ESPACEMENTS BIDIMENSIONNELS ET DONNÉES ENTACHÉES D'ERREURS DANS L'ANALYSE DES PROCESSUS PONCTUELS SPATIAUX

Lionel Cucala

THÈSE

présentée en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DES SCIENCES SOCIALES TOULOUSE I

Discipline : Mathématiques
Spécialité : Statistique

par

Lionel CUCALA

Espacements bidimensionnels et données entachées d'erreurs dans l'analyse des processus ponctuels spatiaux

---

Sous la direction de : Christine THOMAS-AGNAN

Soutenue le 8 décembre 2006 devant le jury composé de :

Avner BAR-HEN	Université Paris XIII	Rapporteur
Michel GOULARD	INRA Toulouse	Examinateur
Nicolas MOLINARI	Université Montpellier I	Examinateur
Jesper MÖLLER	Aalborg University	Rapporteur
Olivier PERRIN	Université Toulouse I	Examinateur
Christine THOMAS-AGNAN	Université Toulouse I	Directrice

REMERCIEMENTS

Je voudrais tout d'abord remercier Christine Thomas-Agnan pour avoir accepté d'encadrer ma thèse. Je lui suis très reconnaissant de la confiance qu'elle m'a accordée dès le début. Depuis trois ans, nos rendez-vous hebdomadaires ont vu naître de nombreuses idées mathématiques qui ont abondamment alimenté cette thèse et son soutien constant a été primordial.

Je tiens ensuite à remercier Avner Bar-Hen et Jesper Møller pour avoir accepté d'être les rapporteurs de cette thèse. Je suis très flatté de l'intérêt qu'ils ont porté à ce travail. Leur relecture attentive du manuscript ainsi que leurs remarques pertinentes ont contribué à améliorer la version finale de ce document.

Je suis très heureux que Nicolas Molinari ait accepté de faire partie de mon jury. La lecture de ces travaux récents sur la recherche d'agrégats m'a beaucoup inspiré et j'espère avoir l'occasion dans un futur proche de collaborer avec lui.

Je souhaite également remercier Michel Goulard et Olivier Perrin de faire partie de mon jury. J'ai eu notamment l'occasion de les côtoyer avec plaisir lors de certains congrès et leurs remarques concernant mes travaux ont souvent été pertinentes.

Un grand merci également à Noel Cressie pour m'avoir accueilli dans son laboratoire à l'Ohio State University. Son cours sur les statistiques spatiales m'a beaucoup apporté et la genèse de la dernière partie de cette thèse lui est directement imputable. Merci à tous ceux qui m'ont permis de supporter plus facilement le rude climat de l'Ohio en hiver, que ce soit les professeurs Ernest Fokoué, Catherine Calder et Hailen Tsing, mes camarades Jim, Hongfei

et Xiaobai, mon colocataire Clint et bien sûr mon compatriote Nicolas dont la présence à mes côtés a été un plaisir constant.

Je voudrais à présent remercier les professeurs du LSP et du GREMAQ que j'ai pu côtoyer pendant ces trois années de thèse et auparavant, notamment Fabrice Gamboa, Philippe Vieu et Frédéric Ferraty qui ont toujours répondu à mes questions avec patience et Anne Ruiz-Gazen pour sa gentillesse et sa franchise.

Ces trois années de thèse m'ont également permis de rencontrer des doctorants avec qui je passe de très bons moments. Les doctorants arrivés l'an dernier, Maxime, Laurent, Florent et Amélie, ont apporté leur bonne humeur pendant la pause quizz de midi. Mes remerciements vont aussi aux doctorants arrivés en thèse la même année que moi ou l'année suivante, qui vont me laisser de très bons souvenirs : Delphine (avec qui ça a été un plaisir de partager le bureau ces deux dernières années), Marielle, Agnès, Solenn, Myriam, Diana, Simon et Arnaud. Je ne saurais oublier les doctorants qui m'ont accueilli à mon arrivée en thèse, et tous les bons moments qu'on a passés : Renaud, Clément, Cécile, Yan, Élie, Abdel, Nicolas et Jean-Pierre. Je souhaite aussi remercier Sébastien, à qui j'ai posé d'innombrables questions sur LaTeX, sur **R**, et je retiens avant tout sa disponibilité et sa bonne humeur. Enfin, je connais Christophe depuis le DEA et on partage le même bureau depuis le début de notre thèse, et je tiens à lui dire quel plaisir j'ai eu de pouvoir faire ma thèse en même temps que lui, pour tous les bons moments passés pendant ces années. Merci d'avoir répondu à toutes les questions, même les moins pertinentes, avec patience. Merci aussi de m'avoir fait découvrir le joli village de Vira.

Les occasions de revoir mes anciens camarades de l'INSA ou assimilés sont toujours un plaisir : merci aux Vals, aux Fabs et à Cyril ainsi qu'à ceux que j'aimerais voir plus souvent, Mohamad et Dala.

Merci également à tous les membres du Castres Football Club que j'ai toujours beaucoup de plaisir à retrouver et que j'espère accompagner dans la victoire comme dans la défaite pendant un maximum de temps.

Enfin, je voudrais remercier toute ma famille, plus particulièrement mes parents qui m'ont toujours soutenu dans les études et qui m'ont permis de les réaliser dans les meilleures conditions possibles, et Harold pour m'avoir permis d'exercer mes talents culinaires chaque semaine. Enfin, je remercie Séverine pour tout ce qu'elle m'apporte.

TABLE DES MATIÈRES

INTRODUCTION

0.1. Les processus ponctuels spatiaux

0.1.1. Présentation. — Dans de nombreux domaines d'application, les jeux de données à analyser se présentent sous la forme de listes d'événements localisés dans l'espace. Ces jeux de données sont appelés semis de points et peuvent correspondre à la disposition de certaines espèces végétales dans une forêt, aux emplacements des épicentres de secousses sismiques enregistrées, à la localisation de trésors archéologiques retrouvés sur un site, aux adresses de patients affectés d'une certaine maladie dans une région, à la répartition de cellules dans un tissu biologique...

Un tel semis de points est noté $\{s_i, i = 1, \cdots, n\}$, où $s_i \in X \subsetneqq \mathbb{R}^d, \forall i = 1, \cdots, n$, X étant borné. Afin de l'analyser, on considérera généralement qu'il s'agit d'une réalisation d'un processus stochastique appelé processus ponctuel spatial : le nombre total d'événements N est aléatoire, ainsi que les variables de localisation $S_i \in X \subsetneqq \mathbb{R}^d$. Chaque localisation peut être adjointe d'un ensemble de variables de différents types (réelles, entières, booléennes...) appelé "marque". Lorsque c'est le cas, le processus ponctuel est dit marqué. Par la suite, nous n'envisagerons que des processus ponctuels non marqués.

Pour un tel modèle de processus ponctuel spatial, notons $\mathcal{X}$ la tribu des boréliens bornés de X et M la mesure aléatoire de comptage qui, à tout élément B de $\mathcal{X}$, associe $M(B)$ le nombre d'événements contenus dans B. Un processus ponctuel spatial est caractérisé de manière unique par les probabilités d'évitement définies par $\mathbb{P}[M(B) = 0], \forall B \in \mathcal{X}$.

On peut également définir une densité jointe f qui associe $f\big((s_1, \cdots, s_n), n\big)$ à chaque réalisation du processus ponctuel (Cressie, 1993, p.622). On a alors

$$\sum_{n=0}^{\infty} \int_{X^n} f\big((s_1, \cdots, s_n), n\big) \nu(ds_1) \cdots \nu(ds_n) = 1$$

où $\nu(.)$ est la mesure de Lebesgue.

0.1.2. Quelques notions importantes. — Afin de caractériser les différents modèles de processus ponctuels, il est nécessaire de définir quelques notions.

Introduisons d'abord la mesure du moment d'ordre 1 de M :

$$\forall B \in \mathcal{X},\ \mu_M(B) = \mathbb{E}[M(B)].$$

De même, on peut définir la mesure du moment d'ordre k de M :

$$\forall (B_1, \cdots, B_k) \in \mathcal{X}^k,\ \mu_M^{(k)}(B_1 \times \cdots \times B_k) = \mathbb{E}[M(B_1) \cdots M(B_k)].$$

Considérons maintenant ces mesures de moment lorsque l'on prend pour B un singleton $\{s\}$, $s \in X$. Notons ds et du des volumes infinitésimaux centrés respectivement en s et u.

L'intensité (de premier ordre) de M est définie par :

$$\forall s \in X,\ \lambda(s) = \lim_{\nu(ds)\to 0} \mu_M(ds)/\nu(ds)$$

et s'interprète comme le nombre moyen d'événements par unité de volume au point s.

L'intensité de second ordre de M est définie par :

$$\forall (s,u) \in X^2,\ \lambda_2(s,u) = \lim_{\nu(ds)\to 0, \nu(du)\to 0} \frac{\mu_M^{(2)}(ds \times du)}{\nu(ds)\nu(du)}$$

et s'interprète comme le nombre moyen de couples d'événements par unité de volume au couple de points $\{s,u\}$.

Un processus ponctuel sera dit stationnaire si sa mesure de comptage M est invariante par translation. On a alors $\lambda(s) = \lambda$, $\forall s \in X$ et il existe une fonction $\lambda_2^\star$ telle que $\lambda_2(s,u) = \lambda_2^\star(s-u)$, $\forall (s,u) \in X^2$. Le processus sera également dit isotrope si sa mesure de comptage est invariante par rotation : il existe alors une fonction $\lambda_2^{\star\star}$ telle que $\lambda_2(s,u) = \lambda_2^{\star\star}(||s-u||)$, $\forall (s,u) \in X^2$, $||.||$ symbolisant la norme euclidienne.

0.1.3. La "Complete Spatial Randomness" (CSR). — De très nombreux modèles de processus ponctuels ont été introduits. Le modèle de base, traduisant la notion d'homogénéité spatiale (traduction de "complete spatial randomness") est le processus de Poisson homogène, appelé ainsi car le nombre d'événements contenus dans $A \subset X$ est une variable aléatoire de Poisson de paramètre $\lambda\nu(A)$. On a $f\big((s_1, \cdots, s_n), n\big) = \frac{e^{-\lambda}\lambda^n}{n!}\frac{1}{\nu(X)^n}$. Par conséquent $\frac{f\left((s_1,\cdots,s_n),n\right)}{\int_{X^n} f\left((s_1,\cdots,s_n),n\right)\nu(ds_1)\cdots\nu(ds_n)} = \frac{1}{\nu(X)^n}$ et donc, sachant N, les localisations S_i sont indépendantes et suivent une loi uniforme sur X.

Ce modèle sert de standard pour ce qu'on entend intuitivement comme une répartition aléatoire de points, c'est pourquoi le problème de tester l'adéquation

d'un jeu de données avec ce modèle est généralement la première étape importante de l'analyse d'un semis de points.

0.1.4. Les hypothèses alternatives. — Nous présentons ici quelques-unes des alternatives à l'homogénéité spatiale étudiées par la suite mais en aucun cas cette liste n'est exhaustive.

0.1.4.1. Les processus de Poisson inhomogènes. — Cette classe des processus de Poisson sert à modéliser tous les processus ponctuels dont les événements sont localisés de manière indépendante. Ils sont entièrement caractérisés par leur fonction intensité λ sur X et satisfont les deux propositions suivantes :

•$\forall B_1, \cdots, B_m$ t.q $B_i \subsetneq X$ et $B_i \cap B_j = \emptyset$ si $i \neq j$, les variables aléatoires

$M(B_1), \cdots, M(B_m)$ sont indépendantes et de loi de Poisson d'espérances

respectives $\int_{B_1} \lambda(s)\nu(ds), \cdots, \int_{B_m} \lambda(s)\nu(ds)$.

• Conditionnellement à $M(X) = n$, les localisations $(s_1, \cdots, s_n)$ sont

indépendamment distribuées selon la densité d-dimensionnelle $\frac{\lambda(s)}{\int_X \lambda(s)\nu(ds)}$

sur X.

0.1.4.2. Les processus de Cox. — Cette famille de processus sert généralement à modéliser des processus dont les événements ont tendance à s'agréger. Le processus de Cox se construit en deux étapes incluant de l'aléa et est pour cette raison souvent appelé "processus doublement stochastique". Les deux étapes sont les suivantes :

- On définit un champ aléatoire Λ sur X, à valeurs dans $\mathbb{R}^+$. λ est une réalisation de ce champ aléatoire.
- On construit un processus de Poisson de fonction intensité λ.

Le processus de Cox est donc un processus de Poisson dont la fonction intensité est aléatoire. Il existe de nombreuses familles de processus de Cox suivant la loi du champ aléatoire Λ. On peut citer par exemple le processus de Neyman-Scott, pour lequel la fonction intensité est proportionnelle à l'estimateur à noyau de densité multidimensionnelle appliqué aux événements d'un processus de Poisson homogène sur X. Un tel processus reproduit l'émergence

d'événements autour de certaines sources (les réalisations du processus de Poisson homogène).

Il faut remarquer que lorsqu'on dispose d'une seule réalisation d'un processus ponctuel, il est impossible de comparer l' hypothèse d'un processus de Poisson inhomogène à l'hypothèse d'un processus de Cox. En effet, rien ne permet de savoir si la fonction intensité d'un éventuel processus de Poisson ayant généré l'échantillon est de nature déterministe ou la réalisation d'un champ aléatoire sous-jacent.

0.1.4.3. Les processus d'inhibition. — Cette famille de processus sert à modéliser des processus dont les événements ont tendance à se repousser. L'exemple le plus trivial de processus d'inhibition est le processus dit "hardcore", qui ne permet pas à deux événements d'être distants d'une longueur inférieure à r (Matérn, 1960).

0.1.4.4. Les processus de Markov. — Cette famille de processus sert généralement à modéliser des processus dont les événements ont tendance à se repousser, mais peut également être utilisée pour modéliser l'agrégation. Un processus est dit Markov de rayon r si l'intensité en un point $s \in X$, conditionnellement à la réalisation du processus sur $X \setminus s$ ne dépend que des événements compris dans la boule centrée en s et de rayon r.

L'exemple le plus trivial de processus de Markov est le processus de Strauss (1975), dont la densité jointe $f\big((s_1, \cdots, s_n), n\big)$ dépend uniquement du nombre de couples d'événements "voisins"

$$\sum_{1 \leq i < j \leq n} \mathbb{1}(||s_i - s_j|| \leq r).$$

0.2. Les tests d'homogénéité spatiale

Pour tester l'hypothèse CSR, il existe de nombreuses techniques que nous pouvons découper en différentes familles.

0.2.1. Les statistiques basées sur les quadrats. — Historiquement les plus anciennes, ces statistiques résultent du découpage du domaine d'observation D en sous-régions disjointes appelées quadrats, $A_1, \cdots, A_q$. On note

$N_1, \cdots, N_q$ les variables aléatoires représentant les nombres d'événements contenus dans les quadrats. Sous l'hypothèse d'homogénéité spatiale, ces variables aléatoires sont indépendantes et $N_i \sim \mathcal{P}\big(\lambda\nu(A_i)\big), \forall i = 1, \cdots, q$. Cette hypothèse est alors testée par le biais d'un test de Khi-deux. Notons que l'on obtient des tests différents selon le choix préalable de l'ensemble des quadrats.

0.2.2. Les statistiques basées sur les distances. — Elles s'appuient sur les distances entre événements ou sur les distances entre points du domaine d'observation et événements. En voici quelques exemples.

• Distance moyenne au plus proche voisin :

Notons $Z_i, i = 1 \cdots N$, les variables aléatoires représentant les distances de chaque événement à son plus proche voisin (i.e. l'événement le plus proche en distance euclidienne), et $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$. On pose

$$T = \frac{\bar{Z} - \mu_{\bar{Z}}}{\sigma_{\bar{Z}}}$$

où $\mu_{\bar{Z}} = 0.5N^{-1/2} + 0.206N^{-1} + 0.164N^{-3/2}$,

et $\sigma_{\bar{Z}} = 0.07N^{-2} + 0.148N^{-5/2}$.

Sous l'hypothèse d'homogénéité spatiale, T suit approximativement une loi normale standard (Donnelly, 1978).

• Fonction de répartition empirique des distances au plus proche voisin :

Notons $\hat{G}_1(z)$ la fonction de répartition empirique (FRE) de $(z_1, \cdots, z_n)$ et $\hat{G}_i(z), i = 2, \cdots, 100$ la FRE des distances au plus proche voisin issue de la ième simulation d'un processus de Poisson homogène sur D.

On pose , $\forall i = 1, \cdots, 100$:

$$\bar{G}_i(z) = \frac{1}{99}\sum_{j \neq i} \hat{G}_j(z)$$

On peut alors définir la statistique associée à chacun des semis de points (réel ou simulé) :

$$U_i = \int [\hat{G}_i(z) - \bar{G}_i(z)]^2 dz$$

où le domaine d'intégration est choisi par l'utilisateur.

On comparera alors U_1, issue du semis réel, à $U_2, \cdots, U_{100}$, issues des semis simulés (Diggle, 1979).

• Fonction de répartition empirique des distances point-événement :

La technique est la même que précédemment sauf que l'on s'appuie sur $w_j, j = 1 \cdots p$, les distances de p points aléatoires à l'événement le plus proche (Diggle et Matérn, 1980) et la fonction de répartition empirique $\hat{H}(.)$ associée.

On comparera V_1 à $V_2, \cdots, V_{100}$, où

$$V_i = \int [\hat{H}_i(w) - \bar{H}_i(w)]^2 dw.$$

0.2.3. Les statistiques hybrides. — Elles s'appuient sur les coordonnées des événements ou sur l'ensemble des distances entre chaque couple d'événements.

• La statistique de Liebetrau (1977) est dérivée de l'estimation de la fonction de covariance de la mesure de comptage associée à un processus de Poisson homogène. Elle s'écrit :

$$Li = \sum_{j=1}^{n-1} \sum_{k=1}^{n-j} [(t_0 - |u_j - u_{j+k}|)_+ (t_0 - |v_j - v_{j+k}|)_+]^2$$

où $h_+ = max(h, 0)$ et $\{(u_j, v_j), j = 1, \cdots, n\}$ sont les coordonnées des événements et $|.|$ représente la valeur absolue.

Ici, la constante t_0 est une borne supérieure de l'échelle à laquelle est analysé le semis de points : deux événements dont les abscisses ou les ordonnées sont trop éloignées ne seront pas pris en compte. On cherche donc à repérer les phénomènes d'attraction ou de répulsion d'une portée limitée.

• La fonction K de Ripley (1976) représente l'espérance du nombre d'événements situés dans un rayon de t d'un événement pris au hasard.

Sous l'hypothèse d'homogénéité spatiale, $K(t) = \pi t^2$.

On utilisera alors la statistique :

$$L_m = \sup_{t \leq t_0} |\sqrt{\hat{K}(t)/\pi} - t|,$$

où $\hat{K}(t)$ est l'estimateur empirique de $K(t)$ défini par

$$\hat{K}(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{1}(||s_i - s_j|| < t)$$

et t_0 est définie comme précédemment. Par simplicité, l'estimateur présenté ici ne contient pas de terme de correction de bord.

0.2.4. Les statistiques basées sur la fonction de répartition empirique des points. — On considère dans ce paragraphe des processus ponctuels dont le domaine d'observation est le carré unitaire $[0, 1]^2$.

La statistique de Cramer-Von Mises est étendue à la 2D par Zimmerman (1993) : son expression est

$$\bar{\omega}^2 = \frac{1}{4n} \sum_{i=1}^{n} \sum_{j=1}^{n} (1-|u_i-u_j|)(1-|v_i-v_j|) - \frac{1}{2} \sum_{i=1}^{n} (u_i^2-u_i-1/2)(v_i^2-v_i-1/2) + \frac{n}{9}$$

et sa loi limite est identifiée par Zimmerman (1994) par une méthode de décomposition en composantes principales.

La statistique de Kolmogorov-Smirnov est également étendue à la 2D par Justel & *al.* (1997). Les calculs s'avérant problématiques, ces mêmes auteurs proposent une version simplifiée dont les performances restent semblables. Les distributions des deux statistiques sont estimées par une procédure de type Monte-Carlo.

0.2.5. Les statistiques basées sur le diagramme de Voronoi et la triangulation de Delaunay. — Ces statistiques s'appuient sur le découpage du domaine d'observation D en cellules de Voronoi C_i définies par $C_i = \{x \in X, \forall j \neq i \quad ||x - s_j|| > ||x - s_i||\}$. La triangulation de Delaunay est celle reliant les couples de points dont les cellules de Voronoi sont voisines. Sous l'hypothèse d'homogénéité spatiale, il est possible d'exprimer la distribution de certaines caractéristiques des cellules de Voronoi ou des triangles de Delaunay, et Chiu (2003) estime la puissance des tests basés sur ces caractéristiques.

0.2.6. Les statistiques basées sur le périodogramme. — Pour les fréquences (ω_p, ω_q), le périodogramme associé au semis de points $\{(u_i, v_i), i = 1 \cdots n\}$ est défini par :

$$\hat{f}(\omega_p, \omega_q) = \Big\{ \sum_{i=1}^{n} \cos\{n(\omega_p u_i + \omega_q v_i)\} \Big\}^2 + \Big\{ \sum_{i=1}^{n} \sin\{n(\omega_p u_i + \omega_q v_i)\} \Big\}^2.$$

Les premières analyses de processus ponctuels à l'aide du périodogramme sont dues à Bartlett (1964) et Mugglestone et Renshaw (1996, 2001) ont élaboré et étudié de nombreux tests de la CSR basés sur le périodogramme.

Conditionnellement au nombre d'événements du processus ponctuel, tester l'hypothèse d'homogénéité spatiale correspond à un test d'adéquation à la loi uniforme (Moller & Waagepetersen, 2004). Lorsque l'on travaille en dimension 1, les grandes familles de tests d'adéquation sont les tests de type Chi-deux, les tests basés sur la fonction de répartition empirique et les tests basés sur les espacements. Intéressons-nous plus particulièrement à cette dernière catégorie, qui est la seule à n'avoir pas été jusqu'à maintenant étendue aux dimensions supérieures.

0.3. La théorie des espacements en 1D

En dimension 1, les espacements représentent les longueurs des intervalles entre observations successives. On dispose d'un échantillon $(X_1, \cdots, X_{n-1})$ classé par ordre croissant : $X_{(1)} < \cdots < X_{(n-1)}$. Le vecteur des espacements

est alors $(D_2, \cdots, D_{n-1})$ où

$$D_i = X_{(i)} - X_{(i-1)}, \quad i = 2, \cdots, n.$$

Lorsque la distribution de l'échantillon initial est de support connu $[a, b]$, on peut y adjoindre les espacements $D_1 = X_{(1)} - a$ et $D_n = b - X_{(n-1)}$.

De nombreux auteurs ont suggéré l'utilisation de ces variables aléatoires dans des tests d'adéquation mais le premier à avoir formalisé celà est Greenwood (1946) qui propose de tester l'adéquation d'un échantillon à la loi uniforme en observant la variance empirique des espacements. De nombreuses autres statistiques basées sur les espacements uniformes, i.e. issus d'un échantillon uniforme $[0, 1]$, ont ensuite été proposées (Pyke, 1965). Une grande partie de ces statistiques peut s'écrire sous la forme $S_n = \sum_{i=1}^{n} g(D_i)$, où g est une fonction mesurable. Les autres statistiques s'appuient sur le classement de ces espacements par ordre croissant (Barton & David, 1956). Il est à noter que grâce à la méthode d'inversion de la fonction de répartition, ces techniques, bien que basées sur les échantillons uniformes, permettent de tester l'adéquation à n'importe quelle loi.

La question de la distribution de ces statistiques sous l'hypothèse d'uniformité est rendue compliquée par la dépendance entre les espacements (leur somme étant égale à la longueur de l'intervalle, à savoir 1) et seules les lois limites sont généralement identifiées. Celles-ci furent dans un premier temps étudiées au cas pas cas, puis Darling (1953) et Le Cam (1958) fournirent une démonstration générale pour les statistiques de type S_n en s'appuyant sur les fonctions caractéristiques. Récemment, le théorème de Le Cam a été étendu aux statistiques basées sur les espacements d'ordre m (sommes de m espacements consécutifs) par Beirlant & *al.* (1991), selon une méthode beaucoup plus directe de décomposition. Quelques études empiriques montrent la puissance des tests basés sur les espacements, notamment contre des alternatives de dépendance (Deheuvels, 1983), même si, d'un point de vue théorique, elles ont peu de puissance contre des alternatives contigües (Guttorp & Lockhart, 1988).

0.4. La théorie des espacements en 2D

Ce paragraphe résume les résultats énoncés dans la partie 1.

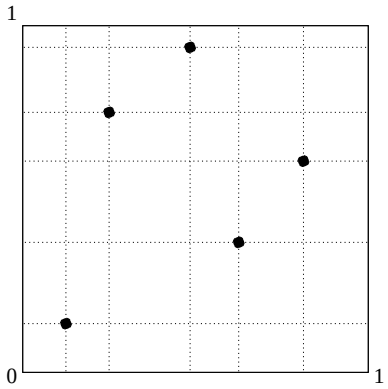

FIGURE 1. Les espacements bidimensionnels

0.4.1. Définition des espacements bidimensionnels. — Placons-nous dans le cadre d'un semis de points $\{(u_i, v_i), i = 1, \cdots, n-1\}$ sur le carré unitaire $[0,1]^2$. L'extension de la notion d'espacements à ce cas bidimemsionnel se comprend aisément d'un point de vue géométrique. En dimension 1, les espacements sont les longueurs des segments générés par le découpage selon les observations, en dimension 2 ce seront les aires des surfaces générées par le découpage selon les observations. Celà est illustré par la figure 1.

Ces espacements seront notés $\{A_{i,j}, i = 1, \cdots, n, j = 1, \cdots, n\}$ et on a la relation suivante

$$\forall i = 1, \cdots, n, \quad \forall j = 1, \cdots, n, \quad A_{i,j} = D_i^u D_j^v$$

où $\{D_i^u, i = 1, \cdots, n\}$ sont les espacements générés par les abscisses et $\{D_j^v, j = 1, \cdots, n\}$ les espacements générés par les ordonnées.

Dans un premier temps, il était également envisagé de ne conserver que les espacements correspondant à des aires dont un des coins est un événement, par exemple toutes les aires dont le coin bas-gauche est un événement. Ces espacements bidimensionnels "corniers", i.e. relatifs à un coin, sont définis de la manière suivante.

$u_{(1)} \leq \cdots \leq u_{(n-1)}$ sont les statistiques d'ordre tirées de $(u_1, \cdots, u_{n-1})$.

$v_{(1)} \leq \cdots \leq v_{(n-1)}$ sont les statistiques d'ordre tirées de $(v_1, \cdots, v_{n-1})$.

$u_0 = v_0 = u_{(0)} = v_{(0)} = 0$

$u_n = v_n = u_{(n)} = v_{(n)} = 1$

Introduisons les permutations c et r de $\{0, \cdots, n-1\}$ telles que :

$$u_{(c(i))} = u_i \text{ et } v_{(r(i))} = v_i$$

On peut alors définir les espacements unidimensionnels :

$$\forall i \in \{1, \cdots, n\}, \quad D_i^u = u_{(c(i))} - u_{(c(i)-1)}$$

$$\forall i \in \{1, \cdots, n\}, \quad D_i^v = v_{(c(i))} - v_{(c(i)-1)}$$

Les espacements corniers seront alors les aires A_i définies par :

$$\forall i \in \{1, \cdots, n\}, \quad A_i = D_i^u D_i^v.$$

En effet, la correspondance entre abscisses et ordonnées est totalement perdue dans le processus de création des espacements $A_{i,j}$, ce qui n'est pas le cas pour les espacements corniers A_i. Cependant, dans la pratique, les tests basés sur ces espacements corniers se révèlent beaucoup moins puissants que ceux basés sur les espacements $A_{i,j}$ et nous préférons régler ce problème de correspondance par une procédure multiple, explicitée ultérieurement.

0.4.2. Les statistiques basées sur les espacements bidimensionnels et leurs lois. — Les espacements bidimensionnels ayant été introduits, on peut trouver un équivalent à toutes les statistiques utilisées en dimension 1. L'équivalent de la statistique S_n sera par exemple $V_n = \sum_{i=1}^n \sum_{j=1}^n g(A_{i,j})$.

