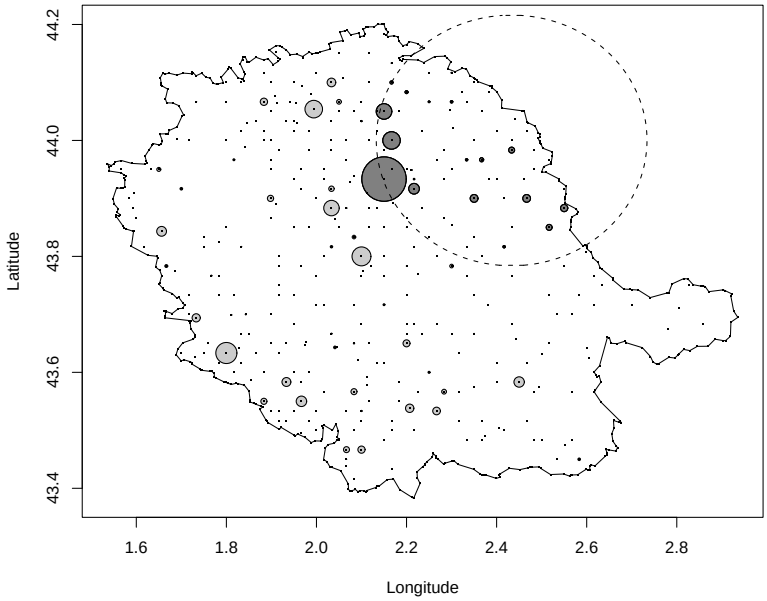


FIG. B.7 – Vraisemblance des cas de cancer de la thyroide par commune dans le département du Tarn (81).

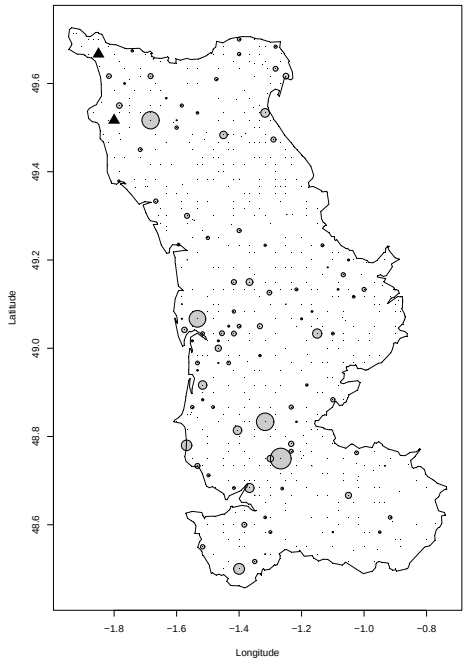


FIG. B.8 – Vraisemblance des cas de cancer de la thyroide par commune dans le département de la Manche (50).

B.5 Visualisation de la vraisemblance ajustée sur l'âge par commune

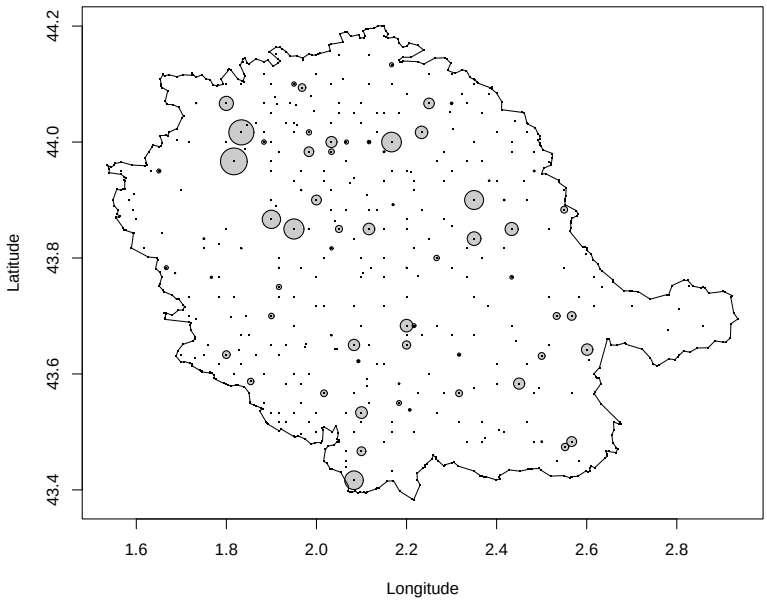


FIG. B.9 – Vraisemblance ajustée sur l'âge des cas de cancer du SNC par commune dans le département du Tarn (81).