Là encore, la loi de la statistique V_n est impossible à obtenir à cause de la structure de dépendance entre espacements. On cherche alors à identifier la loi asymptotique et nous y parvenons en adaptant la technique de décomposition de Beirlant & *al.* (1991).

Notons $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_n)$ deux échantillons i.i.d. suivant une loi exponentielle de moyenne 1, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. On suppose que la fonction g satisfait certaines hypothèses décrites dans la partie 1.

Notons

$$\begin{aligned}
\mu &= \mathbb{E}[g(X_1Y_1)],\\
\eta &= Cov\big(g(X_1Y_1), g(X_1Y_2)\big) = Cov\big(g(X_1Y_1), g(X_2Y_1)\big),\\
c &= Cov\big(g(X_1Y_1), X_1\big) = Cov\big(g(X_1Y_1), Y_1\big),\\
\sigma^2 &= 2(\eta - c^2).
\end{aligned}$$

Théorème 1. — *Sous certaines hypothèses sur g, on a*

$$\frac{1}{n^{3/2}}\{V_n - n^2\mu\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2).$$

D'un point de vue pratique, nous mettons en place une procédure multiple permettant de régler le problème de correspondance précédemment évoqué et applicable pour n'importe quel domaine d'observation D (et non plus uniquement le carré unitaire). Cette procédure est appliquée à de nombreux jeux de données réels ou simulés (suivant une des alternatives précédemment décrites) et il apparaît que les tests basés sur les espacements bidimensionnels sont plus puissants que les autres pour identifier certaines formes d'hétérogénéité.

0.5. La détection d'agrégats

Ce paragraphe résume les résultats énoncés dans la partie 2.

Lorsqu'un semis de points ne satisfait pas à l'hypothèse d'homogénéité spatiale et semble exhiber une certaine agrégation, il est parfois utile de délimiter les zones de plus forte intensité, que nous nommerons agrégats. Cette notion d'agrégat, sur laquelle beaucoup se sont penchées, n'a rien de strictement mathématique et est sujette à interprétation. On s'intéresse aux méthodes, dites spécifiques, qui cherchent à localiser le ou les agrégats éventuels. Nous ne parlerons pas des tests dits "focus" qui cherchent à déterminer la présence ou non d'un agrégat autout d'un point spécifié à l'avance, comme par exemple une mortalité plus importante autour d'une source de pollution (Lawson, 1993).

0.5.1. La dimension 1. — Supposons que l'on observe des événements de même type qui se produisent sur un intervalle de temps donné, comme par

exemple la détection de cas d'une maladie spécifique. Par une simple transformation des données, on se ramène à l'observation d'événements localisés en $\{x_1, \cdots, x_n\}$ sur le segment unitaire $[0, 1]$. Un agrégat est généralement décrit comme une zone délimitée dans laquelle se concentrent un nombre anormalement élevé d'événements. Reste à définir un indice ou une mesure de concentration, et à établir où se situe sa limite de "normalité".

Lorsque deux intervalles sont de même longueur, celui le plus à même d'être un agrégat sera bien sûr celui qui contient le plus grand nombre d'événements mais les choses se compliquent lorsque l'on veut comparer deux intervalles de longueurs différentes.

Naus (1965) a introduit l'idée de statistique "scan" qui consiste à fixer à l'avance la longueur d des agrégats possibles et à recueillir le nombre maximum d'événements dans un intervalle de longueur d. La distribution de cette statistique a été largement étudiée et tabulée (Huntington & Naus, 1975) sous l'hypothèse H_0 que $\{x_1, \cdots, x_n\}$ soit un échantillon i.i.d. uniforme sur $[0, 1]$, fournissant ainsi un test de cette hypothèse ainsi que l'agrégat de longueur d le plus significatif.

Afin de s'affranchir de cette limitation concernant la taille de l'agrégat, Nagarwalla (1996) introduit la statistique de scan de longueur variable. Il choisit comme indice de concentration sur l'intervalle $[a, a+d]$ la statistique du test de vraisemblance de H_0 contre une hypothèse alternative H_1 de densité constante par morceaux (une constante sur l'intervalle $[a, a + d]$, une autre constante ailleurs). Finalement, l'indice de concentration d'un intervalle de longueur d contenant n_0 événements sera

$$G(d, n_0) = \begin{cases} \left(\frac{n_0}{n}\right)^{n_0} \left(\frac{n-n_0}{n}\right)^{n-n_0} \left(\frac{1}{d}\right)^{n_0} \left(\frac{1}{1-d}\right)^{n-n_0} & \text{si } n_0/d > n, \\ 1 & \text{sinon.} \end{cases}$$

On cherche alors l'intervalle de longueur d contenant n_0 événements maximisant $G(d, n_0)$: on définit $\Lambda = \sup_{n_0 \geq 2} G(d, n_0)$. On remarquera que l'on ne considère que les intervalles contenant au moins 2 événements car il apparaît illogique de considérer les intervalles de longueur infinitésimale centrés sur un événement comme des agrégats, bien que leur indice de concentration $G(d, n_0)$ soit infini. Malheureusement, la distribution de Λ sous H_0 est jusqu'ici inconnue et sa vraisemblance doit être estimée par une procédure de Monte-Carlo.

Cette procédure est efficace mais on peut s'interroger sur l'influence de la manière dont est définie H_1 sur le résultat du test. En effet, cette statistique de

scan sera-t-elle aussi efficace pour détecter des structures d'agrégats différentes de la structure constante par morceaux ? Voilà pourquoi nous proposons une famille de statistiques de tests différentes et qui ne nécessite aucune spécification de l'hypothèse alternative.

Comme les espacements, précédemment définis, sont utiles dans les méthodes non-spécifiques de test de la CSR, nous pensons qu'ils peuvent également être à l'origine de méthodes spécifiques d'identification des agrégats. C'est d'ailleurs la voie suivie par Molinari & *al.* (2001) dont la méthode consiste à effectuer une régression constante par morceaux sur l'échantillon des espacements générés par les événements. On peut faire à cette technique le même reproche concernant l'hypothèse alternative qu'à la statistique de scan, ainsi que le fait de ne pas prendre en compte la dépendance entre les espacements.

Notons $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1$ les statistiques d'ordre associées aux variables aléatoires représentant les localisations des événements et $D_i = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1$, les espacements associés.

Notre idée est d'identifier un agrégat sur $[X_{(j)}, X_{(k)}]$ lorsque la statistique $S_{j,k}^{(l)} = \sum_{i=j+1}^{k} D_i^l$ est anormalement faible. En effet, on peut penser que même sous H_0 il n'est pas très anormal de rencontrer un espacement de faible valeur mais que les soupçons concernant H_0 grandissent si les espacements voisins sont également petits. Il paraît donc logique de s'intéresser aux sommes locales d'espacements à l'ordre l. Un choix de l supérieur à 1 permet de donner plus de poids aux espacements les plus petits.

Dans un premier temps, il nous faut identifier la loi de $S_{j,k}^{(l)}$ sous H_0 ? Dans le cas général, cela n'est pas trivial mais on peut obtenir le résultat pour les deux valeurs $l = 1$ et $l = 2$, que nous utiliserons dans la suite de l'étude.

Finalement, quel que soit le choix de l, la statistique retenue sera

$$S^{(l)*} = \min_{1 \leq j < k \leq n} F_{k-j}^{(l)}(S_{j,k}^{(l)})$$

où $F_{k-j}^{(l)}$ représente la fonction de répartition de $S_{j,k}^{(l)}$ sous l'hypothèse H_0.

En effet, quelle que soit la valeur de $k - j$, le nombre d'événements contenus dans l'intervalle, la loi de $F^{(l)}_{k-j}(S^{(l)}_{j,k})$ est la même sous H_0, à savoir la loi uniforme sur $[0, 1]$, et il apparaît donc pertinent de comparer tous les agrégats potentiels par ce biais.

Comme pour la statistique de scan, la loi de $S^{(l)*}$ sous H_0 est inconnue et devra être évaluée par une procédure de simulation.

L'application de la méthode à des jeux de données réels ou simulés montre sa puissance : il est parfois fait le reproche à la statistique de scan d'englober dans les agrégats les plus significatifs des zones trop larges et cela semble moins le cas pour notre méthode.

Enfin signalons que toutes ces méthodes s'adaptent à l'existence d'une densité sous-jacente, comme par exemple une densité de population inhomogène, puisqu'il suffit d'appliquer l'inverse de la fonction de répartition aux données initiales.

0.5.2. La dimension 2. — L'extension à la dimension 2 n'est pas toujours évidente. En effet, alors qu'en dimension 1 les agrégats potentiels ne sont déterminés que par leur point de départ et leur longueur, en dimension 2 il n'y a aucune restriction de forme. Il existe alors plusieurs voies.

La première consiste à se focaliser sur une famille d'agrégats potentiels et à déterminer la significativité de chacun. C'est la voie empruntée par Kulldorff (1997) qui généralise la statistique de scan à la dimension 2. Les agrégats potentiels sont généralement des disques dont les centres appartiennent à une grille prédéfinie. Comme en dimension 1, l'indice de concentration de l'agrégat est basé sur la statistique du test de vraisemblance de H_0 contre une hypothèse alternative H_1 de densité constante par morceaux (une constante sur le disque, une autre constante ailleurs).

Une deuxième consiste à utiliser l'estimateur à noyau de la densité et à considérer les zones dans lesquelles cet estimateur est significativement plus élevé que la densité sous H_0 (Kelsall & Diggle, 1995). La significativité en chaque point, ainsi qu'un test global de H_0, sont établis par une procédure de type Monte-Carlo.

Enfin, il est possible d'introduire un ordonnancement des événements et de considérer toute suite d'événements consécutifs comme un agrégat potentiel.

C'est l'idée introduite par Dematteï & *al.* (2006) et que nous décidons de suivre.

Dans un premier temps, les événements sont ordonnés en commençant par l'événement le plus proche du bord du domaine d'observation et en continuant par l'événement le plus proche parmi tous les événements non encore parcourus. L'opération se répète jusqu'à ce que tous les événements aient été parcourus. On notera les événements ainsi ordonnés $X_{(1)}, \cdots, X_{(n)}$.

Ensuite, on étudie les distances entre les événements ordonnés et on les compare à leurs fonctions de répartition conditionnelles aux distances précédentes sous l'hypothèse H_0. On obtient alors une suite d'indices de proximité entre chacun des couples, dont on connaît la loi sous H_0 (i.i.d. uniforme sur $[0,1]$) et que l'on peut utiliser de manière similaire aux espacements en dimension 1.

Notons ces indices de proximité $U_1, \cdots, U_n$ où U_1 traduit la proximité du bord du domaine d'observation au premier événement, U_2 la proximité du premier au deuxième événement, etc. Comme en dimension 1, notre idée est d'identifier un agrégat englobant $X_{(j)}, \cdots, X_{(k)}$ lorsque la statistique $S_{j,k}^{(l)} = \sum_{i=j+1}^{k} U_i^l$ est anormalement faible.

Il nous faut identifier la loi de $S_{j,k}^{(l)}$ sous H_0 ? Là encore, on obtient le résultat pour les deux valeurs $l = 1$ et $l = 2$, que nous utiliserons dans la suite de l'étude.

Finalement, quel que soit le choix de l, la statistique retenue sera

$$S^{(l)*} = \min_{1 \leq j < k \leq n} F_{k-j}^{(l)}(S_{j,k}^{(l)})$$

où $F_{k-j}^{(l)}$ représente la fonction de répartition de $S_{j,k}^{(l)}$ sous l'hypothèse H_0.

Comme en dimension 1, la loi de $S^{(l)*}$ sous H_0 est inconnue et devra être évaluée par une procédure de simulation. Lorsqu'un agrégat est déclaré significatif, reste en dernier lieu à déterminer sa forme : en effet, nous connaissons seulement les événements qu'il englobe. Deux solutions sont proposées par Dematteï & *al.* (2006). La première consiste à identifier un agrégat à la réunion des cellules de Voronoi des événements qui le composent. La seconde à définir une zone d'influence pour chaque événement comme le cercle centré en cet événement dont l'aire est l'aire d'observation totale divisée par n : l'agrégat est alors la réunion des zones d'influence des événements qui le composent.

Nous appliquons ensuite la méthode à des jeux de données réels ou simulés.

Là encore, cette méthode ainsi que celles décrites précédemment s'adaptent à la présence d'une densité de population inhomogène sous-jacente qui peut être connue ou estimée, mais également aux localisations d'événements dits "de contrôle" (par opposition aux événements "cas" que l'on cherche à analyser).

Il est à noter qu'il existe de nombreuses méthodes de détection d'agrégats lorsque l'on dispose de données groupées, i.e. du nombre d'événements dans chaque zone issue d'un découpage géographique (Lawson, 2001). Ces zones peuvent être par exemple des communes, des cantons, des départements... Ici, nous nous plaçons sous l'hypothèse que l'on dispose des localisations précises des événements et il apparaît nécessaire d'utiliser toute l'information disponible, sans recourir à un découpage parfois peu pertinent.

0.6. L'estimation d'intensité

Le premier chapitre de la partie 3 présente une revue bibliographique sur le thème des méthodes non-paramétriques en statistique spatiale. Dans le domaine de la géostatistique, on s'intéresse notamment aux techniques de prédiction d'un champ aléatoire et d'estimation non-paramétrique du variogramme. Dans le domaine des processus ponctuels, on se concentre principalement sur les méthodes d'estimation des caractéristiques du premier et du second ordre du processus. Ce paragraphe résume les techniques d'estimation d'intensité existantes et expose un résultat asymptotique décrit dans le premier chapitre de la partie 3.

La fonction intensité, définie précédemment, permet de résumer les propriétes de premier ordre d'un processus ponctuel. Cela équivaut à la tendance pour un champ aléatoire. D'après la seconde proposition sur les processus de Poisson inhomogènes, cette notion d'intensité est fortement liée à celle de densité multidimensionnelle et, de fait, ses méthodes d'estimation s'apparentent à celles d'estimation de densité. Dans ce paragraphe, nous nous plaçons sous l'hypothèse Poissonnienne.

On peut clairement distinguer deux familles d'estimateurs : les estimateurs paramétriques et les estimateurs non-paramétriques.

L'estimation paramétrique nécessite le choix d'une famille de fonctions intensité $\mathcal{F} : \{\lambda_c : \int_D \lambda_c(s)\nu(ds) = n\}$. c est le vecteur de paramètres et la condition sur λ_c découle de la propriété $\mathbb{E}N = \int_D \lambda(s)\nu(ds)$. L'estimateur retenu sera la fonction de $\mathcal{F}$ qui maximise la vraisemblance

$$L(c; s_1, \cdots, s_n) = \frac{\lambda(s_1) \cdots \lambda(s_n)}{\big(\int_D \lambda(s)\nu(ds)\big)^n}.$$

Ogata & Katsura (1988) utilisent cette technique pour analyser la localisation de secousses sismiques, en choisissant comme fonctions d'intensité possibles une famille de B-splines cubiques.

Mais généralement le choix d'une famille paramétrique apparaît trop restrictif et la plupart optent pour un estimateur non paramétrique de l'intensité dérivé des estimateurs de densité multidimensionnelle.

L'estimateur à noyau de la densité multidimensionnelle (Silverman, 1986) est

$$\hat{f}_h(s) = \frac{1}{nh^d} \sum_{i=1}^{n} K\big(\frac{s - s_i}{h}\big)$$

où K est un noyau d-dimensionnel d'intégrale 1.

Le rapport $\frac{1}{n}$ permet d'avoir la propriété $\int_{\mathbb{R}^d} \hat{f}_h(s) = 1$. La première version de l'estimateur à noyau de l'intensité est alors

$$\hat{\lambda}_h(s) = \frac{1}{h^d} \sum_{i=1}^{n} K\big(\frac{s - s_i}{h}\big)$$

de telle sorte que $\int_{\mathbb{R}^d} \hat{\lambda}_h(s) = n$.

Il nous faut ensuite prendre en compte le fait que le processus ponctuel ne peut être observé que sur un domaine limité, en l'occurence D. L'estimation d'intensité peut en pâtir et il peut en résulter une sous-estimation, notamment sur les bords du domaine. Ce problème peut se résoudre par l'introduction d'un terme de correction au bord et l'estimateur introduit par Diggle (1985)

$$\hat{\lambda}_{EC,h}(s) = \frac{\frac{1}{h^d} \sum_{i=1}^{n} K\big(\frac{s-s_i}{h}\big)}{\int_D \frac{1}{h^d} K\big(\frac{s-u}{h}\big)\nu(du)} = \frac{\sum_{i=1}^{n} K\big(\frac{s-s_i}{h}\big)}{\int_D K\big(\frac{s-u}{h}\big)\nu(du)}$$

possède notamment la propriété d'être asymptotiquement sans biais dans un cadre asymptotique spécifique décrit dans la partie 3.

Le problème inhérent à toute estimation par noyau est celui de la sélection de largeur de bande. On cherchera généralement à minimiser l'erreur quadratique moyenne intégrée $EQMI(h) = \mathbb{E}\big[\int_D \big(\lambda(s) - \hat{\lambda}_{EC,h}(s)\big)^2 \nu(ds)\big]$.

En estimation de densité, la plupart des méthodes (Scott, 1992) s'appuient sur la décomposition de l'erreur quadratique moyenne intégrée en un terme de biais et un terme de variance puis utilisent les expressions du biais et de la variance asymptotiques. Malheureusement, l'introduction du terme de correction au bord complique singulièrement ces expressions et on utilisera généralement une méthode de validation croisée (Xu & *al.*, 2003). Celle-ci s'appuie sur la décomposition

$$EQMI(h) = \int_D \lambda(s)^2 \nu(ds) - 2\mathbb{E}\Big[\int_D \hat{\lambda}_{EC,h}(s)\lambda(s)\nu(ds)\Big] + \mathbb{E}\Big[\int_D \hat{\lambda}^2_{EC,h}(s)\nu(ds)\Big].$$

Le premier terme est constant, le deuxième peut être estimé par $\sum_{i=1}^n \hat{\lambda}_{EC,h}(s_i)$ et le dernier par $\int_D \hat{\lambda}^2_{EC,h}(s)\nu(ds)$. On retiendra donc la valeur de h qui minimise $\sum_{i=1}^n \hat{\lambda}_{EC,h}(s_i) + \int_D \hat{\lambda}^2_{EC,h}(s)\nu(ds)$.

Enfin on signalera une approche originale baséé sur un modèle hiérarchique bayésien et qui considère une fonction intensité constante par morceaux, les "morceaux" correspondant aux cellules de Voronoi issues d'un ensemble de points évoluant dynamiquement (Heikkinen & Arjas, 1998 and Byers & Raftery, 2002).

0.7. La gestion des erreurs de mesure

Tout échantillon de données est basé sur des mesures qui peuvent s'avérer plus ou moins précises, mais dont l'incertitude ne peut être totalement nulle. Les moyens actuels permettent souvent de multiplier les enregistrements et donc d'obtenir des informations pertinentes sur la nature des erreurs de mesure. Dans de nombreux domaines de la statistique, les chercheurs ont tenté d'incorporer ces erreurs de mesure dans les modèles afin d'être plus proches de la réalité (Fuller, 1987).

Lorsque l'on traite des données géoréférencées, les erreurs de localisation constituent tout ou partie des erreurs de mesure. Si l'on s'intéresse à un champ aléatoire, ces erreurs sur la localisation des données s'ajoutent aux erreurs sur

les mesuses du champ lui-même. Cressie & Kornak (2003) traitent la question du krigeage (prédiction du champ en tous points, basée sur une combinaison linéaire des valeurs observées) en prenant en compte ces deux types d'erreur.

En ce qui concerne les processus ponctuels spatiaux non marqués, les seules données étant les localisations des événements, seules ces erreurs de localisation sont à prendre en compte. Il existe peu de littérature sur ce sujet. Lund & Rudemo (2000), disposant d'un jeu de données recensant observations et "vraies" localisations (sans connaître toutefois la correspondance), estiment les paramètres des erreurs de localisation. Bar-Hen & *al.* (2005) s'intéressent eux à l'estimation de certaines caractéristiques de second-ordre du processus comme la fonction de répartition de la distance au plus proche voisin G. Pour celà, ils supposent un modèle additif

$$z_i = s_i + \epsilon_i, \quad i = 1, \cdots, n$$

où $(z_1, \cdots, z_n)$ est le vecteur des observations et $(\epsilon_1, \cdots, \epsilon_n)$ le vecteur des erreurs de localisation. Les erreurs sont supposées i.i.d. de densité $g(.)$ et indépendantes des "vraies" localisations $(s_1, \cdots, s_n)$. Ils utilisent ensuite une méthode de simulation-extrapolation (SIMEX) mise en place originellement dans le cadre de la régression paramétrique (Cook & Stefanski, 1994). Cette méthode consiste en trois étapes :

- introduire artificiellement des erreurs de mesure de variance σ^2 aux données observées.
- réestimer à chaque fois le paramètre de régression θ par $\hat{\theta}$ (par n'importe quelle méthode : maximum de vraisemblance, moindres carrés...) et le considérer en tant que fonction de σ^2 : $\hat{\theta}(\sigma^2)$.
- extrapoler la valeur de $\hat{\theta}$ quand la variance de l'erreur est nulle.

Cette technique a l'avantage d'être très facile à implémenter et de pouvoir être appliquée à de nombreux modèles pour lesquels la prise en compte des erreurs de mesure rend les choses inextricables. Malheureusement, il est difficile de choisir la méthode d'extrapolation à utiliser dans la troisième étape. De nombreux articles suggèrent, d'un point de vue pratique, d'utiliser une extrapolation quadratique mais il n'existe jusqu'à présent aucune justification théorique.

Si l'on s'intéresse maintenant à l'estimation de la fonction intensité dans le cas bruité, quelles sont les méthodes envisageables ? La première idée consiste à utiliser une méthode de type SIMEX mais, en plus du problème invoqué précédemment, on peut recenser un autre inconvénient spécifique au contexte d'étude. En effet, nous cherchons à estimer la fonction intensité sur le domaine d'observation D, ce qui nous oblige à réaliser une extrapolation pour chaque

point de D (en pratique pour chaque point d'une grille de D) et ce sans prendre en compte l'aspect spatial.

Les similitudes entre l'estimation d'intensité et l'estimation de densité nous amènent naturellement à regarder ce qui se fait dans ce domaine en présence d'observations bruitées. Il existe justement une littérature conséquente sur ce sujet, dans le cas réel, et qui s'appuie sur une méthode de déconvolution issue du modèle additif

$$y_i = x_i + \epsilon_i, \quad i = 1, \cdots, n$$

où $(y_1, \cdots, y_n)$ est le vecteur des observations de densité inconnue $h(.)$, $(x_1, \cdots, x_n)$ le vecteur des "vraies" valeurs de densité inconnue $f(.)$et $(\epsilon_1, \cdots, \epsilon_n)$ le vecteur des erreurs de mesure de densité connue $g(.)$. Comme précédemment, les erreurs sont supposées i.i.d. et indépendantes des "vraies" valeurs. Alors, si on note $\mathcal{F}$ la transformée de Fourier, on a

$$\begin{aligned}
& y_i = x_i + \epsilon_i, \quad i = 1, \cdots, n \\
\Rightarrow \quad & h = f * g \\
\Rightarrow \quad & \mathcal{F}(h)(.) = \mathcal{F}(f)(.)\ \mathcal{F}(g)(.) \\
\Rightarrow \quad & \mathcal{F}(f)(.) = \mathcal{F}(h)(.)/\mathcal{F}(g)(.) \\
\Rightarrow \quad & f = \mathcal{F}^{-1}\big(\mathcal{F}(h)(.)/\mathcal{F}(g)(.)\big).
\end{aligned}$$

L'idée introduite par Stefanski & Carroll (1990) est d'obtenir un estimateur de $f(.)$ en remplaçant dans cette formule $h(.)$ par son estimateur à noyau classique issu des observations. On s'aperçoit que l'existence de la transformée de Fourier inverse est assurée en utilisant un noyau à bande limitée, i.e. dont la transformée de Fourier est à support compact. Cet estimateur se révèle asymptotiquement sans biais et consistent. La question du choix de la largeur de bande se pose à nouveau. Comme dans le cas non bruité, on peut distinguer les méthodes s'appuyant sur les expressions du biais et de la variance asymptotiques de celles de type validation croisée, ces dernières s'avérant moins performantes (Delaigle & Gijbels, 2004). Les premières nécessitent l'estimation de $\int f''(x)^2 dx$ pour laquelle une méthode est donnée par Delaigle & Gijbels (2001).

On peut également noter que cette méthode de déconvolution est aussi utilisée en régression non-paramétrique en présence d'erreurs (Fan & Truong, 1993).

0.8. L'estimation d'intensité dans le cas bruité

Ce paragraphe résume les résultats énoncés dans le deuxième chapitre de la partie 3.

En conservant le même modèle additif que Bar-Hen & *al.* (2005), la méthode de déconvolution s'adapte à l'estimation d'intensité et l'estimateur ainsi construit s'écrit

$$\lambda^{**}_{Y,h}(s) = \frac{\sum_{j=1}^{n} \frac{1}{h^d} K_h^*\left(\frac{s-z_j}{h}\right)}{\int_D \frac{1}{h^d} K_h^*\left(\frac{s-u}{h}\right)\nu(du)}, \forall s \in G'_h,$$

où $K_h^*(t) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{it'y} \mathcal{F}(K)(y)/\mathcal{F}(g)(y/h)dy$, K est un noyau d-dimensionnel et $G'_h = \{s \in \mathbb{R}^d : \int_D \frac{1}{h^d} K_h^*\left(\frac{s-u}{h}\right)\nu(du) \neq 0\}$.

Cet estimateur, qui s'identifie à l'estimateur classique de Diggle, ne possède malheureusement pas les mêmes qualités. Il est notamment asymptotiquement biaisé, sauf dans le cas où le processus ponctuel est Poisson homogène. Celà se produit du fait de l'addition de l'observation sur un domaine limité et de la présence d'erreurs : en effet, la méthode de déconvolution est telle que l'intensité du processus "réel" en un point dépend de l'intensité du processus observé en tout point de $\mathbb{R}^d$. Il est donc tentant de supposer que les relatives faiblesses de l'estimateur sont dûes uniquement à la complexité du problème.

La question du choix de la largeur de bande est débattue. D'un côté, les termes de biais et de variance asymptotiques se révèlent complexes et notamment fortement dépendants du domaine d'observation D : il apparaît difficile de les utiliser pour approximer l'erreur quadratique moyenne intégrée. De l'autre côté, les méthodes classiques de type validation croisée se révèlent inutilisables car l'on ne dispose pas des "vraies" localisations. Voilà pourquoi nous décidons d'occulter la limitation du domaine d'observation, conscient que le terme de correction de bord n'influence que très peu l'estimation aux points éloignés de la frontière, car il est alors proche de 1. Le problème revient désormais à adapter les techniques existantes en estimation de densité unidimensionnelle à la d-dimension. Cela se fait sans difficulté, excepté pour l'estimation nécessaire de $\int f''(s)^2\nu(ds)$: la technique de Delaigle & Gijbels ne s'adaptant pas facilement, nous préférons utiliser la règle de la référence gaussienne, qui consiste à supposer une densité gaussienne pour effectuer cette estimation.

D'un point de vue pratique, l'estimateur introduit se révèle bien plus efficace que l'estimateur naïf de Diggle, ou un estimateur qui n'introduirait aucune correction au bord, prouvant par là que dans ce cadre, déconvolution et correction au bord sont à mener parallèlement.

PARTIE I

TESTS D'HOMOGÉNÉITÉ SPATIALE BASÉS SUR LES ESPACEMENTS BIDIMENSIONNELS

Cette partie reprend, dans un premier chapitre, l'étude asymptotique d'une famille de statistiques basées sur les espacements bidimensionnels (Cucala, 2005) puis, dans le second chapitre, l'application de ces statistiques dans une étude pratique (Cucala & Thomas-Agnan, 2006a).

CHAPITRE 1

LE CAM SPACINGS THEOREM IN DIMENSION TWO

Abstract

The definition of spacings associated to a sequence of random variables is extended to the case of random vectors in $[0,1]^2$. Beirlant & *al.* (1991) give an alternative proof of the Le Cam (1958) theorem concerning asymptotic normality of additive functions of uniform spacings in $[0,1]$. We adapt their technique to the two-dimensional case, leading the way to new directions in the domain of Complete Spatial Randomness (CSR) testing.

1.1. Introduction

Testing the uniformity of real random variables (r.v.) can be done in several ways : using Chi-square tests, tests based on the empirical distribution function (e.d.f.), tests based on spacings ... The latter ones have been extensively studied (Pyke, 1965) and recommended, for example, for the analysis of the local renewal structure of a point process (Deheuvels, 1983a).

In higher dimensions, when dealing with a spatial point pattern $U \in S \subset \mathbb{R}^d$, one first wishes to know whether it satisfies the CSR hypothesis : is the spatial process governing U a homogeneous Poisson process ? This question is equivalent to the following : given the number of points in the pattern (also called events), are these points uniformly and independently distributed in S (Moller & Waagepetersen, 2004) ?

We concentrate here on point patterns distributed in rectangles in $\mathbb{R}^2$, which is similar, after linear transformation of the coordinates, to testing the uniformity in $[0,1]^2$.

Most of two-dimensional uniformity tests are either Chi-square tests or distance-based methods (Cressie, 1993). The first ones depend on the number and location of the quadrats (cells in which events are counted), whereas the last ones require numerous simulations. More recently, there has been some interest in e.d.f.-based methods and extensions of the Cramer-Von Mises test(Zimmerman, 1993) and the Kolmogorov-Smirnov test (Justel & *al*, 1997) to the $[0,1]^2$ case have been established.

On the other hand, spacings theory, so useful for testing uniformity on $\mathbb{R}$, remains almost unworked in higher dimensions even if one may think, as Zimmerman (1993) does, that distances from events to their nearest neighbours can be viewed as two-dimensional analogues of spacings. We shall just mention the results of Deheuvels (1983b) and Janson (1987) concerning the asymptotic distribution of the maximal multidimensional spacing, i.e. the volume of the largest square (or ball) contained in $[0,1]^2$ and avoiding every point of the pattern.

A first application of spacings theory to CSR testing would be to test both x- and y-coordinates' uniformity using a spacings-based method. The rejection of either leads to refuse the two-dimensional uniformity hypothesis. But we can never accept it as a bivariate distribution with uniform marginals need not be uniform. This makes necessary to take into account the joint distribution of the x- and y-coordinates.

In this paper, following this idea, we introduce a new notion of two-dimensional spacings which is related to spacings based on x- and y-coordinates. This relationship then allows us to derive the limiting distribution, under the uniformity hypothesis, of a wide family of statistics based on these spacings. This is done by a direct decomposition method similar to the one Beirlant & *al* (1991) used for one-dimensional spacings.

An application of this asymptotic result is developed by Cucala & Thomas-Agnan (2006a). Two of these statistics, the variance and the absolute mean deviation of the two-dimensional spacings, are selected and used in practice to test for CSR. A multiple procedure is adopted to generalize the tests to point processes in any domain (not necessarily rectangular). Then the power

of these spacings-based tests is compared to the power of existing tests using real and simulated data sets : they appear to be inferior for detecting regularity or clustering but more powerful for detecting certain types of heterogeneity.

1.2. Spacings in $[0,1]^2$

1.2.1. Definition. — Let $\mathsf{U} = \Big((U_1^x, U_1^y), \cdots, (U_{n-1}^x, U_{n-1}^y)\Big)$ be a point pattern in $[0,1]^2$.

Let $\mathsf{U}^{\mathsf{x}} = (U_1^x, \cdots, U_{n-1}^x)$ and $\mathsf{U}^{\mathsf{y}} = (U_1^y, \cdots, U_{n-1}^y)$.

$U_{(1)}^x \leq \cdots \leq U_{(n-1)}^x$ are the order statistics corresponding to U^{x}.

$U_{(1)}^y \leq \cdots \leq U_{(n-1)}^y$ are the order statistics corresponding to U^{y}.

Set $U_0^x = U_0^y = U_{(0)}^x = U_{(0)}^y = 0$ and $U_n^x = U_n^y = U_{(n)}^x = U_{(n)}^y = 1$.

One may define the spacings related to the pattern

$$\begin{aligned} &\text{the x-spacings} \quad D_i^x = U_{(i)}^x - U_{(i-1)}^x, \quad i = 1, \cdots, n \ , \\ &\text{the y-spacings} \quad D_j^y = U_{(j)}^y - U_{(j-1)}^y, \quad j = 1, \cdots, n \ . \end{aligned}$$

A way to take account of how x- and y-spacings vary jointly is then to define the two-dimensional spacings as the areas A_{ij} formed by the grid in Figure 1

$$\forall i \in \{1, \cdots, n\}, \forall j \in \{1, \cdots, n\}, \quad A_{ij} = D_i^x D_j^y.$$

1.2.2. Uniformity hypothesis. — From now on, we will assume the uniformity hypothesis H_0 : the point pattern U is uniformly distributed in $[0,1]^2$.

Let $(D_1, \cdots, D_n)$ be the spacings corresponding to a (n-1) uniform sample on $[0,1]$. Then it is easy to see that

$$\begin{cases} \mathcal{L}(D_1^x, \cdots, D_n^x) = \mathcal{L}(D_1, \cdots, D_n), \\ \mathcal{L}(D_1^y, \cdots, D_n^y) = \mathcal{L}(D_1, \cdots, D_n), \\ (D_1^x, \cdots, D_n^x) \perp\!\!\!\perp (D_1^y, \cdots, D_n^y). \end{cases}$$

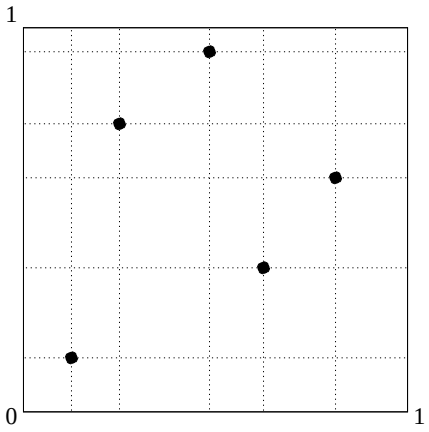

FIGURE 1. Two-dimensional spacings

1.3. Asymptotic normality of additive functions of spacings

1.3.1. Main result. — Many statistics based on one-dimensional spacings are additive functions

$$V_n = \sum_{i=1}^{n} g(nD_i)$$

for a measurable function g.

For the two-dimensional case consider

$$V_n^{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n} g(n^2 A_{ij}).$$

The asymptotic normality of V_n was proved by Beirlant & *al.* (1991) using the following distributional equivalence (Moran, 1947)

$$(nD_1, \cdots, nD_n) \sim (\frac{E_1}{\bar{E}}, \cdots, \frac{E_n}{\bar{E}}) \tag{1}$$

where $(E_1, \cdots, E_n)$ are independent exponentially distributed r.v.'s with mean 1

and $$\bar{E} = \frac{1}{n} \sum_{i=1}^{n} E_i \ .$$

We will use the same technique to prove the asymptotic normality of $V_n^{(2)}$ under H_0.

Introduce $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ two samples of independent exponentially distributed r.v.'s with mean 1. Denote $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$. Then define the statistic

$$G_n = \sum_{i=1}^{n}\sum_{j=1}^{n} g\Big(\frac{X_i}{\bar{X}}\frac{Y_j}{\bar{Y}}\Big).$$

From (1) , $V_n^{(2)}$ and G_n have same distribution.

From now on we will assume g satisfies the following, where φ and ψ are measurable functions

$$g \text{ continuous on } \mathbb{R}^+, \tag{2}$$

$$\mathbb{E}g^2(X_1Y_1) < \infty, \tag{3}$$

$$\forall t_0 \in \mathbb{R}^+, \exists \varphi : \mathbb{R} \to \mathbb{R}, \forall n \in \mathbb{N}, \mathbb{E}\varphi(\bar{X}\bar{Y}) < \infty \text{ and } \forall t < t_0, \forall x \in \mathbb{R}^+, |g(tx)| < \varphi(x), \tag{4}$$

$$\forall t_0 \in \mathbb{R}^+, \exists \psi : \mathbb{R} \to \mathbb{R}, \forall n \in \mathbb{N}, \mathbb{E}\psi(\bar{X}\bar{Y}) < \infty \text{ and } \forall t > t_0, \forall x \in \mathbb{R}^+, \Big|\frac{g(tx)}{g(t)}\Big| < \psi(x). \tag{5}$$

Denote

$$\begin{aligned} \mu &= \mathbb{E}[g(X_1Y_1)], \\ \eta &= Cov\big(g(X_1Y_1), g(X_1Y_2)\big) = Cov\big(g(X_1Y_1), g(X_2Y_1)\big), \\ c &= Cov\big(g(X_1Y_1), X_1\big) = Cov\big(g(X_1Y_1), Y_1\big). \end{aligned}$$

To justify the decomposition, we make the following argument.

Using the same Taylor-expansion as Proschan & Pyke (1964), it appears that, if the function g was differentiable, $n^{-3/2}(G_n - n^2\mu)$ could be asymptotically equivalent in distribution to

$$n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{g(X_iY_j)-\mu-(\bar{X}-1)g'(X_iY_j)(X_iY_j)-(\bar{Y}-1)g'(X_iY_j)(X_iY_j)\}$$
$$=n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big\{g(X_iY_j)-\mu-(X_i-1+Y_j-1)\Big(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}g'(X_kY_l)(X_kY_l)\Big)\Big\}.$$

By partial integration it follows that

$$\mathbb{E}g'(X_1Y_1)(X_1Y_1)=c \quad \Rightarrow \quad \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}g'(X_kY_l)(X_kY_l)\xrightarrow[n\to\infty]{P}c.$$

That is why it seems useful to decompose G_n as follows

$$\frac{1}{n^{3/2}}\big(G_n-n^2\mu\big)=\frac{S_n}{n^{3/2}}+\frac{R_n}{n^{3/2}}$$
$$\text{where } S_n=\sum_{i=1}^{n}\sum_{j=1}^{n}[g(X_iY_j)-\mu-c(X_i-1)-c(Y_j-1)]$$
$$\text{and } R_n=\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[g\Big(\frac{X_i}{\bar{X}}\frac{Y_j}{\bar{Y}}\Big)-g(X_iY_j)\Big]+cn^2(\bar{X}-1)+cn^2(\bar{Y}-1).$$

As $\dfrac{S_n}{n^{3/2}}$ is a two-sample U-statistic with mean 0 and limiting variance

$\sigma^2=2(\eta-c^2)$, one has from Van Der Vaart (1998)

$$\frac{S_n}{n^{3/2}}\xrightarrow[n\to\infty]{d}\mathcal{N}(0,\sigma^2).$$

We will prove in the next section that

$$\frac{\mathbb{E}(R_n^2)}{n^3}\xrightarrow[n\to\infty]{}0,$$

which will yield the following result.

Theorem 1. — *Assume g satisfies (2), (3), (4) and (5). Then*

$$\frac{1}{n^{3/2}}\{G_n - n^2\mu\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma^2).$$

1.3.2. Behaviour of the remainder term. — Using the independence of $\frac{X_i}{\bar{X}}$ and $\bar{X}$, as well as the independence of $\frac{Y_j}{\bar{Y}}$ and $\bar{Y}$, one gets

$$\frac{\mathbb{E}(R_n^2)}{n^3} = T_{1,n} + 2\,T_{2,n} + T_{3,n} \quad \text{where}$$

$$T_{1,n} = \frac{1}{n}\Big[\mathbb{E}\Big\{g^2\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)\Big\} - 2\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_1Y_1)\Big\} + \mathbb{E}g^2(X_1Y_1)\Big], \qquad (6)$$

$$T_{2,n} = \frac{n-1}{n}\Big[\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g\Big(\frac{X_1}{\bar{X}}\frac{Y_2}{\bar{Y}}\Big)\Big\} - 2\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_1Y_2)\Big\} + \mathbb{E}g(X_1Y_1)g(X_1Y_2)\Big], \qquad (7)$$

$$\text{and } T_{3,n} = \frac{(n-1)^2}{n}\Big[\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g\Big(\frac{X_2}{\bar{X}}\frac{Y_2}{\bar{Y}}\Big)\Big\} - 2\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_2Y_2)\Big\} + \mu^2\Big] - 2c^2. \qquad (8)$$

1.3.2.1. Preliminary results. — The marginal and bivariate densities of the spacings D_i are given by Pyke (1965) and lead to

$$\mathbb{E}g^2\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big) = \frac{(n-1)^2}{n^2}\int_0^n\int_0^n g^2(xy)\Big(1-\frac{x}{n}\Big)^{n-2}\Big(1-\frac{y}{n}\Big)^{n-2}dy\,dx, \qquad (9)$$

$$\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g\Big(\frac{X_1}{\bar{X}}\frac{Y_2}{\bar{Y}}\Big)\Big\} = \frac{(n-1)^2(n-2)}{n^3}\int_0^n\int_0^n\int_0^{n-y} g(xy)g(xv)\Big(1-\frac{x}{n}\Big)^{n-2}\Big(1-\frac{y+v}{n}\Big)^{n-3}dv\,dy\,dx, \qquad (10)$$

$$\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g\Big(\frac{X_2}{\bar{X}}\frac{Y_2}{\bar{Y}}\Big)\Big\} = \frac{(n-1)^2(n-2)^2}{n^4}\int_0^n\int_0^n\int_0^{n-x}\int_0^{n-y} g(xy)g(uv)\Big(1-\frac{x+u}{n}\Big)^{n-3}\Big(1-\frac{y+v}{n}\Big)^{n-3}dv\,du\,dy\,dx. \qquad (11)$$

Using the independence of $\frac{X_i}{\bar{X}}$ and $\bar{X}$, one finds

$$\begin{aligned}\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_1Y_1)\Big\} &= \mathbb{E}\Big[g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\bar{X}\,\frac{Y_1}{\bar{Y}}\bar{Y}\Big)\Big|\frac{X_1}{\bar{X}},\frac{Y_1}{\bar{Y}}\Big\}\Big]\\ &= \frac{(n-1)^2}{n^2}\int_0^n\int_0^n g(xy)\mathbb{E}\{g(x\bar{X}y\bar{Y})\}\\ &\quad \Big(1-\frac{x}{n}\Big)^{n-2}\Big(1-\frac{y}{n}\Big)^{n-2}dy\,dx,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_1Y_2)\Big\} &= \mathbb{E}\Big[g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\bar{X}\,\frac{Y_2}{\bar{Y}}\bar{Y}\Big)\Big|\frac{X_1}{\bar{X}},\frac{Y_1}{\bar{Y}},\frac{Y_2}{\bar{Y}}\Big\}\Big]\\ &= \frac{(n-1)^2(n-2)}{n^3}\int_0^n\int_0^n\int_0^{n-y} g(xy)\mathbb{E}\{g(x\bar{X}v\bar{Y})\}\\ &\quad \Big(1-\frac{x}{n}\Big)^{n-2}\Big(1-\frac{y+v}{n}\Big)^{n-3}dv\,dy\,dx, \qquad (12)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_2Y_2)\Big\} &= \mathbb{E}\Big[g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)\mathbb{E}\Big\{g\Big(\frac{X_2}{\bar{X}}\bar{X}\,\frac{Y_2}{\bar{Y}}\bar{Y}\Big)\Big|\frac{X_1}{\bar{X}},\frac{X_2}{\bar{X}},\frac{Y_1}{\bar{Y}},\frac{Y_2}{\bar{Y}}\Big\}\Big]\\ &= \frac{(n-1)^2(n-2)^2}{n^4}\int_0^n\int_0^n\int_0^{n-x}\int_0^{n-y} g(xy)\mathbb{E}\{g(u\bar{X}v\bar{Y})\}\\ &\quad \Big(1-\frac{x+u}{n}\Big)^{n-3}\Big(1-\frac{y+v}{n}\Big)^{n-3}dv\,du\,dy\,dx. \qquad (13)\end{aligned}$$

The following lemma is also needed.

Lemma 1. — *If g continuous on $\mathbb{R}^{+\star}$ and $\mathbb{E}g^2(X_1Y_1)<\infty$, then $\forall t\in[0,+\infty[$*

$$\lim_{n\to\infty}\mathbb{E}g(t\bar{X}\bar{Y}) = g(t).$$

Proof :

Denote by $f_n(u,t)=\dfrac{n^n}{\Gamma(n)}\dfrac{u^{n-1}}{t^n}e^{-nu/t}$ the common density of $t\bar{X}$ and $t\bar{Y}$.

$$\mathbb{E}g(t\bar{X}\bar{Y}) = \mathbb{E}g(\sqrt{t}\bar{X}\sqrt{t}\bar{Y}) = \int_0^\infty \int_0^\infty g(uv) f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) du\, dv$$
$$= \int\int_D g(uv) f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) du\, dv + \int\int_{\bar{D}} g(uv) f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) du\, dv$$

where $D = \{(u,v) \in \mathbb{R}^{+2}; \sqrt{t}/2 < u < 3\sqrt{t}/2; \sqrt{t}/2 < v < 3\sqrt{t}/2\}$.

It is easy to see that :

$\sqrt{t}\bar{X} \xrightarrow[n\to\infty]{P} \sqrt{t}$ and $\sqrt{t}\bar{Y} \xrightarrow[n\to\infty]{P} \sqrt{t}$.

Introduce the function φ : $D \to \mathbb{R}$

$$(x,y) \to \varphi(x,y) = g(xy).$$

From (2) the function φ is continuous and bounded on D , so by the Helly-Bray theorem one concludes

$$\int\int_D g(uv) f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) du\, dv = \mathbb{E}\,\varphi(\sqrt{t}\bar{X}, \sqrt{t}\bar{Y}) \xrightarrow[n\to\infty]{} \varphi(\sqrt{t},\sqrt{t}) = g(t).$$

It remains to prove that $\int\int_{\bar{D}} g(uv) f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) du\, dv \xrightarrow[n\to\infty]{} 0$.

K and $\tilde{K}$ are two constants. From Beirlant & *al* (1991), one has $\forall n \geq 1$

$$f_n(u,\sqrt{t}) \leq K t^{-1/2} n^{1/2} \Big(\frac{u}{\sqrt{t}} e^{1-u/\sqrt{t}}\Big)^{n-1}.$$

So : $(u,v) \in \bar{D} \Rightarrow f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) \xrightarrow[n\to\infty]{} 0$.

Moreover : $n \geq n_0 \geq 5 \Rightarrow f_n(u,\sqrt{t}) f_n(v,\sqrt{t}) \leq \tilde{K} f_{n_0}(u,\sqrt{t}) f_{n_0}(v,\sqrt{t})$.

And taking $m \geq 16$, we get, $\forall u \in [0, \sqrt{t}/2] \cup [3\sqrt{t}/2, +\infty]$

$$\Big(\frac{u}{\sqrt{t}} e^{1-u/\sqrt{t}}\Big)^m < e^{-u/\sqrt{t}}$$

$$\Rightarrow \quad f_{n_0}(u, \sqrt{t}) < K^{1/2} t^{-1/2} n_0^{1/2} e^{-u} \qquad \text{if we take } n_0 > 16\sqrt{t} + 1.$$

As, from (3), $(u, v) \to g(uv)e^{-u}e^{-v} \in \mathcal{L}^1(\mathbb{R}^2)$, one gets

$$\int\int_{\bar{D}} g(uv) f_{n_0}(u, \sqrt{t}) f_{n_0}(v, \sqrt{t}) du\, dv < \infty.$$

Lebesgue's dominated-convergence theorem leads to the conclusion. ■

1.3.2.2. Behaviour of $T_{1,n}$. — $\forall (x, y) \in [0, n]^2$,

$$|g^2(xy)(1 - x/n)^{n-2}(1 - y/n)^{n-2}| \leq g^2(xy)e^4 e^{-x} e^{-y} \quad \in \mathcal{L}^1([0, n]^2) \text{ from (3).}$$

So applying Lebesgue's dominated-convergence theorem to (9) leads to

$$\lim_{n\to\infty} \mathbb{E} g^2\Big(\frac{X_1}{\bar{X}} \frac{Y_1}{\bar{Y}}\Big) = \int_0^\infty \int_0^\infty g^2(xy) e^{-x} e^{-y} dy\, dx = \mathbb{E} g^2(X_1 Y_1). \tag{14}$$

By Cauchy-Schwarz inequality one gets

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}\Big\{\Big| g\Big(\frac{X_1}{\bar{X}} \frac{Y_1}{\bar{Y}}\Big) g(X_1 Y_1)\Big|\Big\} &\leq \{\mathbb{E} g^2(X_1 Y_1)\}^{1/2} \quad \lim_{n\to\infty} \Big\{\mathbb{E} g^2\Big(\frac{X_1}{\bar{X}} \frac{Y_1}{\bar{Y}}\Big)\Big\}^{1/2} \\ &= \mathbb{E} g^2(X_1 Y_1) \quad \text{from (14)} \\ &< \infty. \end{aligned} \tag{15}$$

(3), (14) and (15) lead to

$$\lim_{n\to\infty} T_{1,n} = 0.$$

1.3.2.3. Behaviour of $T_{2,n}$. — $\forall (x, y) \in [0, n]^2$, $\forall v \in [0, n - y]$,

$|g(xy)g(xv)(1 - x/n)^{n-2}(1 - (y + v)/n)^{n-3}| \leq |g(xy)g(xv)e^5 e^{-x} e^{-y} e^{-v}| \quad \in \mathcal{L}^1([0, n]^2 \times [0, n - y])$ from (3).

So applying Lebesgue's dominated-convergence theorem to (10) leads to

$$\lim_{n\to\infty} \mathbb{E} g\Big(\frac{X_1}{\bar{X}} \frac{Y_1}{\bar{Y}}\Big) g\Big(\frac{X_1}{\bar{X}} \frac{Y_2}{\bar{Y}}\Big) = \mathbb{E} g(X_1 Y_1) g(X_1 Y_2). \tag{16}$$

Introduce the function $h_n : \mathbb{R}^{+*} \to \mathbb{R}$

$$t \to h_n(t) = \mathbb{E}g(t\bar{X}\bar{Y}).$$

Lemma 1 gives

$$g(xy)h_n(xv)(1-x/n)^{n-2}(1-(y+v)/n)^{n-3} \xrightarrow[n\to\infty]{} g(xy)g(xv)e^{-x}e^{-y}e^{-v}.$$

Denote $t_1 \in \mathbb{R}^{+*}$. From (5), one gets

$$\exists \psi : \mathbb{R} \to \mathbb{R}, \forall t > t_1, \forall x \in \mathbb{R}^+, \Big|\frac{g(txy)}{g(t)}\Big| < \psi(x)$$

$$\begin{aligned} \Rightarrow \Big|\frac{h_n(t)}{g(t)}\Big| &= \Big|\int_0^\infty \int_0^\infty \frac{g(txy)}{g(t)} f_n(x,1) f_n(y,1) dy\, dx\Big| \\ &< \int_0^\infty \int_0^\infty \psi(xy) f_n(x,1) f_n(y,1) dy\, dx = \mathbb{E}\psi(\bar{X}\bar{Y}) = \Psi(n). \end{aligned}$$

So, $\forall (x,v) \in \mathbb{R}^{+2}, xv > t_1 \Rightarrow |g(xy)h_n(xv)(1-x/n)^{n-2}(1-(y+v)/n)^{n-3}| < |g(xy)g(xv)\Psi(n)e^5e^{-x}e^{-y}e^{-v}| \in \mathcal{L}^1([0,n]^2 \times [0,n-y])$.

From (4), one gets

$$\exists \varphi : \mathbb{R} \to \mathbb{R}, \forall t < t_1, \forall x \in \mathbb{R}^+, |g(tx)| < \varphi(x)$$

$$\Rightarrow \forall t < t_1, |h_n(t)| < \mathbb{E}\varphi(\bar{X}\bar{Y}) = \Phi(n).$$

So, $\forall (x,v) \in \mathbb{R}^{+2}, xv < t_1 \Rightarrow |g(xy)h_n(xv)(1-x/n)^{n-2}(1-(y+v)/n)^{n-3}| < |g(xy)\Phi(n)e^5e^{-x}e^{-y}e^{-v}| \in \mathcal{L}^1([0,n]^2 \times [0,n-y])$.