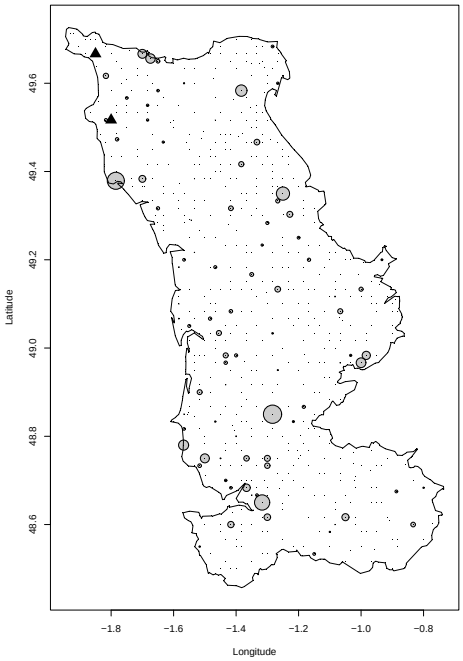


FIG. B.10 – Vraisemblance ajustée sur l'âge des cas de cancer du SNC par commune dans le département de la Manche (50).

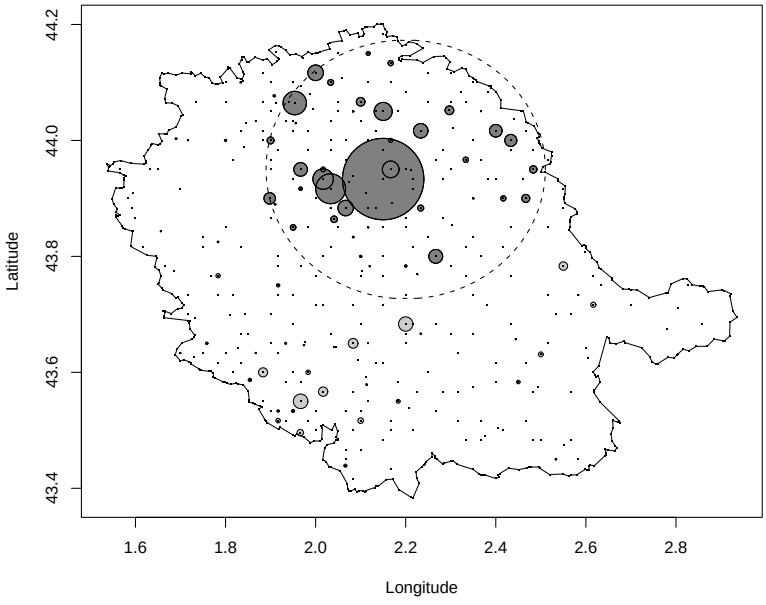


FIG. B.11 – Vraisemblance ajustée sur l'âge des cas de cancers hématologiques par commune dans le département du Tarn (81).

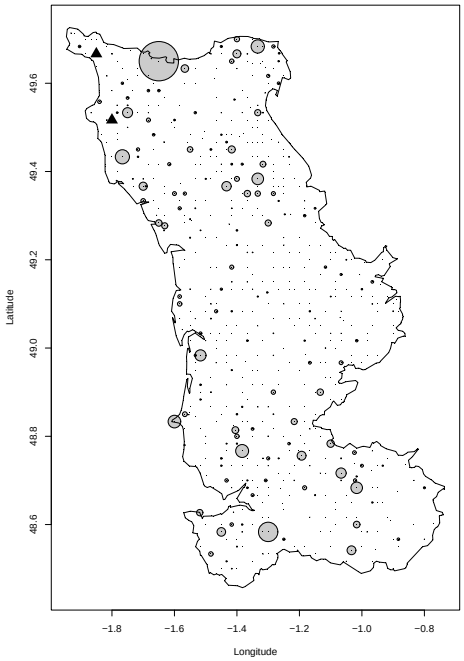


FIG. B.12 – Vraisemblance ajustée sur l'âge des cas de cancers hématologiques par commune dans le département de la Manche (50).