So applying Lebesgue's dominated-convergence theorem to (12) leads to

$$\lim_{n\to\infty} \mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_1Y_2)\Big\} = \mathbb{E}g(X_1Y_1)g(X_1Y_2). \tag{17}$$

(7), (16) and (17) lead to

$$\lim_{n\to\infty} T_{2,n} = 0.$$

1.3.2.4. Behaviour of $T_{3,n}$. — From (11), one gets

$$\begin{aligned}
&\frac{(n-1)^2}{n}\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g\Big(\frac{X_2}{\bar{X}}\frac{Y_2}{\bar{Y}}\Big)\Big\}\\
&= \frac{(n-1)^4(n-2)^2}{n^5}\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty g(xy)g(uv)\Big(1+\frac{3(x+u)-(x+u)^2/2}{n}\Big)\\
&\quad \Big(1+\frac{3(y+v)-(y+v)^2/2}{n}\Big)e^{-x}e^{-y}e^{-u}e^{-v}dv\,du\,dy\,dx + I_n\\
&= \frac{(n-1)^4(n-2)^2}{n^5}\mu^2 + 12\mu\frac{(n-1)^4(n-2)^2}{n^6}\mathbb{E}[X_1 g(X_1Y_1)] + \mathcal{O}(n^{-1}) \qquad (18)\\
&- 2\mu\frac{(n-1)^4(n-2)^2}{n^6}\mathbb{E}[X_1^2 g(X_1Y_1)] - 2\frac{(n-1)^4(n-2)^2}{n^6}\mathbb{E}^2[X_1 g(X_1Y_1)] + I_n
\end{aligned}$$

where $I_n =$

$$\frac{(n-1)^4(n-2)^2}{n^5}\Big\{\int_0^n\int_0^n\int_0^{n-x}\int_0^{n-y} g(xy)g(uv)\Big(1-\frac{x+u}{n}\Big)^{n-3}\Big(1-\frac{y+v}{n}\Big)^{n-3}$$
$$dv\,du\,dy\,dx - \int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty g(xy)g(uv)\Big(1+\frac{3(x+u)-(x+u)^2/2}{n}\Big)$$
$$\Big(1+\frac{3(y+v)-(y+v)^2/2}{n}\Big)e^{-x}e^{-y}e^{-u}e^{-v}dv\,du\,dy\,dx\Big\}.$$

From (13), one gets

$$\begin{aligned}
U_n &= \frac{(n-1)^2}{n}\mathbb{E}\Big\{g\Big(\frac{X_1}{\bar{X}}\frac{Y_1}{\bar{Y}}\Big)g(X_2Y_2)\Big\}\\
&= \frac{(n-1)^4(n-2)^2}{n^5}\int_0^n\int_0^n\int_0^{n-x}\int_0^{n-y}\Big(\int_0^\infty\int_0^\infty g(ab)f_n(a,u)f_n(b,v)da\,db\Big)\\
&\quad g(xy)\Big(1-\frac{x+u}{n}\Big)^{n-3}\Big(1-\frac{y+v}{n}\Big)^{n-3}dv\,du\,dy\,dx\\
&= \frac{(n-1)^4(n-2)^2}{n^5}\int_0^n\int_0^n\int_0^\infty\int_0^\infty g(xy)\Big\{\int_0^{n-x} f_n(a,u)\Big(1-\frac{x+u}{n}\Big)^{n-3}du\Big\}\\
&\quad g(ab)\Big\{\int_0^{n-y} f_n(b,v)\Big(1-\frac{y+v}{n}\Big)^{n-3}dv\Big\}db\,da\,dy\,dx.
\end{aligned}$$

Now the first integral in braces is equal to

$$\frac{n}{n-1}e^{-a}\Big(1-\frac{x}{n}\Big)^{n-3}e^{-ax/(n-x)}\Big(1+\frac{na}{(n-x)(n-2)}\Big)$$

$$
\begin{aligned}
\Rightarrow U_n = {} & \frac{(n-1)^2(n-2)^2}{n^3}\int_0^n\int_0^n\int_0^\infty\int_0^\infty g(xy)g(ab)e^{-a}\Big(1-\frac{x}{n}\Big)^{n-3} \\
& e^{-ax/(n-x)}\left(1+\frac{na}{(n-x)(n-2)}\right)e^{-b}\Big(1-\frac{y}{n}\Big)^{n-3}e^{-ay/(n-y)} \\
& \left(1+\frac{nb}{(n-y)(n-2)}\right)db\,da\,dy\,dx \\
= {} & \frac{(n-1)^2(n-2)^2}{n^3}\int_0^n\int_0^n\int_0^\infty\int_0^\infty g(xy)g(ab)e^{-a}\Big(1-\frac{x}{n}\Big)^{n-3} \\
& e^{-b}\Big(1-\frac{y}{n}\Big)^{n-3}\left(1-\frac{ax}{n-x}+\frac{na}{(n-x)(n-2)}\right) \\
& \left(1-\frac{by}{n-y}+\frac{nb}{(n-y)(n-2)}\right)db\,da\,dy\,dx + J_n
\end{aligned}
$$

where $J_n =$

$$
\begin{aligned}
& \frac{(n-1)^2(n-2)^2}{n^3}\int_0^n\int_0^n\int_0^\infty\int_0^\infty g(xy)g(ab)e^{-a}\Big(1-\frac{x}{n}\Big)^{n-3}e^{-b}\Big(1-\frac{y}{n}\Big)^{n-3} \\
& \left\{e^{-ax/(n-x)}e^{-by/(n-y)}\left(1+\frac{na}{(n-x)(n-2)}\right)\left(1+\frac{nb}{(n-y)(n-2)}\right)\right. \\
& \left.-\left(1-\frac{ax}{n-x}+\frac{na}{(n-x)(n-2)}\right)\left(1-\frac{by}{n-y}+\frac{nb}{(n-y)(n-2)}\right)\right\}db\,da\,dy\,dx.
\end{aligned}
$$

Substituting $\dfrac{ax}{n}\Big(1-\dfrac{x}{n}\Big)^{-1}$ to $\dfrac{ax}{n-x}$, one gets that U_n equals

$$
\begin{aligned}
& \frac{(n-1)^2(n-2)^2}{n^3}\mu^2 + \frac{6(n-1)^2(n-2)^2}{n^4}\mu\mathbb{E}[X_1g(X_1Y_1)] \\
- {} & \frac{2(n-1)^2(n-2)^2}{n^4}\mathbb{E}[X_1g(X_1Y_1)]\int_0^n\int_0^n xg(xy)\Big(1-\frac{x}{n}\Big)^{n-4}\Big(1-\frac{y}{n}\Big)^{n-3}dy\,dx \\
+ {} & \frac{2(n-1)^2(n-2)}{n^3}\mathbb{E}[X_1g(X_1Y_1)]\int_0^n\int_0^n g(xy)\Big(1-\frac{x}{n}\Big)^{n-4}\Big(1-\frac{y}{n}\Big)^{n-3}dy\,dx \\
- {} & \frac{(n-1)^2(n-2)^2}{n^4}\mu\mathbb{E}[X_1^2g(X_1Y_1)] + J_n + K_n + \mathcal{O}(n^{-1}) \qquad (19)
\end{aligned}
$$

where $K_n =$

$$\frac{(n-1)^2(n-2)^2}{n^3}\mu\Big[\int_0^n\int_0^n g(xy)\Big(1-\frac{x}{n}\Big)^{n-3}\Big(1-\frac{y}{n}\Big)^{n-3}dy\,dx$$
$$-\int_0^\infty\int_0^\infty g(xy)e^{-x}e^{-y}\Big(1+\frac{3x-x^2/2}{n}\Big)\Big(1+\frac{3y-y^2/2}{n}\Big)dy\,dx\Big].$$

So, from (8), (18) et (19) we get

$$T_{3,n} = A_n + \mathcal{O}(n^{-1}) + I_n + J_n + K_n - 2c^2 \tag{20}$$

where $A_n = (n-8)\mu^2 + 12\mu\mathbb{E}[X_1 g(X_1Y_1)] - 2\mu\mathbb{E}[X_1^2 g(X_1Y_1)] - 2\mathbb{E}^2[X_1 g(X_1Y_1)]$
$- 2(n-6)\mu^2 - 12\mu\mathbb{E}[X_1 g(X_1Y_1)] + 2\mu\mathbb{E}[X_1^2 g(X_1Y_1)] + (n-2)\mu^2$
$+ 4\mathbb{E}[X_1 g(X_1Y_1)]\int_0^n\int_0^n xg(xy)\Big(1-\frac{x}{n}\Big)^{n-4}\Big(1-\frac{y}{n}\Big)^{n-3}dy\,dx$
$- 4\mathbb{E}[X_1 g(X_1Y_1)]\int_0^n\int_0^n g(xy)\Big(1-\frac{x}{n}\Big)^{n-4}\Big(1-\frac{y}{n}\Big)^{n-3}dy\,dx + \mathcal{O}(n^{-1})$

$$\Rightarrow A_n \xrightarrow[n\to\infty]{} 2\mu^2 - 4\mu\mathbb{E}[X_1 g(X_1Y_1)] + 2\mathbb{E}^2[X_1 g(X_1Y_1)]$$
$$\Rightarrow A_n \xrightarrow[n\to\infty]{} 2c^2. \tag{21}$$

It now suffices to show that $I_n + J_n + K_n = o(1)$.

A Taylor-expansion leads to

$$\forall x \in [1,\sqrt{n}] : \Big(1-\frac{x}{n}\Big)^{n-3} = e^{-x}\Big[1+\frac{3x-x^2/2}{n}+\mathcal{O}\big(n^{-2}(x^2+x^4)\big)\Big]. \tag{22}$$

Denote $b_n = 4\log n$. Choosing n large enough gives

$$\forall x \in [b_n, n], \Big(1-\frac{x}{n}\Big)^{n-3} < e^{-x}. \tag{23}$$

By Cauchy-Schwarz inequality one gets

$$\begin{aligned}\frac{n^3}{(n-1)^2(n-2)^2\mu}|K_n| &= \Big|\int_0^\infty\int_0^\infty g(xy)\Big[\Big\{\Big(1-\frac{x}{n}\Big)^+\Big(1-\frac{y}{n}\Big)^+\Big\}^{n-3}\\ &\quad - e^{-x}e^{-y}\Big(1+\frac{3x-x^2/2}{n}\Big)\Big(1+\frac{3y-y^2/2}{n}\Big)\Big]dy\,dx\Big|\\ &\le [\mathbb{E}g^2(X_1Y_1)]^{1/2}\Bigg[\int_0^\infty\int_0^\infty\Big[\Big\{\Big(1-\frac{x}{n}\Big)^+\Big(1-\frac{y}{n}\Big)^+\Big\}^{n-3}\\ &\quad - e^{-x}e^{-y}\Big(1+\frac{3x-x^2/2}{n}\Big)\Big(1+\frac{3y-y^2/2}{n}\Big)\Big]^2e^xe^ydy\,dx\Bigg]^{1/2}\end{aligned}$$

where $x^+ = x$ if $x > 0$, 0 elsewhere.

Denote E_n as the double integral

$$\begin{aligned}&\int_0^\infty\int_0^\infty\Big[\Big\{\Big(1-\frac{x}{n}\Big)^+\Big(1-\frac{y}{n}\Big)^+\Big\}^{n-3} - e^{-x}e^{-y}\\ &\Big(1+\frac{3x-x^2/2}{n}\Big)\Big(1+\frac{3y-y^2/2}{n}\Big)\Big]^2e^xe^ydy\,dx.\end{aligned}$$

Using (22) and (23) and following the technique of Does & Klaassen (1984), one can prove : $E_n = \mathcal{O}(n^{-4})$

$$\Rightarrow K_n = \mathcal{O}(n^{-1}). \tag{24}$$

The same arguments are used for I_n

$$\Rightarrow I_n = \mathcal{O}(n^{-1}). \tag{25}$$

Using the inequalities $1 - z \ge e^{-z} \ge 1 - z + z^2/2$, $z \ge 0$, the expression in braces in the definition of J_n is bounded below by

$$\begin{aligned}&-\frac{a^2nx}{(n-x)^2(n-2)} - \frac{n^2a^2bx}{(n-2)^2(n-x)^2(n-y)}\\ -\ &\frac{b^2ny}{(n-y)^2(n-2)} - \frac{n^2b^2ay}{(n-2)^2(n-y)^2(n-x)}\end{aligned}$$

and bounded above by

$$
\begin{aligned}
& \frac{a^2x^2}{2(n-x)^2} + \frac{na^2bx^2}{2(n-2)(n-x)^2(n-y)} \\
+ & \frac{a^2b^2x^2y^2}{4(n-x)^2(n-y)^2} + \frac{na^2b^3x^2y^2}{4(n-2)(n-x)^2(n-y)^3} \\
+ & \frac{na^3x^2}{2(n-2)(n-x)^3} + \frac{n^2a^3bx^2}{2(n-2)^2(n-x)^3(n-y)} \\
+ & \frac{na^3b^2x^2y^2}{4(n-2)(n-x)^3(n-y)^2} + \frac{n^2a^3b^3x^2y^2}{4(n-2)^2(n-x)^3(n-y)^3} \\
+ & \frac{b^2y^2}{2(n-y)^2} + \frac{nb^2ay^2}{2(n-2)(n-y)^2(n-x)} \\
+ & \frac{b^2a^2y^2x^2}{4(n-y)^2(n-x)^2} + \frac{nb^2a^3y^2x^2}{4(n-2)(n-y)^2(n-x)^3} \\
+ & \frac{nb^3y^2}{2(n-2)(n-y)^3} + \frac{n^2b^3ay^2}{2(n-2)^2(n-y)^3(n-x)} \\
+ & \frac{nb^3a^2y^2x^2}{4(n-2)(n-y)^3(n-x)^2} + \frac{n^2b^3a^3y^2x^2}{4(n-2)^2(n-y)^3(n-x)^3}\,.
\end{aligned}
$$

Hence one gets

$$
\begin{aligned}
\frac{n^3}{(n-1)^2(n-2)^2}|J_n| \;\leq\; & \frac{1}{2n^2}\int_0^n\int_0^n\int_0^\infty\int_0^\infty |g(xy)g(ab)|a^2x^2e^{-a}e^{-b} \\
& \left(1-\frac{x}{n}\right)^{n-5}\left(1-\frac{y}{n}\right)^{n-3} db\,da\,dy\,dx \\
+ \;& \cdots \\
+ \;& \frac{1}{n(n-2)^2}\int_0^n\int_0^n\int_0^\infty\int_0^\infty |g(xy)g(ab)|ab^2ye^{-a}e^{-b} \\
& \left(1-\frac{x}{n}\right)^{n-4}\left(1-\frac{y}{n}\right)^{n-5} db\,da\,dy\,dx
\end{aligned}
$$

$$
\Rightarrow J_n = \mathcal{O}(n^{-1}). \tag{26}
$$

(20), (21), (24), (25) and (26) lead to

$$
\lim_{n\to\infty} T_{3,n} = 0.
$$

CHAPITRE 2

SPACINGS-BASED TESTS FOR SPATIAL RANDOMNESS

Abstract

We examine tests for the Complete Spatial Randomness (CSR) hypothesis of a point pattern in $\mathbb{R}^2$, based on functions of the spacings between x-ordinates and the spacings between y-ordinates. These tests extend to dimension two the one-dimensional uniformity spacings-based tests. We propose a multiple procedure based on the Rosenblatt transformation to break free from the coordinate system dependence. A real example and a simulation study show that the multiple spacings-based tests are inferior to existing tests for detecting regularity or clustering but more powerful for detecting certain types of heterogeneity, and that the multiple procedure increases the power of many other tests.

2.1. Introduction

When dealing with a spatial point pattern $U \in S \subset \mathbb{R}^d$, one first wishes to know whether it satisfies the CSR hypothesis : is the spatial process governing U a homogeneous Poisson process ? For a single realization, this question can be reformulated as : given the number of points in the pattern (also called events), are these points uniformly and independently distributed in S (Moller & Waagepetersen, 2004) ?

We concentrate at first on point patterns distributed in rectangles in $\mathbb{R}^2$, which is similar, after linear transformation of the coordinates, to testing uniformity in $[0, 1]^2$.

Historically, the first CSR tests were Chi-square tests applied to quadrat counts, i.e. the number of events in disjoint cells. Then appeared many methods based on various distance measurements between events or between sampled points and the nearest event (Cressie, 1993).

More recently, there has been some interest in tests based on the empirical distribution function (e.d.f.). An extension of the Cramer-Von Mises test to the $[0,1]^2$ case has been established and used as a CSR test (Zimmerman, 1993). Likewise, Justel & *al.* (1997) have generalized the Kolmogorov-Smirnov test to the bidimensional case : it has been used for testing normality, not yet as a CSR test.

As noted by Deheuvels (1983), the tests based on spacings are very useful to assess the goodness-of-fit of real random variables even if, theoretically, their performance is poor against standard contiguous alternatives (Guttorp & Lockhart, 1988). The extension of the theory of spacings to the bidimensional case can provide original techniques to test for CSR. Following this idea, Cucala (2005) extends the notion of spacings to dimension two and derives the limiting distribution of test statistics based on these bidimensional spacings.

In this paper, using these results, we build two spacings-based statistics generalizing the well-known Greenwood statistic and Sherman statistic (Pyke, 1965) to the $[0,1]^2$ case. In contrast with some existing tests for CSR, these statistics are computationally simple and do not need to adjust for edge effects.

In Section 2 we introduce the statistics and discuss their null asymptotic distributions. Then we underline the need for a multiple procedure and describe the selected tests in Section 3. The power of the spacings-based tests are compared to the power of some existing tests for CSR using several real data sets in Section 4 and using simulated data sets in Section 5. Finally, in Section 6 we present some possible extensions.

2.2. The test statistics and their null distributions in $[0,1]$

Let $\mathsf{U} = \left((U_1^x, U_1^y), \cdots, (U_{n-1}^x, U_{n-1}^y)\right)$ be a point pattern in $[0,1]^2$.

We first recall the definition of the two-dimensional spacings as defined in Cucala (2005)

$$\forall i \in \{1, \cdots, n\}, \forall j \in \{1, \cdots, n\}, \quad A_{ij} = D_i^x D_j^y$$

where $(D_i^x, i = 1, \cdots, n)$ are the first-order spacings related to the x-ordinates of U and $(D_j^y, j = 1, \cdots, n)$ are the first-order spacings related to the y-ordinates of U.

As the domain is rectangular, the $n-1$ points partition it into n^2 rectangles by drawing horizontal and vertical lines through each point. As the domain has area 1, the expected size of each such rectangle under the uniformity hypothesis is n^{-2} so the tests will proceed by comparing some function g of the area of each rectangle times n^2 to $g(1)$.

As stated by Rao Jammalamadaka and Goria (2004), a test of one-dimensional uniformity based on spacings should correspond to a dispersion measure of these spacings. Two of the first dispersion measures that have been used for this purpose are the variance, leading to the statistic introduced by Greenwood (1946), and the absolute mean deviation, leading to the statistic introduced by Sherman (1950). Similarly, we shall use these dispersion measures to build the following two statistics for testing bidimensional uniformity

$$V_n = \frac{V_n^* - \mathbb{E}V_n^*}{n^{3/2}} \text{ where } V_n^* = \sum_{i=1}^{n}\sum_{j=1}^{n}(n^2 A_{ij} - 1)^2,$$

$$R_n = \frac{R_n^* - \mathbb{E}R_n^*}{n^{3/2}} \text{ where } R_n^* = \sum_{i=1}^{n}\sum_{j=1}^{n}|n^2 A_{ij} - 1|.$$

Under CSR hypothesis, the distributions of these statistics are unknown but their limiting distributions can be derived from Cucala (2005).

We introduce E_1, E_2, E_3 three independent exponentially distributed random variables with mean 1.

Denote

$$\begin{aligned} g_1(t) &= (t-1)^2, \\ \mu_1 &= \mathbb{E}[g_1(E_1E_2)] = 3, \\ \eta_1 &= Cov\big(g_1(E_1E_2), g_1(E_1E_3)\big) = 52, \\ c_1 &= Cov\big(g_1(E_1E_2), E_1\big) = 6, \\ \sigma_1^2 &= 2(\eta_1 - c_1^2) = 32. \end{aligned}$$

As the function g_1 satisfies the required hypotheses (Cucala, 2005), we get the asymptotic distribution of V_n under CSR.

Lemma 1. —

$$V_n = \frac{1}{n^{3/2}}\{V_n^* - 3n^2\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 32).$$

Denote

$$\begin{aligned} g_2(t) &= |t-1|, \\ \mu_2 &= \mathbb{E}[g_2(E_1E_2)] = 4BK(0,2) + 4BK(1,2) \simeq 1.015, \\ \eta_2 &= Cov\big(g_2(E_1E_2), g_2(E_1E_3)\big) = 1 - 8BK(0,2) - 16BK(1,2) \\ &+ 32BK(0,2\sqrt{2}) + 32\sqrt{2}BK(1,2\sqrt{2}) - 16(BK(0,2) + BK(1,2))^2, \\ c_2 &= Cov\big(g_2(E_1E_2), E_1\big) = 8BK(1,2) - 4BK(0,2) - 1, \\ \sigma_2^2 &= 2(\eta_2 - c_2^2) \simeq 0.1634, \\ t &\to BK(\nu,t) \text{ is the modified Bessel function of the second kind with order } \nu. \end{aligned}$$

As the function g_2 satisfies the required hypotheses (Cucala, 2005), we get the asymptotic distribution of R_n under CSR.

Lemma 2. —

$$R_n = \frac{1}{n^{3/2}}\{R_n^* - \mu_2 n^2\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma_2^2).$$

However we would like to assess the quality of these asymptotic distributions by comparing them to the empirical distributions of V_n and R_n for large n. To evaluate this, we generate 10000 samples of size $n-1$ uniformly in $[0,1]^2$ and we compute the associated values of V_n and R_n. Tables 1 and 2 give estimated percentiles of V_n and R_n for different values of n together with percentiles of the limiting distribution.

TABLE 1. Selected percentiles of the distribution of V_n

$P(V_n \leq x)$	Values of x for the following values of n			
	$n = 10$	$n = 50$	$n = 100$	$n = \infty$
0.01	-7.309	-9.697	-10.788	-13.52
0.02	-6.985	-8.984	-9.71	-11.618
0.05	-6.448	-7.895	-8.229	-9.305
0.95	4.92	8.363	9.0976	9.305
0.98	8.317	12.476	12.558	11.618
0.99	10.801	15.431	15.202	13.52

TABLE 2. Selected percentiles of the distribution of R_n

$P(R_n \leq x)$	Values of x for the following values of n			
	$n = 10$	$n = 50$	$n = 100$	$n = \infty$
0.01	-1.111	-1.019	-1.007	-0.94
0.02	-0.998	-0.911	-0.881	-0.83
0.05	-0.825	-0.737	-0.71	-0.665
0.95	0.542	0.611	0.621	0.665
0.98	0.709	0.77	0.776	0.83
0.99	0.821	0.866	0.897	0.94

These results show that, for sample sizes commonly encountered in practice, simulated empirical percentiles should be preferred to the asymptotic distribution for both V_n and R_n. As noted by Gatto & Jammalamadaka (1999), even one-dimensional spacings-based statistics usually present this drawback. In their paper, these authors present a saddlepoint method leading to a very accurate approximation of the distribution of this type of statistics. This method cannot be directly applied to the two-dimensional spacings-based statistics but an extension could be considered in a future work.

2.3. The multiple test procedure in a general domain

Compared to the distance-based methods, the main drawback of the e.d.f. or spacings-based methods is that the result of the test may be very sensitive to the particular x- and y-axes that have been chosen. As an example, one could imagine a simulated point pattern U on the unit square where the x-ordinates $\mathsf{U}^{\mathsf{x}} = \{U_1^x, \cdots, U_{n-1}^x\}$ are independently and uniformly distributed

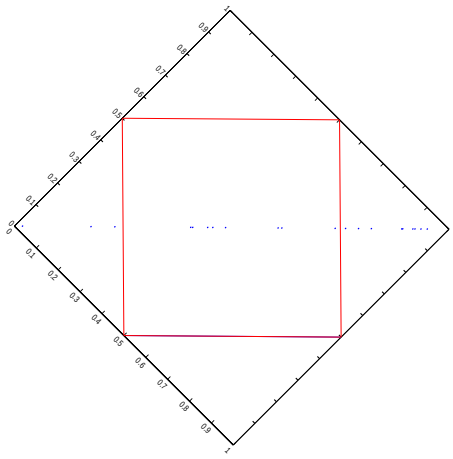

FIGURE 1. The need for rotation

on $[0,1]$ and the y-ordinates $\mathsf{U}^{\mathsf{y}} = \{U_1^y, \cdots, U_{n-1}^y\}$ are equal to the x-ordinates : $\mathsf{U}^{\mathsf{x}} = \mathsf{U}^{\mathsf{y}}$. Both statistics V_n and R_n do not depart from their values under CSR hypothesis. The solution can be to rotate the axes by an angle of $\pi/4$ and to compute V_n and R_n based on the new ordinates. As illustrated in Figure 1, most of the y-spacings are then null and the values of the statistics always lead to the rejection of CSR.

The other important drawback of the e.d.f. and spacings-based methods is the following : in order that the problem be the same as in the unit square, the domain D where the data have been collected must be rectangular. When D is not rectangular, a few coordinate transformations can be applied (D'Agostino & Stephens, 1986). We chose to use the probability integral transformation due to Rosenblatt (1952) : the coordinates (U_i^x, U_i^y) of each event are transformed into (Z_i^x, Z_i^y) where $Z_i^x = \nu\big(D \cap \{(U^x, U^y) : U^x < U_i^x\}\big)/\nu(D)$ and $Z_i^y = \nu\big(D \cap \{(U^x, U^y) : U^x = U_i^x, U^y < U_i^y\}\big)/\nu\big(D \cap \{(U^x, U^y) : U^x = U_i^x\}\big)$,where ν represents the Lebesgue measure. If U is a realization of a homogeneous Poisson process in D, $\mathsf{Z} = \Big((Z_1^x, Z_1^y), \cdots, (Z_{n-1}^x, Z_{n-1}^y)\Big)$ will be a realization of a homogeneous Poisson process in $[0,1]^2$.

This transformation is geometrically illustrated on a single example in Figure 2. In fact, Z^x is just the volume of the hatched area divided by the volume of D, and Z^y is the length between A and U divided by the length between A and B.

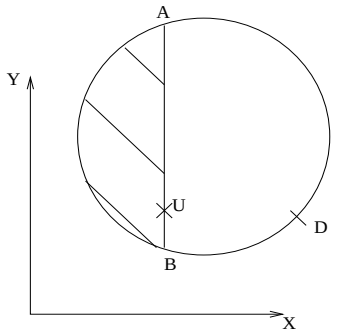

FIGURE 2. The Rosenblatt transformation

Thus, in order to deal with the defects mentioned, we now introduce a multiple step procedure based on the statistic V_n in order to test the CSR hypothesis for a $(n-1)$-point pattern U in any domain D (not necessarily rectangular). Let $a \in \mathbb{N}$ be the number of rotations. For each of the a rotations, let ω be the angle between the x- and y-axes of the new system of coordinates and the x- and y-axes of our original system of coordinates. We initialize ω to 0, $V_{n,max}$ to $-\infty$ and $V_{n,min}$ to ∞.

1) We consider the axes X_ω and Y_ω forming an angle of ω with X and Y and we set the coordinates of the $n-1$ events in this new system of coordinates (the origin remains the same).

2) The coordinates $(U^x_{i,\omega}, U^y_{i,\omega})$ of each event are transformed into $(Z^x_{i,\omega}, Z^y_{i,\omega})$ according to the Rosenblatt transformation described earlier.

3) $\mathsf{Z}_\omega = \left((Z^x_{1,\omega}, Z^y_{1,\omega}), \cdots, (Z^x_{n-1,\omega}, Z^y_{n-1,\omega})\right)$ is a point pattern in $[0,1]^2$ so we can compute the associated statistic $V_{n,\omega}$.

4) $V_{n,max} = \max(V_{n,\omega}, V_{n,max})$ and $V_{n,min} = \min(V_{n,\omega}, V_{n,min})$. Then set $\omega = \omega + \pi/(2a)$.

This procedure is repeated until $\omega = \pi$. Indeed, the statistic $V_{n,\omega+\pi}$ will be the same as $V_{n,\omega}$.

From the a statistics associated to each angle $(V_{n,0}, \cdots, V_{n,\pi-\pi/a})$, it would be possible to use a multiple procedure controlling the family-wise error rate (FWER) i.e. the probability that the CSR hypothesis is rejected for any of the a investigated angles. As we have no information about the correlation

between the a available statistics, we could think for example of the classical Bonferroni procedure or the Simes (1986) procedure. The control of the false discovery rate (Benjamini & Hochberg, 1995) is not relevant here as we are not interested in the proportion of angles leading to CSR rejection.

However, as the exact distribution of the initial statistics is unknown and the multiple procedures mentioned before are often conservative, we decided to focus on the minimum and the maximum of the a statistics and to compare them with their empirical quantiles under CSR obtained by simulation.

Thus, we adopt the following decision rule for the α-significance level test. We reject the CSR hypothesis if and only if
- $V_{n,min}$ is smaller than its empirical $\alpha/2$ quantile or
- $V_{n,max}$ is greater than its empirical $1-\alpha/2$ quantile.

Of course, this procedure is also valid when using the statistic R_n or any statistic sensitive to the system of coordinates, for example the e.d.f.-based statistics.

2.4. Examples

We now compute the values of our new statistics associated to four data sets (from Zimmerman, 1993) and compare them with two e.d.f.-based statistics, $\bar{\omega}^2$ and D_n, and three distance-based statistics, T, Li and L_m.

Traditionnally, the four data sets, Japanese pines, Redwoods, Biological cells and Scouring rushes, are respectively considered as random, aggregated, regular and heterogeneous.

The statistics $\bar{\omega}^2$ (Zimmermann, 1993) and D_n (Justel & *al.*, 1997) are respectively the bivariate extensions of the one-dimensional Cramer-Von Mises statistic and the one-dimensional Kolmogorov-Smirnov statistic. They are both origin-invariant but depend on the coordinate system, and explicit formulas are available.

The statistic T is the standardized empirical mean of the nearest neighbour distances (distances from an event to its nearest neighbour). Under CSR, its distribution is approximately Gaussian (Donnelly, 1978).

The statistic Li is based on the differences between coordinates of all pairs of events (Liebetrau, 1977).

The well-known function $K(t)$ can be interpreted as the expected number of events within a distance t of a randomly chosen event, divided by the average intensity. Under CSR, $K(t) = \pi t^2$. The statistic L_m (Diggle, 1983) measures the discrepancy between Ripley's (1976) estimator of $K(t)$ and the theoretical formula under CSR.

For more details about these statistics, see Zimmerman (1993) and Justel & *al.* (1997). It should be noticed that all these statistics are used within two-tailed tests so as to detect both aggregation and regularity.

Table 3 gives the obtained significance levels. For the T statistic, the levels are computed using the theoretical asymptotic normal distribution. For the others, we use a Monte Carlo procedure with a number of simulated patterns of 99 for the more computationnally demanding L_m and all the multiple procedures, and 999 for V_n, R_n, $\bar{\omega}^2$, D_n and Li. For the multiple procedure MV_n associated to V_n, the significance value is taken to be $2\min(p_{min}, p_{max}) \wedge 1$, where p_{min} (resp. p_{max}) is the p-value associated to $V_{n,min}$ (resp. $V_{n,max}$). Same for the multiple procedures MR_n, $M\bar{\omega}^2$, MD_n and MLi respectively associated to R_n, $\bar{\omega}^2$, D_n and Li.

TABLE 3. Attained significance levels for various tests of CSR

Test statistic	Results for the following data sets			
	Japanese pines	Redwoods	Biological cells	Scouring rushes
V_n	< 0.002	0.06	0.092	0.044
MV_n	0.04	< 0.02	0.04	0.04
R_n	< 0.002	< 0.002	0.768	0.036
MR_n	< 0.02	< 0.02	0.20	0.16
$\bar{\omega}^2$	0.660	0.694	0.006	0.002
$M\bar{\omega}^2$	0.72	0.4	0.004	< 0.02
D_n	0.272	0.914	0.012	0.058
MD_n	1	0.44	0.04	< 0.02
Li	0.782	< 0.002	0.002	0.122
MLi	0.52	< 0.02	0.04	0.40
L_m	0.88	< 0.02	< 0.01	0.36
T	0.915	< 0.002	< 0.002	0.936

The spacings-based statistics behave differently according to the dispersion measure. The variance V_n does not seem very useful to identify aggregation whereas the absolute mean deviation seems insensitive to regurality. They are quite efficient to detect the heterogeneity of the Scouring rushes. But the most striking are the significance levels obtained by the Japanese pines data set. It clearly seems that this data set, which is considered to be completely random by most authors, has a specific structure underlined by the spacings-based statistics. Indeed, when looking carefully at the data set, it appears that many points have very close x or y-ordinates. As mentioned by Stoyan & Stoyan (1994), it may come from the fact that these trees were planted many years ago as a regular grid, and this regularity becomes less precise generation after generation. The spacings-based statistics are the only one to detect this problem so we may think they are more sensitive to points which are gathered around lines parallel to the x or the y-axis. This type of behaviour is also exhibited by an inhomogeneous Poisson process with intensity $\lambda(x, y) = \lambda_1(x)\lambda_2(y)$ and

$$\lambda_1(x) = \max_{x_1,\cdots,x_m} \exp(-c_x|x - x_i|),$$

$$\lambda_2(y) = \max_{y_1,\cdots,y_l} \exp(-c_y|y - y_j|).$$

We will call it a grid-based heterogeneous Poisson process, where $\bar{x} = [x_1, \cdots, x_m]$ and $\bar{y} = [y_1, \cdots, y_l]$ are the x- and y-attraction vectors, which may represent the original plantation grid, and c_x and c_y the x- and y-attraction intensities, which may depend on the age of the forest. Figure 3 shows a realization of such a process.

Similarly we may also introduce the grid-based heterogeneous Poisson process with angle ω, whose attraction vectors form an angle of ω with the x- and y-axes.

2.5. Simulation study

In order to check the observations made on the real data sets, we compute the empirical powers of the tests based on the statistics mentioned before against four different types of processes : a simple sequential inhibition process (SSIP), representing a regular alternative to CSR ; a Poisson cluster process (PCP), representing an aggregated alternative to CSR ; an inhomogeneous

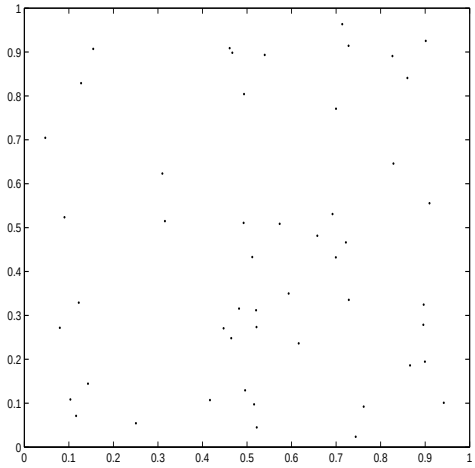

FIGURE 3. A realization of a grid-based heterogeneous Poisson process in the unit square with $\bar{x} = \bar{y} = [0.1, 0.3, 0.5, 0.7, 0.9]$ and $c_x = c_y = 20$.

planar trend Poisson process (IPTPP), as defined by Zimmerman (1993), and a grid-based heterogeneous Poisson process, as defined in the previous section. A brief description of these processes and methods for simulating them can be found below and more details are available in Diggle (1983).

The critical values of the spacings-based tests are computed using the empirical quantiles obtained beforehand as in Section 2. The multiple procedure applied to V_n, R_n, $\bar{\omega}^2$, D_n and Li are described in Section 3. We then derive the powers of the respecting tests by simulating 1000 independent sets of 50 points according to each process.

Table 4 gives the empirical powers obtained for SSIPs with various minimum interevent distances ϵ : events are simulated according to an uniform distribution in $[0, 1]^2$ but not retained if the distance to any previous event is less than ϵ.

Table 5 gives the empirical powers obtained for PCPs with various parameter values μ, ρ and t. The number of clusters follows a Poisson distribution with mean ρ and the cluster centres are uniformly distributed in $[0, 1]^2$. Then, for each cluster, the number of events follows a Poisson distribution with mean μ and these events are uniformly distributed in a circle of radius t.

Table 6 gives the empirical powers obtained for IPTPPs with various parameter values θ_1 and θ_2. The intensity linearly depends on the x- and y-ordinates with respective trends θ_1 and θ_2. The events are simulated by an acceptance-rejection method (Gentle, 2002). We can also imagine an IPTPP with linear trends forming an angle ω with the x- and y- axes. Table 7 gives the results obtained for such processes with an angle ω of $\pi/5$.

For each of these processes we only report the powers of the tests derived from the spacings-based statistics V_n and R_n and the e.d.f.-based statistics $\bar{\omega}^2$ and D_n, and the most powerful of the other methods mentioned earlier. The results clearly indicate that the spacings-based methods are much weaker than the distance-based tests against regularity and clustering, and much weaker than the e.d.f.-based tests against planar trend inhomogeneity. We can also remark that the Kolmogorov-Smirnov statistic D_n, which had not been used for testing CSR till now, has the same characteristics as the Cramer-Von Mises statistic $\bar{\omega}^2$ but its power is slightly inferior. Moreover, we find out that the multiple procedure issued from the statistic Li is the most powerful against both regularity and clustering. Indeed, it obtains slightly better results than the test only based on the statistic Li depending from the original system of coordinates, and than the distance-based tests L_m and T.

TABLE 4. Estimated power of tests for CSR against a simple sequential inhibition process in the unit square

ϵ	Estimated power of the following :								
	R_n	MR_n	V_n	MV_n	$\bar{\omega}^2$	$M\bar{\omega}^2$	D_n	MD_n	MLi
0.07	0.046	0.058	0.050	0.064	0.072	0.077	0.026	0.014	**0.387**
0.09	0.061	0.097	0.071	0.088	0.121	0.125	0.014	0.015	**0.817**
0.11	0.082	0.114	0.095	0.124	0.206	0.231	0.015	0.004	**0.993**

TABLE 5. Estimated power of tests for CSR against a Poisson cluster process in the unit square

μ	ρ	t	Estimated power of the following :								
			R_n	MR_n	V_n	MV_n	$\bar{\omega}^2$	$M\bar{\omega}^2$	D_n	MD_n	MLi
10	10	0.3	0.515	0.739	0.636	0.871	0.759	0.798	0.393	0.594	**0.954**
10	10	0.35	0.408	0.663	0.511	0.766	0.719	0.765	0.370	0.579	**0.907**
20	5	0.5	0.463	0.642	0.576	0.786	0.782	0.809	0.434	0.644	**0.851**

Results for the grid-based heterogeneous Poisson process, as defined earlier, are given in Table 8, when the angle ω is null, and in Table 9, with an angle

TABLE 6. Estimated power of tests for CSR against an inhomogeneous planar trend Poisson process in the unit square

θ_1	θ_2	Estimated power of the following :								
		R_n	MR_n	V_n	MV_n	$\bar{\omega}^2$	$M\bar{\omega}^2$	D_n	MD_n	L_m
4	4	0.080	0.104	0.097	0.157	0.621	**0.662**	0.351	0.559	0.187
6	6	0.107	0.168	0.136	0.237	0.811	**0.835**	0.468	0.743	0.262
8	8	0.121	0.210	0.194	0.312	0.896	**0.903**	0.564	0.811	0.334

TABLE 7. Estimated power of tests for CSR against an inhomogeneous planar trend Poisson process with an angle of $\pi/5$ in the unit square

θ_1	θ_2	Estimated power of the following :								
		R_n	MR_n	V_n	MV_n	$\bar{\omega}^2$	$M\bar{\omega}^2$	D_n	MD_n	L_m
4	4	0.082	0.123	0.106	0.151	0.625	**0.652**	0.539	0.586	0.175
6	6	0.086	0.159	0.134	0.245	0.815	**0.839**	0.712	0.777	0.250
8	8	0.109	0.248	0.173	0.348	0.896	**0.899**	0.791	0.848	0.333

ω of $\pi/5$. Here k represents the length of both the x- and y-attraction vectors $\bar{x} = \bar{y} = [1/2k, 3/2k, \cdots, (2k-1)/2k]$ and $c = c_x = c_y$ represents both the x and y-attraction intensities.

TABLE 8. Estimated power of tests for CSR against a grid-based heterogeneous Poisson process in the unit square

m	c	Estimated power of the following :								
		R_n	MR_n	V_n	MV_n	$\bar{\omega}^2$	$M\bar{\omega}^2$	D_n	MD_n	T
7	25	**0.667**	0.525	0.429	0.202	0.028	0.032	0.040	0.074	0.399
7	30	**0.889**	0.820	0.725	0.415	0.021	0.032	0.037	0.051	0.496
9	45	**0.967**	0.942	0.703	0.363	0.022	0.033	0.051	0.068	0.243

It appears that the spacings-based tests have no competitors for detecting grid-based heterogeneous Poisson processes. More precisely, when the direction of the grid can be suspected by an human eye, the test based on the statistic R_n is more powerful than any other, especially than the most powerful of the distance-based methods : T. On the other hand, when applying an automatic procedure, the multiple procedure based on R_n is of course less powerful than in the previous situation but still performs better than any other.

TABLE 9. Estimated power of tests for CSR against a grid-based heterogeneous Poisson process with an angle of $\pi/5$ in the unit square

m	c	Estimated power of the following :								
		R_n	MR_n	V_n	MV_n	$\bar{\omega}^2$	$M\bar{\omega}^2$	D_n	MD_n	T
7	25	0.059	**0.541**	0.055	0.221	0.023	0.029	0.091	0.080	0.364
7	30	0.046	**0.841**	0.032	0.395	0.029	0.033	0.087	0.069	0.481
9	45	0.053	**0.954**	0.440	0.375	0.044	0.038	0.071	0.054	0.220

2.6. Conclusion

As clustering and regularity are concepts closely linked to the distances between events, the tests that are directly based on these distances perform better than many against aggregated and regular alternatives. However, it appears that the statistic Li, an estimator of the variance function of the spatial point process, gives even better results and its use in a multiple procedure based on Rosenblatt transformation gets rid of the angle and domain shape restrictions.

On the other hand, detecting heterogeneity requires to observe the global characteristics of a point pattern such as the e.d.f. or the spacings' dispersion, so the tests based on these are more successful in detecting inhomogeneous alternatives. But the heterogeneity can take different shapes : Zimmerman (1993) defines the planar trend heterogeneity and proposes an appropriate statistic to detect it ; similarly we define the grid-based heterogeneity, and the statistics we introduce seem the most appropriate to detect for example whether a forest results from a human plantation. These spacings-based tests exhibit similarities with tests based on statistically equivalent blocks (Alam & *al.*, 1993) or directional tests introduced by Lawson (1988), but our method concentrates the information brought by the point pattern into two uniform spacings samples whereas the former methods concentrate it into a single uniform spacings sample.

As a conclusion, when studying a point pattern, one should always use a spacings-based method when distance-based or e.d.f.-based tests do not indicate departure from CSR. A significant departure would then let us suspect that the underlying process is a grid-based heterogeneous Poisson process. Its parameters can be estimated by a maximum-likelihood approach and the model validated by a residual approach (Baddeley & *al.*, 2005).

The spacings-based tests could also be extended to three-dimensional point patterns. In fact, one can define the three-dimensional spacings as the products of the spacings along the x-, y- and z-axes and, following the technique of Beirlant & *al.* (1991), Le Cam spacings theorem could certainly be extended to dimension three.

Finally, one could also think of adapting the same type of tests using high-order spacings, as it has been done by a few authors for dimension one (Cressie, 1976). The distribution theory increases in complexity but the power of the tests may also.

PARTIE II

DÉTECTION D'AGRÉGATS POUR DONNÉES PONCTUELLES

Cette partie décrit, dans un premier chapitre, des techniques d'identification d'agrégats adaptées à des données temporelles puis, dans un deuxième chapitre, l'extension de ces techniques à des données spatiales.

Dans la partie précédente, nous nous sommes intéressés à des tests dits globaux, permettant de tester l'hypothèse d'homogénéité spatiale contre toute autre alternative. Une des hypothèses alternatives fréquemment rencontrées est le phénomène d'agrégation, lorsque les événements ont tendance à se concentrer de manière anormalement élevée. Ce phénomène d'agrégation peut s'expliquer de deux différentes manières. La première consiste à dire que les événements sont indépendants mais leur apparition est fortement liée à une ou plusieurs covariables dont le niveau est plus élevé dans certaines zones. Par exemple, une certaine espèce végétale peut être plus présente sur des sols atteignant un certain degré d'hygrométrie, et donc aux alentours d'une source d'eau. Cela revient à penser que le processus ponctuel observé est un processus de Poisson inhomogène. La seconde explication consiste à dire que des événements non observés, dits "parents" donnent chacun naissance à plusieurs événements observés, tous rassemblés autour de leur parent. Par exemple, une plante peut produire des graines donnant naissance à plusieurs plantes dans la zone alentour. Cela revient à penser que le processus ponctuel observé est un processus de Cox. Malheureusement, lorsque l'on ne dispose que d'une seule réalisation, il est impossible de distinguer ces deux hypothèses alternatives (Cressie, 1993). Dans cette partie, nous n'opèrerons donc aucune distinction entre ces deux phénomènes d'agrégation.

Comme nous l'avons dit précédemment, les tests globaux permettent de rejeter l'hypothèse d'homogénéité spatiale mais il est souvent utile, lorsqu'on est en présence d'un phénomène d'agrégation, de détecter les zones de forte concentration, appelées agrégats, afin de mieux analyser les raisons d'une telle concentration. Lorsque l'on soupçonne une zone précise de contenir un nombre anormalement élevé d'événements, il existe des tests dits "focus" permettant de tester cette hypothèse. On peut citer les travaux de Lawson (1993) permettant de tester si l'apparition d'une maladie donnée est plus élevée autour d'une source de pollution. Mais dans la plupart des cas, nous n'avons pas d'idée préconçue sur les agrégats possibles et il faut mettre en place des procédures d'identification des agrégats. C'est ce que nous développons dans cette partie.

Enfin, notamment dans les études d'épidémiologie, il est souvent nécessaire de s'adapter à l'inhomogénéité (temporelle et/ou spatiale) de la population observée. En effet, la concentration de malades dans une certaine zone doit

être relativisée par rapport à la concentration de population dans cette même zone. Il est donc absolument indispensable de mettre en place des techniques prenant en compte cette inhomogénéité sous-jacente.

CHAPITRE 3

TEMPORAL CLUSTER DETECTION BASED ON SPACINGS

Abstract

In this chapter we propose new techniques for identifying clusters in temporal point processes. They rely on the analysis of unidimensional spacings and are independent of any alternative hypothesis. One of them is found to be more efficient than existing methods for identifying and recovering one-cluster alternatives.

3.1. Introduction

Let $X_1, \cdots, X_n$ be random variables which denote the times of occurrence of n events in an interval $[0, T]$. Without loss of generality, we set $T = 1$. The first objective of this work is to identify the zone in which events are most concentrated, usually named cluster. The second objective is to test whether the events are totally randomly distributed (i.e. independently and uniformly on $[0, 1]$), denoted as the null hypothesis H_0, against the alternative that they cluster within the predetermined zone. This problem arises naturally in epidemiological applications when one wants to assess the outbreak of an unknown disease and to analyse for example whether it is infectious or dependent on a seasonal environmental factor.

Many procedures are applicable to grouped data, i.e. when the observation period is divided into subintervals and only the number of events within each interval is known. See for example the test introduced by Tango (1984) relying on the division of the time interval in equal subintervals. These procedures

may also be used when individual data are available but the loss of information may be important and the test result be dependent on the arbitrarily chosen division.

The most popular of the methods for individual data is the scan statistic, first introduced by Naus (1965). Originally it was simply the maximum number of events observed within an interval of fixed length d. Later on, Nagarwalla (1996) extended the method so as to compare intervals having different lengths. The test is the generalized likelihood ratio test for a uniform null distribution against a piecewise constant alternative. We may wonder how powerful it is against clustering alternatives of a different kind, as bell-shaped densities that appear more realistic.

A method, due to Kelsall & Diggle (1995a), relies on a kernel intensity estimation of the point process and identifies the clusters as the intervals in which the intensity estimate is higher than expected under the null hypothesis. This technique is unfortunately very dependent on the bandwidth choice.

Many other tests, so called global tests, are powerful to reject the uniform hypothesis but do not identify the most significative cluster. A lot of these tests rely on spacings (Pyke, 1965). Recently, Molinari & *al.* (2001) introduced a cluster detection method based on applying a piecewise constant regression model to the spacings issued from the events' occurence times. This method allows for multiple cluster detection but do not take into account the spacings dependence and, as for the scan statistic, seems to be more adapted to piecewise constant cluster alternatives.

In this chapter, we introduce different cluster detection methods based on spacings and not relying on any particular alternative hypothesis. In a first part, we describe the data transformation process. Then we introduce the test statistic based on a concentration index allowing to compare all intervals and we describe how it can be adapted to the population inhomogeneity. Finally the test is applied to real and simulated data sets and compared to different existing techniques.

3.2. The data transformation

Denote $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1$ the order statistics associated to $(X_1, \cdots, X_n)$ and $D_i = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1$, the associated spacings. These spacings, giving an information about the closeness of two events, are all identically distributed under H_0 but dependent. In order to get rid of this dependence, we introduce the modified spacings

$$\begin{aligned} \forall i = 1, \cdots, n, \quad \tilde{D}_i &= F(D_i \,|D_1, \cdots, D_{i-1}) \\ &= F(X_{(i)} - X_{(i-1)} \,|X_{(i-1)}) = 1 - \Big(\frac{1 - X_{(i)}}{1 - X_{(i-1)}}\Big)^{n-i+1} \end{aligned}$$

where $F(.|D_1, \cdots, D_{i-1})$ represents the conditional distribution function of D_i given $D_1, \cdots, D_{i-1}$ under H_0. This conditional distribution function comes naturally from Theorem 2.7 by David (1981, p.18) stating that, under H_0, given $X_{(i-1)} = x_{(i-1)}$, $(X_{(i)}, \cdots, X_{(n)})$ have the same distribution as order statistics from a $(n - i + 1)$-sample uniformly distributed on $[x_{(i-1)}, 1]$. These modified spacings are now independent and uniformly distributed on $[0, 1]$ under H_0.

3.3. The test statistic

When applying the scan statistic with variable window (Nagarwalla, 1996), one identifies a cluster when too many events are concentrated on too small an interval. Here, we will identify a cluster when there is a concentration of exceedingly small intervals, that is when the sum of consecutive (modified or not) spacings will be too small. Denote $S^{(0)}_{j,k} = \sum_{i=j+1}^{k} D_i$ and $S^{(1)}_{j,k} = \sum_{i=j+1}^{k} \tilde{D}_i$. We may also rely on the modified spacings raised to the power 2 (so as to give more weight to little spacings) and introduce $S^{(2)}_{j,k} = \sum_{i=j+1}^{k} \tilde{D}_i^2$. In order to compare two intervals containing different numbers of events, we introduce the concentration indicators

$$\Lambda^{(l)}_{j,k} = F^{(l)}_{j,k}(S^{(l)}_{j,k})$$

where $F^{(l)}_{j,k}(.)$ represents the distribution function of $S^{(l)}_{j,k}$ under H_0, that we can identify for the different values of l.

3.3.1. The case $l = 0$. — Remark that $S^{(0)}_{j,k} = X_{(k)} - X_{(j)}$ and follows a Beta distribution as noted by David (1981, p.12). Its density is :

$$f^{(0)}_{j,k}(w) = \frac{1}{B(k-j, n-k+j+1)} w^{k-j-1}(1-w)^{n-k+j} \mathbb{1}(0 \leq w \leq 1)$$

and its distribution function is the incomplete Beta function :

$$F^{(0)}_{j,k}(s) = B_{inc}(s, k-j, n-k+j+1).$$

3.3.2. The case $l = 1$. — The density function of a sum of m independent $[0,1]$-uniform random variables is given by Mitra (1971) :

$$f_m(x) = \frac{1}{(m-1)!} \sum_{i=0}^{\lfloor x \rfloor} [C^i_m (x-i)^{m-1}] \mathbb{1}(0 \leq x \leq m)$$

and thus the distribution function of $S^{(1)}_{j,k}$ under H_0 is

$$\begin{aligned} F^{(1)}_{j,k}(t) &= \frac{1}{(k-j-1)!} \Big[\sum_{i=0}^{\lfloor t \rfloor} \big[(-1)^i C^i_{k-j} \frac{(t-i)^{k-j} - (\lfloor t \rfloor - i)^{k-j}}{k-j}\big] \\ &+ \sum_{j=0}^{\lfloor t \rfloor - 1} \sum_{i=0}^{\lfloor t \rfloor - 1 - j} \big[(-1)^i C^i_n \frac{(\lfloor t \rfloor - i - j)^{k-j} - (\lfloor t \rfloor - i - j - 1)^{k-j}}{k-j}\big] \Big]. \end{aligned}$$

3.3.3. The case $l = 2$. — We are looking for the distribution function of a sum of m independent $[0,1]$-uniform random variables raised to the power 2. When $m = 1$, one gets $\forall x \in [0,1], F_1(x) = \mathbb{P}(U^2 < x) = \mathbb{P}(U < \sqrt{x}) = \sqrt{x}$ where U stands for a $[0,1]$-uniform random variable. The corresponding density is $f_1(x) = \frac{1}{2\sqrt{x}} \mathbb{1}_{[0,1]}(x)$.

When $m = 2$, one gets the density using the relation

$$f_2(x) = \int_{-\infty}^{\infty} f_1(x-t) f_1(t) dt = \begin{cases} \pi/4 & \text{if } 0 \leq x \leq 1, \\ -1/2 \arcsin\left(\frac{x-2}{x}\right) & \text{if } 1 < x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

So the corresponding distribution function is

$$F_2(x) = \begin{cases} 0 & \text{if } x < 0, \\ \pi/4x & \text{if } 0 \leq x \leq 1, \\ \sqrt{x-1} - x/2 \arcsin\left(\frac{x-2}{x}\right) & \text{if } 1 < x \leq 2, \\ 1 & \text{if } x > 2. \end{cases}$$

Unfortunately, it seems that no analytical expression can be written for a general m except for the interval $[0,1]$. We will first prove that, $\forall m \geq 1, \forall x \in [0,1], f_m(x) = \left(\frac{\pi}{4}\right)^{m/2}/\Gamma(m/2)x^{m/2-1}$. Let us prove it using induction. The expression is true for $m = 1$. Then suppose it is true for a given m. We get

$$\begin{aligned} \forall x \in [0,1], f_{m+1}(x) &= \int_{-\infty}^{\infty} f_m(x-t) f_1(t) dt \\ &= \int_0^x \left(\frac{\pi}{4}\right)^{m/2}/\Gamma(m/2)(x-t)^{m/2-1} \frac{1}{2\sqrt{t}} dt \\ &= \left(\frac{\pi}{4}\right)^{m/2}/\Gamma(m/2)/2 \int_0^x \frac{(x-t)^{m/2-1}}{\sqrt{t}} dt \\ &= \left(\frac{\pi}{4}\right)^{m/2}/\Gamma(m/2) \frac{\sqrt{\pi}\Gamma(m/2)}{2\Gamma\left((m+1)/2\right)} x^{(m+1)/2-1} \\ &= \left(\frac{\pi}{4}\right)^{(m+1)/2}/\Gamma((m+1)/2) x^{(m+1)/2-1}. \end{aligned}$$

Thus the distribution function on the interval $[0,1]$ is

$$\forall x \in [0,1], F_{j,k}^{(2)}(t) = \frac{2\left(\frac{\pi}{4}\right)^{m/2}}{m\Gamma(m/2)} t^{m/2}.$$

Outside of this interval, we use an Edgeworth expansion to the third order as suggested by Brandt (1995) :

$$\forall x \in \mathcal{R}, F_{j,k}^{(2)}(t) \simeq \Psi(y) + \frac{1}{6\sqrt{2\pi}} \frac{16}{945} \left(\frac{4}{45}\right)^{-3/2} m^{-1/2}(1-y^2) \exp(-y^2/2)$$

where $y = \frac{t-m/3}{\sqrt{4m/45}}$ and $\Psi(.)$ stands for the distribution function of the standard gaussian distribution.

Whatever the value of l, we can compare all the intervals $[X_{(j)}, X_{(k)}]$ by comparing their concentration indicators $\Lambda_{j,k}^{(l)}$. The most significative cluster will be the one minimizing $\Lambda_{j,k}^{(l)}$ and we denote $\Lambda^{(l)} = \min_{1 \leq j < k \leq n} \Lambda_{j,k}^{(l)}$.

3.4. The adaptation to the population inhomogeneity

As said in the introduction of this part devoted to cluster detection, it is often necessary to take into account the inhomogeneity of the observed population. For example, if one observes the cases of a certain disease in a given hospital during a certain period of time, he may adapt the procedure to a highest population in the summer due to tourism (Molinari & *al.*, 2001).

This population inhomogeneity can be expressed through a function $g(t), t \in [0, 1]$ giving the amount of population at each time t. Then the adaptation is done by setting

$$\forall i = 1, \cdots, n+1, \quad D_i = \frac{\int_{X_{(i-1)}}^{X_{(i)}} g(t)dt}{\int_0^1 g(t)dt}.$$

Otherwise, it may be expressed through "controls", representing for example each time a sane people is observed, and denoted $\{c_1, \cdots, c_m\}$. Then the adaptation is done by setting

$$\forall i = 1, \cdots, n+1, \quad D_i = \frac{\sum_{j=1}^{m} \mathbb{1}(c_j \in [X_{(i-1)}, X_{(i)}])}{m}.$$

3.5. Data analysis

3.5.1. Real data. — We decide to apply our method for $l = 0, 1, 2$ together with the scan statistic with variable window (Nagarwalla, 1996) and the spacings regression method (Molinari & *al.*, 2001) to a classical data set published by Knox (1959) and describing birth defects observed in an hospital between 1950 and 1955. We observe 35 events and the observation days are scaled so that they lie in $[0, 1]$.

These data have already been analysed using the scan statistic by Nagarwalla, but limiting the observation to clusters containing at least $n_0 = 5$ events. The choice of this parameter n_0 is quite arbitrary and we want to observe its influence on results. First we set it to $n_0 = 5$ and we get the results recorded in Table 1, where the initials m.s.c. stand for "most significative cluster" i.e. the interval maximizing the concentration indicator.

TABLE 1. Tests applied to Knox data with $n_0 = 5$

Statistic	m.s.c.	observed value	empirical quantile
$\Lambda^{(0)}$	$[8, 22]$	$0.0201 * 10^{-3}$	**0.003**
$\Lambda^{(1)}$	$[8, 21]$	$1.536 * 10^{-3}$	0.070
$\Lambda^{(2)}$	$[9, 13]$	$1.343 * 10^{-3}$	0.079
Λ	$[8, 22]$	43968	**0.009**
$\Lambda^{(reg)}$	$[5, 35]$	$3.38 * 10^{-2}$	0.950

To assess the significance of the most likely cluster, one should know the distributions of the statistics Λ, $\Lambda^{(0)}$, $\Lambda^{(1)}$, $\Lambda^{(2)}$ and $\Lambda^{(reg)}$ under H_0. Unfortunately these are unknown so we choose to conduct a Monte-Carlo procedure to carry out the significance test. We simulate 999 uniform 35-samples on $[0, 1]$ and compute the observed value of the statistics. The last column of Table 1 records the empirical quantiles of the value of the statistic obtained from the data set among all the values obtained by simulation. Thus we conclude that only the scan statistic and our method using non-modified spacings undoubtedly lead to rejecting H_0 and the decision is much more difficult when using $\Lambda^{(1)}$ or $\Lambda^{(2)}$.

Now we decide to set $n_0 = 2$, that is to observe all possible clusters (as a single event can hardly be considered as a cluster). Then the results, recorded in Table 2, are slightly different.

TABLE 2. Tests applied to Knox data with $n_0 = 2$

Statistic	m.s.c.	observed value	empirical quantile
$\Lambda^{(0)}$	$[8, 22]$	$0.0201 * 10^{-3}$	**0.004**
$\Lambda^{(1)}$	$[8, 21]$	$1.536 * 10^{-3}$	0.151
$\Lambda^{(2)}$	$[9, 13]$	$1.343 * 10^{-3}$	0.167
Λ	$[8, 22]$	43968	0.135
$\Lambda^{(reg)}$	$[5, 35]$	$3.38 * 10^{-2}$	0.953

For all methods, the most significative cluster remains the same but only the test relying on $\Lambda^{(0)}$ leads to rejecting H_0. It seems that the test based on $\Lambda^{(0)}$ is the one that is less influenced by the choice of the parameter n_0 so that the decision based on this test is only dependent on data.