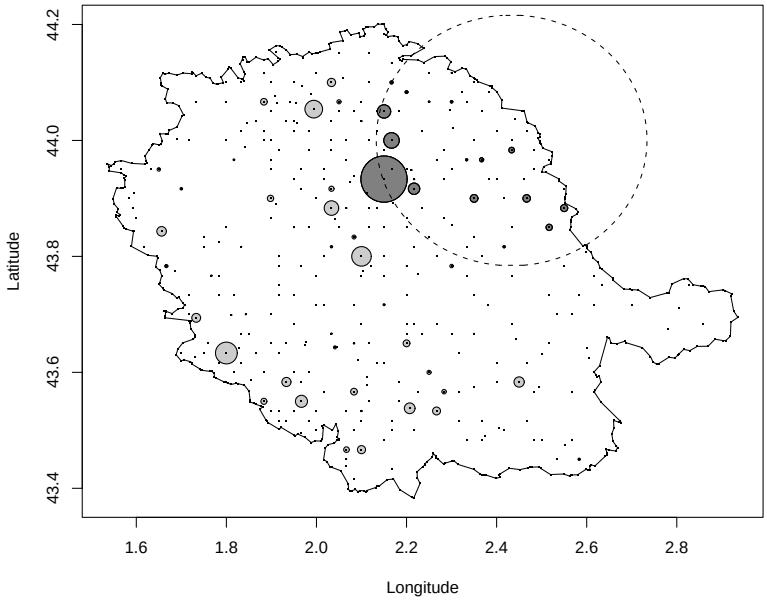


FIG. B.13 – Vraisemblance ajustée sur l'âge des cas de cancer de la thyroide par commune dans le département du Tarn (81).

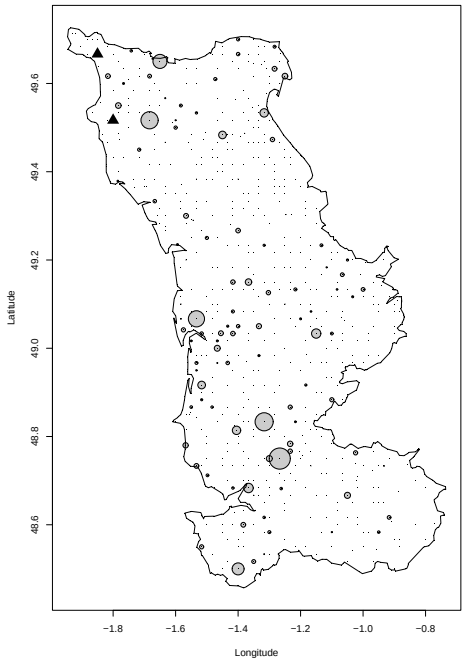


FIG. B.14 – Vraisemblance ajustée sur l'âge des cas de cancer de la thyroide par commune dans le département de la Manche (50).

Résumé

L'objectif de ce travail est de proposer des solutions nouvelles dans le domaine de la détection de clusters d'évènements de santé. Ce type d'analyse est traditionnellement utilisé dans la surveillance de maladies dont l'étiologie est incertaine afin de localiser et mettre en évidence des agrégats ayant une densité anormalement élevée dans le temps et/ou dans l'espace. La détermination de ces clusters constitue généralement une étape préliminaire à la recherche de facteurs de risque.

Nous proposons une revue des méthodes existantes ainsi que notre contribution dans différentes directions. Deux approches sont proposées dans le cadre temporel permettant pour l'une d'éviter l'utilisation de simulations et pour l'autre de prendre en compte les données dont l'information temporelle est incomplète. Nous avons également mis au point une méthode de détection de clusters spatiaux de forme arbitraire permettant d'analyser des données dont on connaît la localisation géographique exacte. Cette approche a été appliquée sur des données particulières, celles obtenues par Imagerie par Résonance Magnétique fonctionnelle. Les perspectives d'analyse spatio-temporelle sont finalement évoquées.

Mots clés : détection de clusters, agrégats spatiaux et/ou temporels, sélection de modèles, changements structurels multiples, données ponctuelles, évènements de santé, données IRMf.

Abstract

The aim of this work is to propose new solutions in the field of health event cluster detection. This type of analysis is traditionally used in the survey of diseases of which the aetiology is unknown in order to locate and detect aggregates which have an abnormally high density in time and/or in space. The determination of these clusters generally represents a preliminary stage in search of risk factors.

We propose a review of existing methods as well as our contribution in various directions. Two approaches are proposed in the temporal frame. The first allows to avoid the use of simulations. The second makes it possible to take into account data of which the temporal information is incomplete. We have also developped a method for the detection of arbitrarily shaped spatial clusters in order to be able to analyse data of which the exact geographic location is known. This approach has been applied on particular data, those obtained by functional Magnetic Resonance Imaging. The perspectives of spatio-temporal analysis are finally evoked.

Key words : cluster detection, spatial and/or temporal aggregates, model selection, multiple structural changes, case event data, health events, fMRI data.