3.5.2. Simulated data. — The same methods are now applied to data sets containing 35 events simulated according to different one-cluster alternatives.

A flat cluster is obtained via the density function

$$f_1(x) = \begin{cases} \frac{r}{0.8+0.2r} & \text{if } 0.4 \leq x \leq 0.6, \\ \frac{1}{0.8+0.2r} & \text{if } x \in [0, 0.4] \cup [0.6, 1], \\ 0 & \text{otherwise.} \end{cases}$$

A bell-shaped cluster is obtained via the density function

$$f_2(x) = \begin{cases} \frac{10}{r+9}\{1 + (r-1) * [1 - 100(x-0.5)^2]\} & \text{if } 0.4 \leq x \leq 0.6, \\ \frac{10}{r+9} & \text{if } x \in [0, 0.4] \cup [0.6, 1], \\ 0 & \text{otherwise.} \end{cases}$$

In both cases, the parameter r is the ratio between the maximum density and the minimum density on the definition domain. Figure 1 represents these two families of density functions when r takes different values.

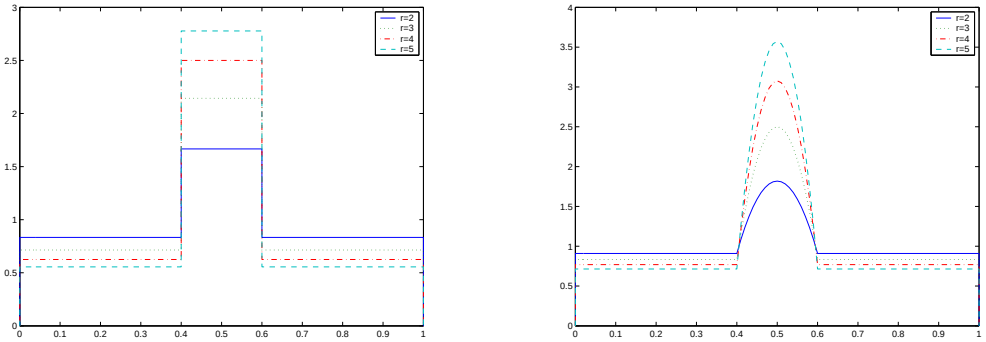

Figure 1 : Left figure : Profile of f_1. Right figure : Profile of f_2

We compare the power of the global tests associated to each method but moreover we want to check whether the significative clusters exhibited by the different methods match the "true" cluster, that is the interval in which the density is higher, $[0.4, 0.6]$ in our examples. When the global test leads to rejecting H_0, we set $\forall t \in [X_{(i)}, X_{(i+1)}]$, $\hat{c}(t) = 1$ if the interval $[X_{(i)}, X_{(i+1)}]$

belongs to the most significative cluster and 0 otherwise. When there is no significative cluster, we set $\forall t \in [0,1], \hat{c}(t) = 0$. We also set $c(t) = \mathbb{1}_{[0.4,0.6]}(t)$. A classification index, inspired by the Rand index (Saporta & Younnès, 2004) is

$$R = \int_0^1 \mathbb{1}\big(\hat{c}(t) = c(t)\big)dt$$

which gives the "percentage" of the segment length where the classification is correct.

For each alternative, 100 data sets are simulated. The nominal level of the global test is set to 5%. Tables 3 and 4 give the results obtained with the minimal cluster size $n_0 = 5$ and the two different cluster shapes. The percentages are the means of the Rand indexes recorded. Between parentheses are the empirical powers of the global tests associated to each method.

TABLE 3. Empirical power of tests for CSR against a flat cluster with $n_0 = 5$

r	Mean correlation index of the following :				
	$\Lambda^{(0)}$	$\Lambda^{(1)}$	$\Lambda^{(2)}$	Λ	$\Lambda^{(reg)}$
2	81.12 % (0.19)	80.87 % (0.12)	**81.17 % (0.17)**	80.45 % (0.18)	78.93 % (0.15)
3	**86.50 % (0.59)**	84.47 % (0.49)	84.00 % (0.53)	85.06 % (0.50)	79.10 % (0.09)
4	**88.95 % (0.83)**	86.28 % (0.68)	84.66 % (0.69)	88.14 % (0.81)	79.31 % (0.08)
5	**94.09 % (0.97)**	89.18 % (0.68)	86.68 % (0.89)	92.07 % (0.95)	80.97 % (0.15)
10	96.95 % (1)	92.61 % (0.99)	72.23 % (0.99)	**97.08 % (1)**	81.72 % (0.16)

TABLE 4. Empirical power of tests for CSR against a bell-shaped cluster with $n_0 = 5$

r	Estimated power of the following :				
	$\Lambda^{(0)}$	$\Lambda^{(1)}$	$\Lambda^{(2)}$	Λ	$\Lambda^{(reg)}$
2	**80.72 % (0.17)**	80.30 % (0.14)	80.52 % (0.13)	80.06 % (0.19)	79.82 % (0.09)
3	**82.41 % (0.29)**	81.43 % (0.21)	81.28 % (0.22)	81.10 % (0.22)	79.01 % (0.08)
4	**84.87 % (0.55)**	83.13 % (0.46)	82.52 % (0.45)	82.72 % (0.41)	79.08 % (0.08)
5	**88.37 % (0.78)**	85.50 % (0.66)	84.75 % (0.69)	85.43 % (0.61)	80.07 % (0.05)
10	**93.52 % (1)**	90.67 % (1)	76.69 % (1)	92.96 % (0.99)	80.82 % (0.13)

Table 5 gives the results obtained with the minimal cluster size $n_0 = 2$ (no cluster size constraint) and the bell-shaped cluster.

TABLE 5. Empirical power of tests for CSR against a bell-shaped cluster with $n_0 = 2$

r	Estimated power of the following :				
	$\Lambda^{(0)}$	$\Lambda^{(1)}$	$\Lambda^{(2)}$	Λ	$\Lambda^{(reg)}$
2	**81.21 % (0.21)**	81.08 % (0.18)	81.16 % (0.20)	80.13 % (0.16)	79.14 % (0.10)
3	**82.85 % (0.47)**	82.21 % (0.43)	82.22 % (0.45)	82.11 % (0.43)	79.65 % (0.12)
4	**85.97 % (0.54)**	84.50 % (0.54)	84.41 % (0.52)	84.66 % (0.53)	81.11 % (0.13)
5	**89.78 % (0.85)**	85.95 % (0.70)	84.23 % (0.72)	88.27 % (0.79)	80.19 % (0.09)
10	**93.52 % (1)**	90.82 % (1)	76.22 % (1)	93.39 % (1)	81.20 % (0.13)

Among the spacings-based methods, the one based on $\Lambda^{(0)}$ appears to be the most powerful and it seems that relying on independent modified spacings is not relevant. It also appears that raising the modified spacings to the power 2 does not lead to better results, even when the cluster is bell-shaped so that very little spacings should be observed.

The test based on the scan statistic Λ is almost always less powerful than the one based on $\Lambda^{(0)}$, both in terms of rejecting H_0 (as measured by the global power) and detecting the cluster (as measured by the Rand index) This ranking is much more evident when we do not set any restriction about the number of events in the clusters ($n_0 = 2$). We did expect such a result when the true cluster structure is different from the cluster structure the scan statistic is based on, but it is also true (even if the differences are less important) when the cluster structure is piecewise constant.

The spacings regression method does not seem very efficient in detecting one-cluster alternatives but it has the great advantage of adapting to multiple cluster detection. The scan statistic and the statistics we propose here are able to classify the possible clusters using a chosen significativity index and to test H_0 against a one-cluster alternative but it would be very useful to extend them to multiple clustering. This can be obtained by just adding a selection model criterion to our method and will be the subject of a future work.

CHAPITRE 4

SPATIAL CLUSTER DETECTION BASED ON SPACINGS

Abstract

In this chapter we propose new techniques for identifying clusters in spatial point processes. They rely on an ordering of the events introduced by Dematteï & *al.* (2005) and the introduction of area spacings having the same distribution as unidimensional spacings. The methods we propose are independent of any alternative hypothesis. Their efficiency is assessed using data sets simulated according to one-cluster alternatives.

4.1. Introduction

Let $X_1, \cdots, X_n$ be random variables which denote the spatial coordinates of the occurrence of n events in A, a bounded subset of $\mathbb{R}^2$. Without loss of generality, we set $|A| = 1$ where $||$ is the Lebesgue measure.The first objective of this work is to identify the zone in which events are most concentrated, usually named cluster. The second objective is to test whether the events are totally randomly distributed (i.e. independently and uniformly on A), denoted as the null hypothesis H_0, against the alternative that they cluster within the predetermined zone. This problem is the spatial version of the one studied in the preceding chapter.

Many procedures are applicable to grouped data, i.e. when the obsevation domain is divided into subdomains and only the number of events within each subdomain is known (Lawson, 2001). The division of the observation domain is often arbitrary and should not be used when individual locations are known.

Existing methods are often adapted from temporal cluster detection techniques. As in the temporal setting, the most popular is the scan statistic adapted to the spatial setting by Kulldorff (1997). It relies on the likelihood test statistic of H_0 against a piecewise-constant density alternative. To apply this method, one needs to define the possible clusters family, for example all discs centered on points of a predefined grid on A.

Let us also mention the method introduced by Kelsall & Diggle (1995b) based on kernel intensity which is the spatial adaptation to the method (Kelsall & Diggle, 1995a) mentioned in the preceding chapter.

Recently, Dematteï & *al.* (2005) set up a spatial version of a multiple temporal cluster detection technique introduced by Molinari & *al.* (2001). In this chapter, our objective is to adapt the technique introduced in the preceding chapter to the spatial setting. First, we describe the data transformation process leading to area spacings whose distribution is the same as unidimensional spacings. Introducing the test statistic is then straightforward and the adaptation to the population inhomogeneity differs very slightly from the temporal setting. Finally the test is applied to simulated data sets.

4.2. The data transformation

The starting point of this data transformation is the ordering of the events introduced by Dematteï & *al.* (2005). The first event, whose location is called $X_{(1)}$, is arbitrarily chosen as the closest one to the boundary of the observation domain A, denoted ∂A, using the euclidian distance. Denote $D_1 = d(X_{(1)}, \partial A)$, where $d(.,.)$ denotes the euclidian distance. Then the second event, whose location is called $X_{(2)}$ is the closest to the first one among all not yet ordered events :

$$X_{(2)} = argmin_{(X_i, i=1\cdots,n: X_i \neq X_{(1)})} d(X_i, X_{(1)}).$$

Denote $D_2 = d(X_{(1)}, X_{(2)})$. All events are then ordered similarly :

$$\forall j = 3, \cdots, n, \quad X_{(j)} = argmin_{(X_i, i=1\cdots,n: \forall k=1,\cdots,j-1,\ X_i \neq X_{(k)})} d(X_i, X_{(j-1)})$$

and $D_j = d(X_{(j-1)}, X_{(j)})$.

Figure 1 represents the trajectory through all ordered events for a simulated point pattern on $[0,1]^2$ ($n = 3$).

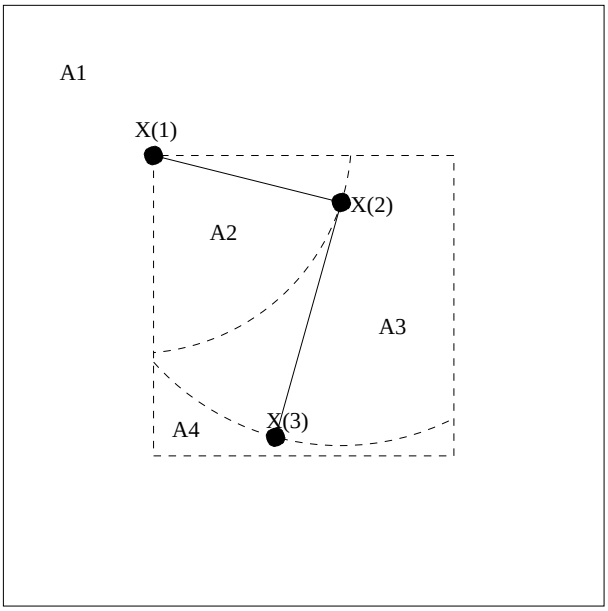

FIGURE 1. The data transformation

Denote A_1 the subdomain of A of points closest than D_1 from the boundary. Its area is the area that has to be explored before finding the first event, when starting from the boundary of A. Actually, under H_0, the distribution of $|A_1|$ is the same than the distribution of a unidimensional spacing generated by a uniform $[0,1]$ n-sample (Pyke, 1965), independently of the geometry of the domain A. Indeed, if we denote $S_{1,r} = \{x \in A : d(x, \partial A) \leq r\}$, we get under H_0

$$\begin{aligned} \forall r > 0, P(|A_1| \leq |S_{1,r}|) = 1 - P(|A_1| > |S_{1,r}|) &= 1 - P(\forall i = 1, \cdots, n,\ X_i \in A \setminus S_{1,r}) \\ &= 1 - |A \setminus S_{1,r}|^n = 1 - (1 - |S_{1,r}|)^n. \end{aligned}$$

Similarly, denote A_2 the subdomain of $A \setminus A_1$ of points whose distance to $X_{(1)}$ is less than D_2, and $S_{2,r} = \{x \in A \setminus A_1 : d(x, X_{(1)}) \leq r\}$. Under H_0,

$$\begin{aligned} \forall r > 0, \qquad & P(|A_2| \leq |S_{2,r}| \ \big|A_1) = 1 - P(|A_2| > |S_{2,r}| \ \big|A_1) \\ = \ & 1 - P(\forall i = 1, \cdots, n \ s.t \ X_i \neq X_{(1)},\ X_i \in A \setminus S_{2,r} \big| X_i \in A \setminus A_1) \\ = \ & 1 - \Big(\frac{|A \setminus S_{2,r}|}{|A \setminus A_1|}\Big)^{n-1} = 1 - \Big(\frac{1 - |S_{2,r}|}{1 - |A_1|}\Big)^{n-1}. \end{aligned}$$

This is the same distribution as the distribution of a uniform spacing conditionnally to another one, so A_2 has also the same distribution as a uniform

spacing. The same method can be used recursively for every $|A_i|$. Finally, if we denote $A_{n+1} = A \setminus (A_1 \cup \cdots \cup A_n)$, we can set that $\{|A_1|, \cdots, |A_{n+1}|\}$ has the same distribution as the spacings generated from a uniform $[0,1]$ n-sample. Let us denote $\{|A_1|, \cdots, |A_{n+1}|\}$ as the area spacings generated from $X_1, \cdots, X_n$.

As in the temporal setting, it is possible to transform these area spacings so as to get rid of the dependence. We introduce the modified area spacings

$$\begin{aligned} \forall i = 1, \cdots, n, \quad \tilde{A}_i &= F(|A_i| \,\big|\, A_1, \cdots, A_{i-1}) \\ &= 1 - \Big(\frac{1 - \sum_{j=1}^{i} |A_j|}{1 - \sum_{j=1}^{i-1} |A_j|}\Big)^{n-i+1}. \end{aligned}$$

4.3. The test statistic

As the area spacings are equivalent in the spatial setting to the spacings in the temporal setting, the statistics we introduce will be similar to the one introduced in the previous chapter. We denote $S_{j,k}^{(0)} = \sum_{i=j+1}^{k} A_i$, $S_{j,k}^{(1)} = \sum_{i=j+1}^{k} \tilde{A}_i$ and $S_{j,k}^{(2)} = \sum_{i=j+1}^{k} \tilde{A}_i^2$ and the concentration indicator for a zone containing the events $X_{(j)}, \cdots, X_{(k)}$ is

$$\Lambda_{j,k}^{(l)} = F_{j,k}^{(l)}(S_{j,k}^{(l)})$$

where $F_{j,k}^{(l)}(.)$ is defined as in the previous chapter for the different values of l.

Whatever the value of l, we compare all the concentration indicators $\Lambda_{j,k}^{(l)}$. The most significative cluster will be the one minimizing $\Lambda_{j,k}^{(l)}$ and we denote $\Lambda^{(l)} = \min_{1 \le j < k \le n} \Lambda_{j,k}^{(l)}$.

Thus one knows which are the events included in the most significative cluster but now we need to define the cluster shape. Two possibilities are given by Dematteï & *al.* (2005). The first one consists in identifying a cluster with the union of the Voronoi cells associated to its events. The second one consists in defining an influence zone for each event as the circle centered on it and whose area is the total observation area divided by the number of events : the cluster is then the union of the influence zones of its events.

4.4. The adaptation to the population inhomogeneity

As said in the introduction of this part devoted to cluster detection, it is often necessary to take into account the inhomogeneity of the observed population. For example, if one observes the cases of a certain disease in a certain country, he may adapt the procedure to a highest population in the largest cities.

This population inhomogeneity can be expressed through a function $g(t), t \in A$ giving the amount of population in the neighbourhood of t. Then the adaptation is done by replacing $|A_i|$ with

$$\frac{\int_{A_i} g(t)\nu(dt)}{\int_A g(t)\nu(dt)}.$$

Otherwise, it may be expressed through "controls", representing for example each location a sane people is observed, and denoted $\{c_1, \cdots, c_m\}$. Then the adaptation is done by replacing $|A_i|$ with

$$\frac{\sum_{j=1}^{m} \mathbb{1}(c_j \in A_i)}{m}.$$

4.5. Data analysis

The method to assess the significativity of the most likely cluster is the same as in the temporal setting, as manipulating area spacings is equivalent to manipulating unidimensional spacings. This means that the quantiles of $\Lambda^{(l)}, l = 1, 2, 3$, are the sames than in the temporal setting and can be obtained through simulations of point patterns on $[0, 1]$. Contrary to the quantiles of the spatial scan statistic, the quantiles of $\Lambda^{(l)}$ under H_0 do not depend on the observation domain A and do not need to be recalculated when this changes or when the possible clusters family changes.

We decide to apply our method with $l = 0, 1, 2$ to data sets simulated according to one-cluster alternatives.

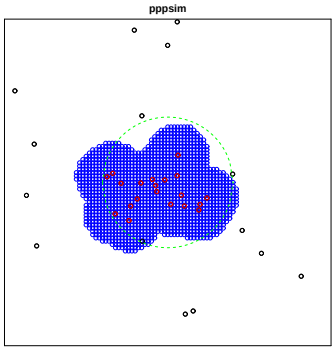

FIGURE 2. Result of the method

A flat cluster is obtained via the density function

$$f_1(x,y) = \begin{cases} \frac{r}{1+0.04\pi(r-1)} & \text{if } (x-0.5)^2+(y-0.5)^2 < 0.2^2, \\ \frac{1}{1+0.04\pi(r-1)} & \text{else if } (x,y) \in [0,1]^2, \\ 0 & \text{otherwise,} \end{cases}$$

where the parameter r is the ratio between the maximum density and the minimum density on the definition domain.

We simulate a 35-point pattern according to this density and Figure 2 shows the result of the method with $l = 0, r = 10$ and $n_0 = 2$. The green circle represents the "true cluster", the red points are the events contained in the most likely cluster, found to be significative, and the blue zone is the most likely cluster obtained when using "influence zones" as defined previously.

The correspondence index between the true cluster and its estimate, defined similarly as in the temporal setting, is almost 95% in this example.

Table 1 gives the results obtained when simulating 100 such data sets.

TABLE 1. Empirical power of tests for CSR against a flat cluster with $n_0 = 2$

r	Estimated power of the following :		
	$\Lambda^{(0)}$	$\Lambda^{(1)}$	$\Lambda^{(2)}$
10	**86.11 % (0.97)**	82.13 % (0.92)	77.72 % (0.92)

As in the temporal setting, it seems that the method with $l = 0$, relying on dependent area spacings, is the most efficient. This simulation study is to be continued, and it will be interesting to compare our method to existing ones described previously.

As said in the preceding chapter, it would be also worth adapting these methods to multiple clustering by adding a selection model criterion.

Finally we may think that the area spacings we introduced could also be efficient in a global test for H_0. For example, we may wonder whether a test based on the variance of these area spacings would be more powerful than existing tests, described in the introduction, against some alternatives.

PARTIE III

ESTIMATION NON-PARAMÉTRIQUE DE L'INTENSITÉ ET ADAPTATION AU CAS BRUITÉ

Cette partie reprend, dans un premier chapitre, une étude des techniques non-paramétriques en statistique spatiale, et notamment l'estimation d'intensité d'un processus ponctuel (Cucala & Thomas-Agnan, 2006b).Puis, dans le second chapitre, on se focalise sur l'adaptation d'un estimateur à noyau de l'intensité à des données bruitées (Cucala, 2006a).

CHAPITRE 5

MÉTHODES NON-PARAMÉTRIQUES POUR DONNÉES SPATIALES

5.1. Spécificité des données géoréférencées

Les données géoréférencées sont des données comportant une dimension spatiale, c'est à dire pour lesquelles une information géographique est attachée à chaque unité statistique. L'information géographique est en général la position de l'unité sur une carte ou dans un référentiel spatio-temporel. Ces données nécéssitent un traitement statistique adapté au risque d'une perte d'information, d'erreurs de spécifications, d'estimations non convergentes et non efficaces.

Un des outils de modélisation des données géoréférencées est le champ aléatoire. Lorsqu'une caractéristique $Z(s, \omega)$ d'une unité statistique est mesurée en un site s pour la réalisation ω, on notera Z_s la variable aléatoire associée, où l'indice s varie dans une partie $\mathcal{D}$ de $\mathbb{R}^d$. Comme pour les séries temporelles ($d = 1$), les hypothèses d'indépendance et d'égalité de répartition marginale entre les variables attachées à un ensemble de sites n'est pas satisfaite et le champ peut présenter à la fois

- une autocorrélation spatiale : les variables Z_s et Z_t étant d'autant plus corrélées que les sites s et t sont voisins,
- une hétérogénéité spatiale : la répartition marginale de Z_s varie avec s.

Mais à la différence des séries temporelles, les notions de passé et de futur n'ont pas leur pendant en spatial et il n'y a pas d'ordre naturel dans $\mathbb{R}^d$.

On distingue trois grands types de données géoréférencées :

- les données de type "géostatistique" : les variables Z_s sont observées en des points discrets et **déterministes** s_i de $\mathcal{D}$, un domaine de $\mathbb{R}^d$ contenant

un rectangle de volume strictement positif. Dans ce cas, Z_s a un sens pour s variant **continûment**.
- les données de type "économétrique" : la variable aléatoire Z_s est attachée à une zone, représentée par son centroïde s, pour une collection dénombrable de zones inclues dans $\mathcal{D} \in \mathbb{R}^d$. Dans ce cas, Z_s n'a de sens que sur une zone, comme par exemple le "revenu par foyer fiscal" qui n'a de sens qu'en moyenne sur une zone telle qu'une commune ou un canton. Les données latticielles observées sur un réseau régulier, indéxées par $(\mathbb{N}^*)^d$ peuvent entrer dans ce cadre.
- les données de type "processus ponctuel" ou "semis de points" : elles se distinguent des deux autres cas par le fait que la position s des observations de Z_s est à présent aléatoire. Si seule la position est observée, il s'agit d'un simple processus ponctuel. Si de plus, la variable Z_s (appelée marque) est observée en ces positions aléatoires, il s'agit d'un processus ponctuel marqué.

Nous allons décrirer dans ce chapitre quelques techniques non paramétriques dans le cadre des modèles pour données spatiales qui s'apparentent à celles étudiées par ailleurs dans ce manuel.

5.2. Approches non paramétriques en géostatistique

En géostatistique , on décompose généralement le champ aléatoire étudié $\{Z_s, s \in D \subset \mathbb{R}^d\}$ de la manière suivante

$$Z_s = \mu(s) + \delta_s \tag{1}$$

où $\mu(.) = \mathbb{E}Z_.$ est un champ déterministe appelé champ moyen ou tendance, et δ_s un champ aléatoire appelé champ erreur. On dispose d'un ensemble d'observations de Z, $\{(s_i, Z_{s_i}), i = 1, \cdots, n\}$ et les principaux problèmes à résoudre sont alors l'estimation du champ moyen $\mu(s)$ en tout point du domaine D, la détermination de la structure de covariance du champ erreur, la prédiction du champ en un point non observé. Z_{s_i} sera parfois noté plus simplement Z_i. La structure de covariance est définie par la fonction d'autocovariance

$$R(s,t) = \mathbb{C}ov(Z_s, Z_t).$$

On fait généralement une hypothèse de stationnarité (au sens fort : invariance de la loi du vecteur $Z_{s_1}, \ldots, Z_{s_k}$ par translation, ou à l'ordre deux : invariance

par translation des moments d'ordre un et deux) qui implique le fait que R est une fonction de $s - t$

$$R(s, s+h) = Cov(Z_s, Z_{s+h}) = \rho(h). \tag{2}$$

Lorsque cette hypothèse n'est pas plausible, on élargit à une hypothèse de stationnarité des incréments ou stationnarité intrinsèque à l'ordre un. On n'exige pas dans ce cas l'existence d'un moment d'ordre un pour le champ lui-même mais seulement pour les accroissements du champ et l'on demande que

$$\mathbb{E}(Z_{s+h} - Z_s) = 0, \tag{3}$$

$$\mathbb{V}ar(Z_{s+h} - Z_s) = 2\gamma(h) = \mathbb{E}(Z_{s+h} - Z_s)^2. \tag{4}$$

La fonction γ s'appelle alors le semi-variogramme et 2γ le variogramme. Cette notion peut se généraliser au cas d'incréments d'ordre supérieur.

5.2.1. Estimation de la tendance. — Prenons le modèle (1) et considérons le problème de l'estimation de la tendance μ pour lequel ont été élaborées de nombreuses techniques nonparamétriques. Ce modèle n'est autre qu'un modèle de régression à effets fixes et l'on va naturellement retrouver les estimateurs classiques vus en régression non-paramétrique.

Le plus simple, basé sur l'hypothèse de stationnarité sous laquelle μ est constante, consiste à estimer μ en tout point par la moyenne (ou la médiane, plus robuste aux valeurs aberrantes) des observations. On peut aussi considérer la localisation géographique comme une variable explicative de la variable aléatoire Z et effectuer une régression non-paramétrique en utilisant des estimateurs de type Nadaraya-Watson

$$\hat{\mu}(s_0) = \frac{\sum_{i=1}^{n} K(d_{0,i}) Z_{s_i}}{\sum_{i=1}^{n} K(d_{0,i})}$$

où $d_{i,j} = ||s_i - s_j||$ et $K(.)$ est un noyau généralement décroissant pour permettre de donner plus de poids aux observations les plus proches. Il est également possible de faire varier les poids selon s_0 et de les adapter à la densité d'observations au voisinage (Cleveland & Devlin, 1988). Enfin, on peut également estimer $\mu(.)$ à l'aide de splines de régression (Wahba, 1990).

Tournons nous à présent vers des estimateurs plus spécifiques à la dimension deux. Les méthodes de polissage sont initialement élaborées pour des données latticielles mais s'étendent à tout type de données en identifiant chaque point

d'observation au nœud le plus proche d'un quadrillage judicieusement choisi. Elles s'appuient sur une décomposition additive du champ moyen en un terme global et des termes dépendant chacun d'une coordonnée. En 2D, on obtient pour $s = (x, y)' : \mu(s) = a + c(x) + r(y)$, et, pour tout point s_{ij} du quadrillage, $\mu(s_{ij}) = a + c_i + r_j$.

L'un des algorithmes les plus populaires est le polissage médian, introduit par Tukey (1977) et qui a pour avantage sur le polissage moyen de générer des résidus non biaisés. L'idée principale est de retirer la médiane de chaque colonne de données i (i.e. données réparties sur une même colonne), puis de chaque ligne de données j, et d'itérer jusqu'à convergence. Les effets médians a, c_i et r_j sont alors estimés par les sommes des médianes retirées à chaque itération. Le champ moyen estimé $\hat{\mu}(.)$ est obtenu sur le quadrillage, et peut être reconstruit en tout point par interpolation planaire.

Une autre idée consiste à utiliser la triangulation de Delaunay (celle maximisant l'angle minimum de tous les triangles) des points d'observation $\{s_1, \cdots, s_n\}$. La valeur $\hat{\mu}(s_0)$ est obtenue par interpolation planaire basée sur les sommets du triangle de Delaunay auquel appartient s_0. Une autre méthode nommée "natural neighbor interpolation" (Sibson, 1981) s'appuie elle sur la tessellation de Voronoi qui n'est autre que le dual de la triangulation de Delaunay.

5.2.2. Prédiction non paramétrique d'un champ aléatoire. — La méthode de prédiction la plus connue en géostatistique s'appelle le krigeage. A partir des réalisations de Z_s en n sites $s_i, i = 1, \cdots, n$, on souhaite prédire Z_s en un site s où il n'y a pas eu de mesure en évaluant également l'erreur de prédiction. On peut aussi vouloir prédire plus généralement une variable du type $\int_A w(s) Z_s ds$: par exemple, dans le cas de données hydrologiques, la variable cible peut être la quantité totale de polluant dans une partie A du territoire. Le krigeage suppose que le champ est intrinsèquement stationnaire à un certain ordre.

En supposant la structure de covariance connue, le krigeage consiste à rechercher le "meilleur prédicteur linéaire sans biais" (nous utiliserons le sigle anglosaxon BLUP). On prédit Z_s par une combinaison linéaire Z^* des $Z_i = Z_{s_i}$ en lui imposant les contraintes suivantes :

- être sans biais : $\mathbb{E}(Z^*) = \mathbb{E}(Z_s)$,
- avoir une erreur quadratique de prédiction minimum parmi les prédicteurs Z^* linéaires sans biais :

$$\min \mathbb{E}(Z^* - Z_s)^2.$$

Avec divers degrés de généralité, plusieurs auteurs ont montré que le prédicteur de Y_t ainsi obtenu, considéré comme une fonction de la variable t est une spline de lissage. Citons en particulier les résultats de Kimeldorf and Wahba (1970a et b), Duchon (1976), Matheron (1981), Thomas-Agnan (1991), Kent & Mardia (1994).

Biau & Cadre (2004) s'intéressent à la prédiction d'un champ aléatoire Z indexé par $(\mathbb{N}^\star)^d$, espace latticiel de dimension d, par des méthodes à noyau. Pour $i \in (\mathbb{N}^\star)^d$, notons S_i la localisation géographique du nœud d'indice i, Z_i la valeur de Z au nœud i et $\tilde{Z}_i$ le vecteur aléatoire regroupant les valeurs de Z dans un voisinage fixé ν_i de i. Notons également $\mathcal{I}$ l'ensemble des sites sur lesquels on souhaite connaître Z et $\mathcal{S} \subset \mathcal{I}$ l'ensemble des sites d'observation de Z.

Biau & Cadre (2004) définissent la prédiction $\hat{Z}_j$ de Z au point $j \in \mathcal{I} - \mathcal{S}$ comme l'espérance conditionnelle $\mathbb{E}(Z_j \mid \tilde{Z}_j)$ de Z_i sachant $\tilde{Z}_i$:

$$\hat{Z}_j = \frac{\sum_{i \in \mathcal{S}, \nu_i \subset \mathcal{S}} Z_i K\left(\frac{\tilde{Z}_j - \tilde{Z}_i}{h}\right)}{\sum_{i \in \mathcal{S}, \nu_i \subset \mathcal{S}} K\left(\frac{\tilde{Z}_j - \tilde{Z}_i}{h}\right)}.$$

On peut remarquer que la somme ne s'effectue ici que sur les nœuds auxquels à la fois la variable à régresser Z et le régresseur $\tilde{Z}$ sont connus. Leur démarche s'appuie sur l'intuition que Z_i est plus fortement lié aux valeurs au voisinage $\tilde{Z}_i$ qu'à la seule localisation géographique S_i. Les auteurs parviennent à démontrer la consistance uniforme et la normalité asymptotique de leur estimateur sous certaines conditions de mélange du champ $(Z_i, \tilde{Z}_i)_{i \in (\mathbb{N}^\star)^d}$, obtenant ainsi des résultats similaires à ceux valables dans le cadre de séries temporelles (Bosq, 1998).

5.2.3. Estimation non paramétrique de covariances et variogrammes de champs aléatoires. — Dans le modèle (1), considérons pour le cas d'un champ stationnaire Z_s à valeurs réelles, de moyenne $\mu(s) = \mu$, le problème de l'estimation de la structure de covariance ρ définie par (2) à partir d'une ou plusieurs réalisations du champ en un nombre fini de points (design) irrégulièrement espacés $s_1, \cdots, s_n$. La difficulté de ce problème tient au fait qu'une telle fonction de covariance est par nature semi définie positive, c'est à dire que, quel que soit l'entier n, quels que soient les n réels $a_1, \ldots, a_n$ et l'ensemble de n

positions $s_1, \ldots, s_n$ de D, elle doit satisfaire

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \rho(s_i - s_j) \geq 0. \tag{5}$$

Notons que la covariance empirique définie par

$$\hat{\rho}(h) = \frac{1}{\mid N(h) \mid} \sum_{(s_i, s_i+h)\in N(h)} (Z_{s_i} - \bar{Z})^2,$$

où

$$N(h) = \{(s_i, s_j) : s_i - s_j = h\},$$

et $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_{s_i}$ ne satisfait pas cette condition. Rappelons que d'après le théorème de Bochner (Rudin, 1990), une fonction continue de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$ satisfait la condition 5 si et seulement si sa transformée de Fourier est une mesure positive bornée. Hall & *al.* (1994) pour un processus indexé par le temps et Hall & Patil (1994) pour un champ indexé par $\mathbb{R}^d$ avec $d > 1$ proposent un estimateur à noyau inspiré de la formule de Nadaraya-Watson. Pour imposer la condition de semi définie positivité à cet estimateur, Hall & *al.* (1994) utilisent la condition nécéssaire et suffisante de Bochner. Pour un noyau de densité K et une fenêtre h, ils définissent un estimateur préliminaire par

$$\hat{\rho}_H(s) = [\sum_i \sum_j \hat{Z}_{ij} K\{(s - s_{ij})/h\}][\sum_i \sum_j K\{(s - s_{ij})/h\}]^{-1}, \tag{6}$$

où $\hat{Z}_{ij} = \{Z_{t_i} - \overline{Z}\}\{Z_{t_j} - \overline{Z}\}$, $s_{ij} = s_i - s_j$ et $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_{s_i}$. Cet estimateur est ensuite tronqué au point T_1 et relié linéairement à 0 au point T_2 pour donner $\hat{\rho_1}$. Pour contraindre la transformée de Fourier $\mathcal{F}(\hat{\rho_1})$ à être positive, celle-ci est ensuite tronquée à zéro dès la première fréquence où elle devient négative. Pour $\hat{\theta} = \inf\{\theta \geq 0 : \mathcal{F}(\hat{\rho_1})(\theta) \leq 0\}$, l'estimateur final est alors

$$\tilde{\rho}_H(t) = (2\pi)^{-1} \int_{-\hat{\theta}}^{\hat{\theta}} \mathcal{F}(\hat{\rho_1})(\theta) \cos(\theta t) d\theta. \tag{7}$$

Sous certaines conditions de régularité, Hall & *al.* (1994) montrent que cet estimateur est convergent en moyenne quadratique et en norme sup. Cette approche nécessite le calibrage de trois paramètres h, T_1, T_2 et se revèle difficile à implémenter en pratique. Notons que Masry (1983) propose un autre type d'estimateur à noyau dans le cas d'un "design" aléatoire.

Dans le cas d'un champ indexé par $\mathbb{R}$, Elogne & *al.* (2006) introduisent un estimateur de ρ basé sur une interpolation du processus. Le fait d'interpoler le processus leur permet de se ramener à un processus observé continûment

et d'utiliser alors l'estimateur de Parzen (1961). L'interpolation des processus aléatoires est étudiée et exploitée par ailleurs par exemple dans Weba (1992) et Seleznjev (2000) pour des interpolations splines et Klamer & Masry (1982) pour des polynômes d'interpolation de Lagrange. Elogne (2006) étend cette définition au cas spatial. Les méthodes d'interpolation envisagées sont d'une part l'interpolation linéaire par morceaux et l'interpolation par spline cubique naturelle dans le cas où s varie dans $\mathbb{R}$, l'interpolation linéaire sur une triangulation de Delaunay de l'enveloppe convexe Ω dans le cas où s varie dans $\mathbb{R}^2$.

Plus précisément, notons $\tilde{Z}$ le champ interpolé. Dans le cas d'un champ indexé par $\mathbb{R}$, on peut définir d'abord l'estimateur suivant de μ

$$\tilde{\mu}_n = \frac{1}{s_n - s_1} \int_{s_1}^{s_n} \tilde{Z}_t dt, \tag{8}$$

et par suite l'estimateur suivant de ρ

$$\begin{aligned} \tilde{\rho}_n(t) &= \frac{1}{s_n - s_1} \int_{-\infty}^{\infty} (\tilde{X}_s - \tilde{\mu}_n)(\tilde{X}_{s+|t|} - \tilde{\mu_n}) ds & (9) \\ &= \frac{1}{s_n - s_1} \int_{s_1}^{s_n - |t|} (\tilde{X}_s - \tilde{\mu_n})(\tilde{X}_{s+|t|} - \tilde{\mu_n}) ds. & (10) \end{aligned}$$

Dans le cas d'un champ indexé par $\mathbb{R}^2$, si l'on note Ω_n l'enveloppe convexe des lieux $s_1, \cdots, s_n$, et $\left|\Omega_n\right|$ l'aire de Ω_n, l'estimateur de μ est donné par

$$\tilde{\mu}_n = \left|\Omega_n\right|^{-1} \int_{\Omega_n} \tilde{Z}^*(s) ds,$$

et celui de ρ par

$$\tilde{\rho}_n(s) = \left|\Omega_n\right|^{-1} \int_{\Omega_n \cap (\Omega_n - s)} \big(\tilde{Z}(u) - \tilde{\mu}_n\big)\big(\tilde{Z}(u+s) - \tilde{\mu}_n\big) du, \tag{11}$$

où $\Omega_n - s$ represente le translaté de Ω_n par $-s$, $\Omega_n \cap (\Omega_n - s)$ l'ensemble des points tels que u et $u + s$ appartiennent à Ω_n.

Il est alors facile de démontrer que l'estimateur $\tilde{\rho}(s)$ est une fonction semi définie positive de s. Elogne & *al.* (2006) montrent la convergence en moyenne

quadratique ponctuelle de ρ sous des hypothèses de régularité sur ρ et sur le champ Z.

Dans le cas d'un champ non stationnaire, on fait en général l'hypothèse plus large de stationnarité intrinsèque et on modélise plutôt le variogramme (voir (3) que la covariance. De même que la covariance doit satisfaire la condition de semi définie positivité, le variogramme doit être conditionnellement semi défini négatif c'est à dire que quel que soit l'entier n, quels que soient les n réels $a_1, \ldots, a_n$ satisfaisant $\sum_{i=1}^{n} a_i = 0$ et l'ensemble de n positions $s_1, \ldots, s_n$ de D, il doit satisfaire

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \rho(s_i - s_j) \leq 0. \tag{12}$$

Le variogramme empirique défini par

$$\hat{\gamma}(h) = \frac{1}{\mid N(h) \mid} \sum_{(s_i, s_i+h) \in N(h)} (Z_{s_i+h} - Z_{s_i})^2,$$

où $N(h)$ est défini comme précédemment, ne satisfait pas cette condition. Dans le cas isotrope, c'est à dire lorsque $\gamma(h)$ ne dépend que de $\parallel h \parallel$, Shapiro & Botha (1991) proposent de l'ajuster par moindres carrés pondérés à une version discrétisée de la représentation spectrale obtenue par Yaglom (1957) à partir du théorème de Bochner. En approximant la fonction de répartition spectrale par une fonction en escalier ayant un nombre fini de sauts $y_1, \cdots, y_m$ aux noeuds $z_1, \cdots, z_n,$, l'approximation du variogramme s'écrit

$$\bar{\gamma}(r) = c_0 - \sum_{j=1}^{m} \cos(r z_j) y_j,$$

et l'on ajuste les coefficients $(y_1, \cdots, y_m, c_0)$ en minimisant, pour des poids w_i

$$\sum_{i=1}^{N} w_i (\hat{\gamma}(h_i) - \bar{\gamma}(h_i))^2,$$

où les h_i sont les distances pour lesquelles $N(h)$ est non vide. L'optimisation se fait par programmation quadratique avec contraintes linéaires. On obtient simultanément une estimation de la variance c_0, ce qui permet d'en déduire une estimation de ρ dans le cas stationnaire. Cette approche nécessite cependant un design régulier et les vitesses de convergence ne sont pas explicitées. Par ailleurs, si l'on ne rajoute pas de contraintes sur les dérivées pour régulariser la solution, celle-ci tend à présenter des oscillations non souhaitables. Genton

& Gorsich (2002) proposent un choix des noeuds permettant d'obtenir une solution plus régulière sans ces contraintes.

Dans un cadre spatio-temporel avec p répétitions temporelles et non stationarité spatiale, Guillot & *al.* (2000) et Bel (2002) proposent deux estimateurs obtenus par régularisation par noyau de la covariance empirique ou du variogramme empirique. Si $Z_{s_i}^k$ désigne la répétition k ($k = 1, \cdots, p$) de la variable Z_{s_i}, l'estimateur de Guillot & *al.* réalise un lissage par noyau à la Gasser & Müller (1987) de la covariance empirique

$$\hat{\rho}(s_i, s_j) = \frac{1}{p} \sum_{k=1}^{p} (Z_{s_i}^k - \bar{Z}_{s_i})(Z_{s_j}^k - \bar{Z}_{s_j})$$

donné par

$$\tilde{\rho}(s,t) = \sum_{i,j} \hat{\rho}(s_i, s_j) \int_{D_i \times D_j} \frac{1}{\epsilon^4} K(\frac{x-u}{\epsilon}, \frac{y-v}{\epsilon}) du dv,$$

où D_i est la partition de Voronoï induite par les sites s_i. La semi définie positivité de $\tilde{\rho}$ est alors une conséquence de la semi définie positivité du noyau K lui-même. De façon similaire, l'estimateur de Bel réalise un lissage par noyau à la Nadaraya-Watson de la forme

$$\tilde{\rho}(s,t) = \sum_{i,j} \hat{\rho}(s_i, s_j) \frac{K(\frac{x-s_i}{\epsilon}, \frac{y-s_j}{\epsilon})}{\sum_{k,l} K(\frac{x-s_k}{\epsilon}, \frac{y-s_l}{\epsilon})},$$

où K est un noyau à valeurs positives et séparable, i.e. $K(u,v) = H(u)H(v)$, ce qui suffit à assurer la semi définie positivité de $\tilde{\rho}$. Ces deux approches ne comportent pas de résultats asymptotiques.

Citons aussi le travail de Sampson & Guttorp (1992) qui dans le cadre d'un modèle spatio-temporel avec stationarité temporelle, utilisent une déformation de l'espace des positions pour se ramener à un problème stationnaire dans l'espace et isotrope.

5.3. Approches non paramétriques pour processus ponctuels

Plaçons nous à présent dans le cadre d'un jeu de données de type semis de points (non marqué) $\{s_i, i = 1, \cdots, n\}$, où les variables de localisation $S_i \in X \subseteq \mathbb{R}^d$ sont elles-mêmes aléatoires, ainsi que leur nombre N. Pour un tel modèle de processus ponctuel spatial, notons $\mathcal{X}$ la tribu des boréliens bornés

de X et M la mesure aléatoire de comptage qui, à tout élément B de $\mathcal{X}$, associe $M(B)$ le nombre d'événements contenus dans B. Un processus ponctuel spatial est caractérisé de manière unique par les probabilités d'évitement définies par $\mathbb{P}[M(B) = 0], \forall B \in \mathcal{X}$.

Afin de caractériser les différents modèles de processus ponctuels, il est nécessaire de définir quelques notions.

Introduisons d'abord la mesure du moment d'ordre 1 de M :

$$\forall B \in \mathcal{X},\ \mu_M(B) = \mathbb{E}[M(B)].$$

De même, on peut définir la mesure du moment d'ordre k de M :

$$\forall (B_1, \cdots, B_k) \in \mathcal{X}^k,\ \mu_M^{(k)}(B_1 \times \cdots \times B_k) = \mathbb{E}[M(B_1) \cdots M(B_k)].$$

Considérons maintenant ces mesures de moment lorsque l'on prend pour B un singleton $\{s\}$, $s \in X$. Notons ds et du des volumes infinitésimaux centrés respectivement en s et u, et $\nu(.)$ la mesure de Lebesgue.

L'intensité (de premier ordre) de M est définie par :

$$\forall s \in X,\ \lambda(s) = \lim_{\nu(ds)\to 0} \mu_M(ds)/\nu(ds).$$

L'intensité de second ordre de M est définie par :

$$\forall (s,u) \in X^2,\ \lambda_2(s,u) = \lim_{\nu(ds)\to 0, \nu(du)\to 0} \frac{\mu_M^{(2)}(ds \times du)}{\nu(ds)\nu(du)}.$$

Un processus ponctuel sera dit stationnaire si sa mesure de comptage M est invariante par translation. On a alors $\lambda(s) = \lambda$, $\forall s \in X$ et il existe une fonction $\lambda_2^\star$ telle que $\lambda_2(s,u) = \lambda_2^\star(s-u)$, $\forall (s,u) \in X^2$.

5.3.1. Caractéristiques du second ordre d'un processus ponctuel. — De très nombreux modèles de processus ponctuels ont été introduits. Le modèle de base est le processus de Poisson homogène, appelé ainsi car le nombre d'événements total N est une variable aléatoire de Poisson de paramètre λ, tandis que, sachant N, les localisations S_i sont indépendantes et suivent une loi uniforme sur X. Ce modèle sert de standard pour ce qu'on entend intuitivement comme une répartition aléatoire de points, c'est pourquoi le problème de tester

l'adéquation d'un jeu de données avec ce modèle, dite hypothèse d'homogénéité spatiale (“ Complete Spatial Randomness “ en anglais), est la première étape importante de l'analyse d'un semis de points.

Pour cela, il existe quelques techniques nonparamétriques permettant notamment d'identifier des alternatives d'agrégation (lorsque les événements ont tendance à se concentrer) ou de régularité (lorsque les événements ont tendance à se repousser).

On définit la distribution au plus proche voisin $G(r)$ comme la probabilité que la distance entre un événement pris au hasard et l'événement le plus proche soit inférieure à r. Sous l'hypothèse d'homogénéité spatiale, correspondant au processus de Poisson homogène, on a $G(r) = 1 - \exp(-\lambda\pi r^2)$. Notons $R_i = \min_{j\neq i} ||S_i - S_j||$. On peut alors comparer cette distribution théorique à la fonction de répartition empirique des distances au plus proche voisin $\hat{G}(r) = \frac{1}{N}\sum_{i=1}^{N} I(R_i < r)$ (Hanisch, 1984). Il est également possible d'effectuer un test de type Monte-Carlo en vérifiant si la fonction $\hat{G}$ est bien à l'intérieur d'une enveloppe de simulation obtenue en générant des processus de Poisson homogènes (Besag & Diggle, 1977). Si $\hat{G}$ a tendance à être au-dessus de l'enveloppe, on conclura à une alternative d'agrégation, si elle a tendance à être en-dessous à une alternative de régularité. Les mêmes techniques peuvent être utilisées en considérant la distribution au plus proche événement $F(r)$ qui est la probabilité que la distance entre un point choisi au hasard et l'événement le plus proche soit inférieure à r.

Pour tout processus ponctuel stationnaire, définissons la fonction K, dite de Ripley, par la fonction qui, à un réel positif t, associe l'espérance du nombre d'événements à une distance inférieure à h d'un événement arbitraire :

$$K(t) = \frac{1}{|X|}\mathbb{E}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{I(||S_i - S_j|| < t)}{\lambda^2},$$

$|X|$ représentant le volume du domaine X. Cette fonction K est indépendante du domaine X. Sous l'hypothèse d'homogénéité spatiale, on a $K(t) = \pi t^2$. Comme précédemment, cette fonction peut être approchée par son estimateur empirique :

$$\hat{K}(t) = \frac{1}{\hat{\lambda}N}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} I(||S_i - S_j|| < t)$$

où $\hat{\lambda}$ est généralement l'estimateur issu de la méthode des moments $\hat{\lambda} = N/|X|$. La même démarche qu'avec les distributions G et F peut alors être utilisée (Ripley, 1976).

Une large famille de processus ponctuels stationnaires est celle des processus ponctuels de Markov, dont la densité f est définie par l'intermédiaire d'une fonction de potentiel Ψ. On a :

$$f\big((s_1, \cdots, s_n), n\big) = \frac{C}{n!} \exp\big[- \sum_{1 \leq i < j \leq n} \sum \Psi(||s_i - s_j||)\big]$$

où C est une constante de normalisation.

Des valeurs négatives de la fonction Ψ se traduisent par une tendance à l'agrégation entre les événements, des valeurs positives par une tendance à la régularité. Cette fonction de potentiel peut être notamment estimée par une technique nonparamétrique introduite par Diggle & *al.* (1987).

Notons $g(t) = \frac{\lambda_2^\star(t)}{\lambda^2} - 1$. La fonction de corrélation directe est la fonction $c(.)$ solution de l'équation d'Ornstein-Zernike :

$$c(t) = g(t) - \lambda(c * g)(t).$$

Sa résolution nécessite d'abord l'estimation de $g(.)$, donc de λ et $\lambda_2^\star(.)$. Un estimateur nonparamétrique de $\lambda_2^\star(.)$ s'obtient en utilisant la relation valable pour tout processus stationnaire :

$$\lambda_2^\star(t) = \frac{\lambda^2 \Gamma(1 + d/2)}{d\pi^{d/2} t^{d-1}} K'(t), \; \forall t > 0$$

et les estimateurs de λ et K évoqués précédemment (Fiksel, 1988). Un estimateur nonparamétrique de la fonction Ψ est alors obtenu par le biais de l'approximation de Percus-Yevick :

$$\Psi(t) \simeq \frac{g(t) + 1}{g(t) + 1 - c(t)}.$$

5.3.2. Intensité d'un processus ponctuel. — Parmi les modèles non stationnaires, le plus simple est certainement le processus de Poisson inhomogène, qui génère des événements de manière indépendante, mais selon une fonction d'intensité non constante sur X. A partir de maintenant, plaçons-nous dans ce cadre. Il existe une relation directe entre l'intensité du processus ponctuel, $\lambda(.)$,

et la densité d-dimensionnelle $f(.)$ de toute localisation S_i conditionnellement à N :

$$\forall s \in X,\ f(s) = \frac{\lambda(s)}{\int_X \lambda(s)\nu(ds)}.$$

Ainsi, on peut obtenir des estimateurs de la fonction d'intensité $\lambda(.)$ en utilisant les techniques nonparamétriques d'estimation de densité multivariée. Diggle (1985) introduit un estimateur à noyau de $\lambda(s)$ donné par :

$$\hat{\lambda}_{EC,h}(s) = \frac{\sum_{i=1}^{N} h^{-d} K\big(\frac{s-S_i}{h}\big)}{\int_X h^{-d} K\big(\frac{s-u}{h}\big)\nu(du)}$$

où le dénominateur est un terme de correction au bord nécessaire lorsque le domaine d'observation est limité. Le choix de la largeur de bande permettant de minimiser l'erreur quadratique moyenne intégrée :

$$EQMI(h) = \mathbb{E}\Big\{ \int_X \big(\hat{\lambda}_{EC,h}(s) - \lambda(s)\big)^2 \nu(ds)\Big\}$$

se fait selon les mêmes méthodes que dans le cadre de l'estimation de densité (Scott, 1992).

Notons $r_k(s)$ la distance de s au $k^{\text{ième}}$ plus proche événement, et $V_k(s)$ le volume de la boule de rayon $r_k(s)$. Un estimateur aux plus proches voisins de $\lambda_k(s)$ (Silverman, 1986) est alors :

$$\hat{\lambda}_k(s) = \frac{k}{V_k(s)}.$$

On peut également utiliser l'estimateur hybride :

$$\hat{\lambda}_k(s) = \frac{\sum_{i=1}^{n} r_k(s)^{-d} K\big(\frac{s-S_i}{r_k(s)}\big)}{\int_X r_k(s)^{-d} K\big(\frac{s-u}{r_k(s)}\big)\nu(du)}$$

qui nécessite comme le précédent de choisir l'entier k optimal.

Néanmoins, contrairement au cadre de l'estimation de densité, le nombre d'événements N est ici aléatoire, suivant une loi de Poisson de paramètre $m = \int_X \lambda(s)\nu(ds)$, et non un paramètre que l'on peut contrôler, ce qui entraîne quelques différences. Pour analyser cela, plaçons-nous dans le cas où $X = \mathbb{R}$.

Pour estimer $\lambda(s)$, on utilise l'estimateur à noyau :

$$\hat{\lambda}_h(s) = \sum_{i=1}^{N} \frac{1}{h} K\Big(\frac{s - S_i}{h}\Big)$$

car la correction de bord n'a plus lieu d'être dans ce cas.

Si on note $Y = \sum_{i=1}^{N} g(S_i)$ où $g()$ est une fonction mesurable, on a (Kutoyants, 1998) :

$$\mathbb{E}Y = \int_{\mathbb{R}} g(s)\lambda(s)\nu(ds)$$

et

$$\mathrm{Var}Y = \int_{\mathbb{R}} g^2(s)\lambda(s)\nu(ds).$$

On obtient alors :

$$\mathbb{E}\hat{\lambda}_h(s) = \lambda(s) + \frac{h^2}{2}\lambda''(s)\int_{\mathbb{R}} u^2 K(u)du + O(h^4)$$

et

$$\mathrm{Var}\hat{\lambda}_h(s) = \frac{1}{h}\lambda(s)\int_{\mathbb{R}} K^2(u)du + O(1).$$

L'erreur quadratique moyenne intégrée est de la forme :

$$EQMI(h) = \frac{h^4}{4}\Big(\int_{\mathbb{R}} \lambda''(s)ds\Big)^2\Big(\int_{\mathbb{R}} u^2K(u)du\Big)^2 + \frac{1}{h}\int_{\mathbb{R}}\lambda(s)ds\int_{\mathbb{R}} K^2(u)du + O(h^8) + O(1)$$

et aucun choix de h ne permet de la faire tendre vers 0. Autrement dit, comme le soulignent Diggle & Marron (1988), il apparaît impossible d'obtenir un estimateur de $\lambda(.)$ qui soit consistant.

Diggle & Marron (1988) parviennent à exprimer $EQMI(h)$ lorsque l'on utilise un noyau uniforme et que l'on considère un processus de Cox (Cressie, 1993, p.657) : ils constatent alors la convergence de $EQMI(h)/m^2$.

Voulant nous affranchir de cette contrainte sur le noyau et prenant en compte le caractère inconnu de m (néanmoins estimable par n, puisque $\mathbb{E}(N) = m$) , nous préférons nous intéresser à l'estimation de $\lambda_0(s) = \frac{\lambda(s)}{m}$, qui s'apparente

à la notion de densité. On utilisera alors l'estimateur :

$$\hat{\lambda}_{0,h}(s) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h} K\Big(\frac{s - S_i}{h}\Big) \mathbb{1}[N \neq 0]$$

et on se ramènera au problème initial par la relation $\hat{\lambda}_h(s) = N\hat{\lambda}_{0,h}(s)$.

Lemme 1. — *Si on note* $Z = \frac{1}{N}\sum_{i=1}^{N} g(S_i)\mathbb{1}[N \neq 0]$ *et* $A(m) = \mathbb{E}\big[\frac{1}{N}\mathbb{1}[N \neq 0]\big] = \sum_{k=1}^{\infty} \frac{e^{-m}m^k}{kk!}$, *on a :*

$$\mathbb{E}Z = \int_{\mathbb{R}} g(s)\lambda_0(s)\nu(ds)(1 - e^{-m})$$

et

$$VarZ = \int_{\mathbb{R}} g^2(s)\lambda_0(s)\nu(ds)A(m) - \Big(\int_{\mathbb{R}} g(s)\lambda_0(s)\nu(ds)\Big)^2 \big(A(m) - e^{-m} + e^{-2m}\big).$$

Preuve :

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}[\mathbb{E}g(S_i)|N = k > 0] \\
&= \sum_{k=1}^{\infty} \mathbb{E}g(S_i) \times \mathbb{P}(N = k) \\
&= \sum_{k=1}^{\infty} \int_{\mathbb{R}} g(s)\lambda_0(s)\nu(ds)\frac{e^{-m}m^k}{k!} \\
&= \int_{\mathbb{R}} g(s)\lambda_0(s)\nu(ds) \times \sum_{k=1}^{\infty} \frac{e^{-m}m^k}{k!} \\
&= \int_{\mathbb{R}} g(s)\lambda_0(s)\nu(ds)(1 - e^{-m})
\end{aligned}$$

et

$$\begin{aligned}\mathrm{Var}Z &= \mathbb{E}\Big[\mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{g(S_i)g(S_j)}{n^2}|N=k>0\Big]-(\mathbb{E}Z)^2\\ &= \sum_{k=1}^{\infty}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\mathbb{E}[g(S_i)g(S_j)]}{k^2}\frac{e^{-m}m^k}{k!}-(\mathbb{E}Z)^2.\end{aligned}$$

$$\text{Or } \mathbb{E}[g(S_i)g(S_j)]=\left\{\begin{array}{cc}\int_{\mathbb{R}}\int_{\mathbb{R}}g(s)g(t)\lambda_0(s)\lambda_0(t)\nu(ds)\nu(dt) & \text{if } i\neq j\\ \int_{\mathbb{R}}g^2(s)\lambda_0(s)\nu(ds) & \text{if } i=j\end{array}\right.$$

donc

$$\begin{aligned}\mathrm{Var}Z &= \sum_{k=1}^{\infty}\Big[\Big(\frac{k-1}{k}\int_{\mathbb{R}}\int_{\mathbb{R}}g(s)g(t)\lambda_0(s)\lambda_0(t)\nu(ds)\nu(dt)\\ &+ \frac{1}{k}\int_{\mathbb{R}}g^2(s)\lambda_0(s)\nu(ds)\Big)\frac{e^{-m}m^k}{k!}\Big]-(\mathbb{E}Z)^2\\ &= \Big(\int_{\mathbb{R}}g(s)\lambda_0(s)\nu(ds)\Big)^2\times\sum_{k=1}^{\infty}\frac{(k-1)e^{-m}m^k}{kk!}\\ &+ \int_{\mathbb{R}}g^2(s)\lambda_0(s)\nu(ds)\times\sum_{k=1}^{\infty}\frac{e^{-m}m^k}{kk!}-(\mathbb{E}Z)^2\\ &= \Big(\int_{\mathbb{R}}g(s)\lambda_0(s)\nu(ds)\Big)^2\times\Big(\mathbb{E}\big[\frac{N-1}{N}\mathbb{1}[N\neq 0]\big]-\mathbb{E}\big[\mathbb{1}[N\neq 0]\big]^2\Big)\\ &+ \int_{\mathbb{R}}g^2(s)\lambda_0(s)\nu(ds)\times\mathbb{E}\big[\frac{1}{N}\mathbb{1}[N\neq 0]\big]\\ &= \int_{\mathbb{R}}g^2(s)\lambda_0(s)\nu(ds)A(m)-\Big(\int_{\mathbb{R}}g(s)\lambda_0(s)\nu(ds)\Big)^2\big(A(m)-e^{-m}+e^{-2m}\big).\end{aligned}$$

■

On obtient alors :

$$\begin{aligned}\mathbb{E}\hat{\lambda}_{h,0}(s) &= (1-e^{-m})\int_{\mathbb{R}}\frac{1}{h}K\big(\frac{x-s}{h}\big)\lambda_0(s)\nu(ds)\\ &= \lambda_0(s)(1-e^{-m})+\frac{h^2}{2}(1-e^{-m})\lambda_0''(s)\int_{\mathbb{R}}u^2K(u)du+(1-e^{-m})O(h^4)\end{aligned}$$

et

$$\begin{aligned}\mathrm{Var}\hat{\lambda}_h(s) &= \int_{\mathbb{R}} \frac{1}{h^2} K^2\big(\frac{x-s}{h}\big)\lambda_0(s)\nu(ds)A(m) \\ &- \Big(\int_{\mathbb{R}} \frac{1}{h} K\big(\frac{x-s}{h}\big)\lambda_0(s)\nu(ds)\Big)^2\big(A(m) - e^{-m} + e^{-2m}\big) \\ &= \frac{1}{h}\lambda_0(s)\int_{\mathbb{R}} K^2(u)duA(m) + O(1).\end{aligned}$$

L'erreur quadratique moyenne intégrée est de la forme :

$$EQMI(h) = e^{-2m}\int_{\mathbb{R}} \lambda_0^2(s)ds - h^2e^{-m}(1-e^{-m})\int_{\mathbb{R}} \lambda_0(s)\lambda_0''(s)ds\int_{\mathbb{R}} u^2K(u)du + \frac{h^4}{4}(1-e^{-m})^2$$
$$\int_{\mathbb{R}} \lambda_0''(s)^2ds\Big(\int_{\mathbb{R}} u^2K(u)du\Big)^2 + O(h^6) + O(h^4e^{-m}) + \frac{1}{h}\int_{\mathbb{R}} K^2(u)duA(m) + O(1).$$

On a $A(m) = e^{-m}\sum_{k=1}^{\infty}\frac{m^k}{kk!} < e^{-m}\sum_{k=1}^{\infty}\frac{2m^k}{(k+1)!} = 2/m \Rightarrow A(m) \xrightarrow[m\to+\infty]{} 0$. Donc l'erreur quadratique moyenne intégrée tend vers 0 si on choisit h tel que $h \xrightarrow[m\to+\infty]{} 0$ et $\frac{A(m)}{h} \xrightarrow[m\to+\infty]{} 0$. L'ordre de grandeur de h optimal est : $h = O(A(m)^{1/5})$.

On a alors une erreur quadratique moyenne intégrée approximativement égale à :

$$EQMIA(h) = \frac{h^4}{4}(1-e^{-m})^2\int_{\mathbb{R}} \lambda_0''(s)^2ds\Big(\int_{\mathbb{R}} u^2K(u)du\Big)^2 + \frac{A(m)}{h}\int_{\mathbb{R}} K^2(u)du.$$

La valeur de h qui minimise cette quantité est :

$$h^* = \left(\frac{A(m)\int_{\mathbb{R}} K^2(u)du}{(1-e^{-m})^2\int_{\mathbb{R}} \lambda_0''(s)^2ds\Big(\int_{\mathbb{R}} u^2K(u)du\Big)^2}\right)^{1/5}$$

et conduit à $EQMIA(h^*) = O\big(A(m)^{4/5}(1-e^{-m})^{2/5}\big)$.

Le terme $\int_{\mathbb{R}} \lambda_0''(s)^2ds$ est inconnu et son estimation est largement traitée dans le cadre de l'estimation de densité. Néanmoins, nous allons ici le supposer connu

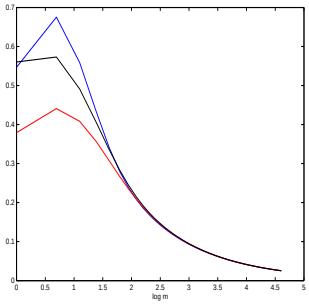

FIGURE 1. Profils des 3 termes asymptotiques

afin d'étudier l'influence de la variabilité du nombre d'événements. On supposera que h^* est estimée par $\hat{h^*} = \frac{1}{IN(1-e^{-N})^2}$, où $I = \frac{\int_{\mathbb{R}} \lambda_0''(s)^2 ds \left(\int_{\mathbb{R}} u^2 K(u) du \right)^2}{\int_{\mathbb{R}} K^2(u) du}$, $1/n$ estime $A(m) = \mathbb{E}\left(\frac{1}{N} \mathbb{1}[N \neq 0]\right)$ et $(1-e^{-n})^2$ estime $(1-e^{-m})^2$.

On peut alors calculer, par la même technique que précédemment :

$$EQMIA(\hat{h^*}) = O\Big(\Big[\sum_{j=1}^{\infty} \frac{e^{-m} m^j}{j^{2/5} j! (1-e^{-j})^{4/5}} \Big]^2 \Big) + O\Big(\sum_{j=1}^{\infty} \frac{e^{-m} m^j (1-e^{-j})^{4/5}}{j^{4/5} j!} \Big).$$

Ces deux termes, le premier représentant le biais et le second la variance, semblent très comparables à $O\big(A(m)^{4/5}(1-e^{-m})^{2/5}\big)$, comme le montrent les profils des fonctions $m \to \left[\sum_{j=1}^{\infty} \frac{e^{-m} m^j}{j^{2/5} j! (1-e^{-j})^{4/5}} \right]^2$ (bleu), $m \to \sum_{j=1}^{\infty} \frac{e^{-m} m^j (1-e^{-j})^{4/5}}{j^{4/5} j!}$ (rouge) et $m \to A(m)^{4/5}(1-e^{-m})^{2/5}$ (noire), représentés Figure 1. Ceci souligne la faible influence de la variabilité du nombre d'événements sur ces résultats asymptotiques.

Si l'on considère maintenant un domaine d'observation X borné, l'estimateur utilisé sera celui de Diggle, $\hat{\lambda}_{EC,h}(.)$. Asymptotiquement, lorsque la largeur de bande h tend vers 0, les estimateurs $\hat{\lambda}_{EC,h}(.)$ et $\hat{\lambda}_h(.)$ sont équivalents et on s'attend donc à retrouver la même erreur asymptotique, ce qui est montré par Cucala (2006).

Comme un processus stationnaire, un processus non stationnaire peut présenter des tendances à l'agrégation ou à la régularité. Afin de les identifier, Baddeley & *al.* (2000) ont récemment étendu la notion de fonction K pour des processus

non stationnaires et introduit la fonction :

$$K_{\text{inhom}}(t) = \frac{1}{|X|}\mathbb{E}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{I(||S_i - S_j|| < t)}{\lambda(S_i)\lambda(S_j)}.$$

Un estimateur de cette fonction s'obtient par le biais de l'estimateur de l'intensité $\hat{\lambda}_{EC,h}(.)$ par :

$$\hat{K}_{\text{inhom}}(t) = \frac{1}{|X|}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{I(||S_i - S_j|| < t)}{\hat{\lambda}_{EC,h}(S_i)\hat{\lambda}_{EC,h}(S_j)}.$$

Par simplicité, l'estimateur présenté ici ne contient pas de terme de correction de bord.

L'hypothèse d'un processus de Poisson inhomogène, pour lequel on obtient $K_{\text{inhom}}(t) = \pi t^2$, pourra être testée exactement comme l'hypothèse d'homogénéité spatiale l'est par l'intermédiaire de la fonction K.

CHAPITRE 6

INTENSITY ESTIMATION FOR PERTURBED POINT PROCESSES

Abstract

This article proposes new kernel estimators of the intensity function of spatial point processes taking into account position errors. The asymptotic properties of these estimators are derived. A simulation study compares their results to the results of the classical kernel estimator and shows that the edge-corrected deconvoluting kernel estimator is the most appropriate.

6.1. Introduction

In the theory of kernel density estimation, many authors have considered the problem of estimating the density from noisy observations. Indeed, one may consider that each measurement reflects the true value polluted by the addition of a stochastic error. This problem is usually handled by a deconvolution method, either when the distribution of the errors is known (Stefanski & Carroll, 1990) or unknown (Diggle & Hall, 1993).

When dealing with a spatial point pattern, a systematic exploratory tool is the intensity function, which is the equivalent of the trend for geostatistical data. Some authors (Ogata & Katsura, 1986) propose a parametric estimation of the intensity function but nonparametric estimators are more frequent. The most common nonparametric estimators of the intensity function are derived from the multivariate density estimation theory : mainly kernel and nearest-neighbour estimators (Cressie, 1993). Recently, a new approach based on a

hierarchical Bayesian model and the Voronoi tessellation has also been proposed (Heikkinen & Arjas, 1998 and Byers & Raftery, 2002).

All of these methods use locational data of the events, which are often difficult to collect and subsequently whose measurements are subject to errors. Lund & Rudemo (2000) try to make inference on such point processes observed with noise while Bar-Hen & *al.* (2005) study the influence of measurement errors on descriptive statistics used for testing the complete spatial randomness. In this paper, we propose a new kernel estimator of the intensity function which takes into account the location errors by a deconvolution method. For simplicity, we develop it in the case of bidimensional point processes.

Each kernel method is subject to a crucial choice, which is the bandwidth selection, much more important than the kernel choice itself (Silverman, 1986). This choice has been extensively discussed in the literature and original procedures have been proposed either for the deconvolution kernel density estimation problem (Delaigle & Gijbels, 2004) or for the kernel intensity estimation problem (Xu & *al.*, 2003).

Section 2 is an introduction to the perturbed point processes. We then define the new estimator and discuss its properties in Section 3. We present an asymptotic study in Section 4 and an adaptation of an existing bandwidth selection procedure to this specific problem in Section 5. The usefulness of the estimator is assessed by its application to simulated data in Section 6.

6.2. Perturbed point processes

Consider a Poisson point process $\mathbf{Y}$ in $\mathbb{R}^2$ with intensity function $\lambda_Y()$.

We only observe the point pattern $Z = \{z_1, \cdots, z_N\}$ in the domain $D \subset \mathbb{R}^2$ according to the model :

$$z_i = y_i + \epsilon_i, \tag{1}$$

where $(y_i : i = 1, \cdots, N)$ are events issued from the process $\mathbf{Y}$ and $(\epsilon_i : i = 1, \cdots, n)$ are i.i.d. with known isometric density function $g(.)$ and represent

the location errors. We will also assume that the errors ϵ_i are independant from the true locations y_i.

This additive error model is very common in statistics, for example in the regression framework (Carroll, Maca & Ruppert, 1999). It has been used in the point process framework by Bar-Hen & *al.* (2005). As in their paper, we denote by $\mathbf{Y}$ the unperturbed (true) point process and by $\mathbf{Z}$ the perturbed (observed) point process.

Our goal is to estimate the intensity function $\lambda_Y(s)$ for every point $s \in D$.

6.3. The deconvoluting kernel intensity estimators

Denote $\lambda_Z(.)$ the intensity function of the perturbed process $\mathbf{Z}$.

Based on the observations Z, the edge-corrected kernel estimator for $\lambda_Z(.)$ is (Diggle, 1985) :

$$\forall s \in \mathbb{R}^2, \hat{\lambda}_{Z,h}(s) = \begin{cases} \frac{\sum_{j=1}^{n} \frac{1}{h^2} K\left(\frac{s-z_j}{h}\right)}{\int_D \frac{1}{h^2} K\left(\frac{s-u}{h}\right) \nu(du)} & \text{if } \int_D \frac{1}{h^2} K\left(\frac{s-u}{h}\right) \nu(du) \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $K(.)$ is a kernel function and ν represents the Lebesgue measure.

From now on, we will assume that

$$\int_{\mathbb{R}^2} ||u|| |K(u)| \nu(du) < \infty. \tag{2}$$

The denominator $p_h(s) = \int_D \frac{1}{h^2} K\left(\frac{s-u}{h}\right) \nu(du)$ ensures that this estimator is asymptotically unbiased when $h \to 0$ and its practical usefulness has been shown (Zheng & *al.*, 2004). Denote $G_h = \{s \in \mathbb{R}^2 : p_h(s) \neq 0\}$.

The bidimensional Fourier transform of $g(.)$ is

$$\mathcal{F}(g)(t) = \int_{\mathbb{R}^2} e^{-it'z} g(z) \nu(dz)$$

where $z = (z_{(1)} z_{(2)})', t = (t_{(1)} t_{(2)})'$ and $t'z = t_{(1)} z_{(1)} + t_{(2)} z_{(2)}$.

Assume

$$\forall t \in \mathbb{R}^2, |\mathcal{F}(g)(t)| > 0. \tag{3}$$

In the density estimation framework, Stefanski & Carroll (1990) introduced a deconvoluting estimator adapted to noisy observations. Without taking into account the limited domain constraint, an estimator of $\lambda_Y(s)$ inspired by this is

$$\begin{aligned}\lambda^*_{Y,h}(s) &= \sum_{j=1}^{n} \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} e^{is't} \Big\{ \int_{\mathbb{R}^2} e^{-it'z} \frac{1}{h^2} K\big(\frac{z-z_j}{h}\big) \nu(dz)/\mathcal{F}(g)(t) \Big\} \nu(dt) \\ &= \sum_{j=1}^{n} \frac{1}{h^2} K^*_h\Big(\frac{s-z_j}{h}\Big),\end{aligned}$$

where $K^*_h(t) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} e^{it'y} \mathcal{F}(K)(y)/\mathcal{F}(g)(y/h) dy$.

Here, the choice of a band-limited kernel K, combined to (3), ensures that the inverse Fourier transform can be applied.

Now, from (1) we get $\lambda_Z(.) = \lambda_Y(.) * g(.) \Rightarrow \mathcal{F}(\lambda_Z)(.) = \mathcal{F}(\lambda_Y)(.)\ \mathcal{F}(g)(.)$ and a natural estimator of $\lambda_Y(s)$ is then

$$\begin{aligned}\hat{\lambda}_{Y,h}(s) &= \mathcal{F}^{-1}\big(\mathcal{F}(\hat{\lambda}_{Z,h})(t)/\mathcal{F}(g)(t)\big)(s) \\ &= \sum_{j=1}^{n} \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} e^{is't} \Big\{ \int_{G_h} \frac{e^{-it'z} \frac{1}{h^2} K\big(\frac{z-z_j}{h}\big)}{p_h(z)} \nu(dz)/\mathcal{F}(g)(t) \Big\} \nu(dt).\end{aligned}$$

Unfortunately, due to the presence of the edge-correction term, it is not clear to find a condition concerning the kernel K ensuring that the inverse Fourier transform can be applied. This is a main difference with the estimator $\lambda^*_{Y,h}$ previously introduced as it prevents its practical use.

A way of adapting the estimator $\lambda^*_{Y,h}$ to the limited domain context is to define

$$\lambda^{**}_{Y,h}(s) = \begin{cases} \frac{\lambda^*_{Y,h}(s)}{p^*_h(s)} & \text{if } p^*_h(s) \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $p^*_h(s)$ is an edge-correction term judiciously chosen, which will be discussed in the next section. We will denote $G'_h = \{s \in \mathbb{R}^2 : p^*_h(s) \neq 0\}$.

Whatever estimator we may choose, we can remark that, if the support of the density function g is $\mathbb{R}^2$, the estimation of λ_Y in each point of D requires the estimation of λ_Z in each point of $\mathbb{R}^2$. Here λ_Z may be highly underestimated in $\bar{D}$ as events in this domain are not observed. We believe that this makes the problem difficult and that the quality of any estimator will be affected by this drawback.

6.4. The asymptotic framework

In kernel density estimation in $\mathbb{R}^d$, the asymptotic framework is usually the following : the sample size n tends to infinity and the bandwidth h tends to 0 s.t. $nh^d \to \infty$, allowing the estimated density in every point to depend on an expected number of observations tending to infinity. In the point process theory, one often assumes that the expectation of the number of observed events N will tend to infinity with the size of the domain D : this is described as the increasing-domain asymptotics. However, in this case, with a given intensity, letting the bandwidth h tend to 0 implies that the estimated intensity in every point will depend on an expected number of events tending to 0.

The solution adopted by Diggle & Marron (1988) is to set up an increasing-intensity asymptotic framework. Denote $m_Y = \int_{\mathbb{R}^2} \lambda_Y(s)\nu(ds)$. Letting m_Y tend to infinity, Cucala & Thomas-Agnan (2006) obtain a consistent kernel estimator of $\frac{\lambda_Y(s)}{m_Y}$ in the error-free unbounded-domain case (no measurement error, unbounded domain). We decide to adopt the same scheme here so we will study the asymptotic behaviour of $\hat{\lambda}^0_{Y,h}(s) = \frac{\hat{\lambda}_{Y,h}(s)}{N}\mathbb{1}[N \neq 0]$ and $\lambda^{**0}_{Y,h}(s) = \frac{\lambda^{**}_{Y,h}(s)}{N}\mathbb{1}[N \neq 0]$ when m_Y tends to infinity.

Following the idea of Lahiri & *al.* (1999), it is also possible to set up a mixed asymptotic framework in which both the intensity and the observation domain increase to infinity, the first at a faster rate than the second.

Finally, let us mention two other alternative asymptotic frameworks. The first relies on replacing the increasing-domain asymptotic framework by an increasing-time asymptotic framework, as defined by Ellis (1991) : the length of the observation time T tends to infinity s.t. $Th^d \to \infty$ and the intensity is assumed to be constant in time. On the other hand, Kutoyants (1998) considers several realizations of the process on a finite domain and lets this number of realizations tend to infinity.

6.4.1. Preliminary results. — Denote $m_Z = \int_D \lambda_Z(s)\nu(ds)$ and $\lambda_Z^0(s) = \frac{\lambda_Z(s)}{m_Z}$, $\lambda_Y^0(s) = \frac{\lambda_Y(s)}{m_Z}$. It has been shown (Cucala & Thomas-Agnan, 2006) that, if we denote the random variable $X = \frac{1}{N}\sum_{i=1}^{N} f(Z_i)\mathbb{1}[N \neq 0]$, where $f()$ is any given measurable function,

$$\mathbb{E}X = (1 - e^{-m_Z}) \int_D f(s)\lambda_Z^0(s)\nu(ds) \tag{4}$$

and

$$\mathrm{Var}X = \int_D f^2(s)\lambda_Z^0(s)\nu(ds)A(m_Z) - \Big(\int_D f(s)\lambda_Z^0(s)\nu(ds)\Big)^2 \big(A(m_Z) - e^{-m_Z} + e^{-2m_Z}\big) \tag{5}$$

where $A(m_Z) = e^{-m_Z}\sum_{k=1}^{\infty} \frac{m_Z^k}{kk!} = \mathbb{E}\Big[\frac{1}{N}\mathbb{1}[N \neq 0]\Big]$.

6.4.2. Asymptotic bias of the estimator $\hat{\lambda}_{Y,h}$. — Even if the Fourier transform leading to this estimator is not ensured to be finite, we would like to know if a suitable choice of the kernel K can lead to an unbiased estimator.

From (4), we have $\mathbb{E}\hat{\lambda}_{Y,h}^0(s) = \dfrac{1 - e^{-m_Z}}{m_Z(2\pi)^2} \times$

$$\int_{\mathbb{R}^2} e^{is't} \Big\{ \int_{G_h} \frac{e^{-it'z}}{p_h(z)} \int_D \frac{1}{h^2} K\Big(\frac{z - x}{h}\Big)\lambda_Z(x)\nu(dx)\nu(dz)/\mathcal{F}(g)(t) \Big\}\nu(dt)$$

which finally leads to (see Appendix)

$$
\begin{aligned}
&\frac{m_Z(2\pi)^2}{1-e^{-m_Z}}\,\mathbb{E}(\hat{\lambda}^0_{Y,h}(s)) \\
&= \int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{G_h} e^{-it'(z-\epsilon)}\lambda_Y(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) \\
&- h\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{G_h}\frac{\int_{B_{z,h}} u_{(1)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial\lambda_Y}{\partial s_{(1)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) \\
&- h\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{G_h}\frac{\int_{B_{z,h}} u_{(2)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial\lambda_Y}{\partial s_{(2)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) \\
&+ \frac{h^2}{2}\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{G_h}\frac{\int_{B_{z,h}} u^2_{(1)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) \\
&+ h^2\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{G_h}\frac{\int_{B_{z,h}} u_{(1)}u_{(2)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) \\
&+ \frac{h^2}{2}\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{G_h}\frac{\int_{B_{z,h}} u^2_{(2)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) + O(h^3).
\end{aligned}
$$

Let $m_Z \to \infty$ and $h \to 0$.

First case : If the kernel K has a compact support, then as $h \to 0$, $G_h \to D$ in a monotone way and $\forall z \in D, B_{z,h} \to \mathbb{R}^2$ in a monotone way. Thus the expectation is asymptotically equal to

$$
\frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{D} e^{-it'(z-\epsilon)}\lambda^0_Y(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) + O(h^2).
$$

Second case : If the kernel K does not have a compact support, then $G_h = \mathbb{R}^2$ and $\forall z \in D, B_{z,h} \to \mathbb{R}^2$ in a monotone way. But $\forall z \in \bar{D}, B_{z,h}$ has no limit. Thus the bias is asymptotically equal to

$$\frac{1}{(2\pi)^2}\Bigg\{-h\int_{\mathbb{R}^2}e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\bar{D}}\frac{\int_{B_{z,h}}u_{(1)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial\lambda_Y}{\partial s_{(1)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2}e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)$$

$$-h\int_{\mathbb{R}^2}e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\bar{D}}\frac{\int_{B_{z,h}}u_{(2)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial\lambda_Y}{\partial s_{(2)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2}e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)$$

$$+\frac{h^2}{2}\int_{\mathbb{R}^2}e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\frac{\int_{B_{z,h}}u_{(1)}^2K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2}e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)$$

$$+h^2\int_{\mathbb{R}^2}e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\frac{\int_{B_{z,h}}u_{(1)}u_{(2)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2}e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)$$

$$+\frac{h^2}{2}\int_{\mathbb{R}^2}e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\frac{\int_{B_{z,h}}u_{(2)}^2K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2}e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)+O(h^3)\Bigg\}.$$

But now the terms depending on h, as $h\frac{\int_{B_{z,h}}u_{(1)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}$, are asymptotically equivalent to constants. Indeed, for example,

$$h\frac{\int_{B_{z,h}}u_{(1)}K(u)\nu(du)}{\int_{B_{z,h}}K(u)\nu(du)}=\frac{\int_{B_{z,1}}u_{(1)}K(u/h)\nu(du)}{\int_{B_{z,1}}K(u/h)\nu(du)}\in[\min_{B_{z,1}}u_{(1)};\max_{B_{z,1}}u_{(1)}].$$

So it seems that, whatever kernel one chooses, it is not possible that the estimator $\hat{\lambda}^0_{Y,h}(s)$ is asymptotically unbiased.

6.4.3. Asymptotic bias of the estimator $\lambda^{**}_{Y,h}$. — In this paragraph, we demonstrate the difficulty of finding an edge correction factor p^*_h so that the estimator $\lambda^{**}_{Y,h}$ is unbiased. But we propose a judicious choice of this edge correction factor leading to a tractable estimator and to asymptotic unbiasedness in the case of no measurement error and in the case of constant intensity.

We have, $\forall s\in G'_h$,

$$\mathbb{E}\lambda^{**0}_{Y,h}(s)=\frac{1-e^{-m_Z}}{m_Z(2\pi)^2p^*_h(s)}\int_{\mathbb{R}^2}e^{is't}\Big\{\int_{\mathbb{R}^2}e^{-it'z}\int_D\frac{1}{h^2}K\Big(\frac{z-x}{h}\Big)\lambda_Z(x)\nu(dx)\nu(dz)/\mathcal{F}(g)(t)\Big\}\nu(dt).$$

And the asymptotic expectation is, when K is a band-limited kernel

$$\begin{aligned}
&\frac{(2\pi)^{-2}}{p_h^*(s)}\Bigg\{\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\int_{B_{z,h}} K(u)\nu(du)e^{-it'(z-\epsilon)}\lambda_Y^0(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)\\
&- h\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\bar{D}}\int_{B_{z,h}} u_{(1)}K(u)\nu(du)e^{-it'(z-\epsilon)}\frac{\partial\lambda_Y^0}{\partial s_{(1)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)\\
&- h\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\bar{D}}\int_{B_{z,h}} u_{(2)}K(u)\nu(du)e^{-it'(z-\epsilon)}\frac{\partial\lambda_Y^0}{\partial s_{(2)}}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)\\
&+ \frac{h^2}{2}\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\int_{B_{z,h}} u_{(1)}^2K(u)\nu(du)e^{-it'(z-\epsilon)}\frac{\partial^2\lambda_Y^0}{\partial {s_{(1)}}^2}(z-\epsilon)\nu(dz)e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)\\
&+ h^2\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\int_{B_{z,h}} u_{(1)}u_{(2)}K(u)\nu(du)e^{-it'z}\frac{\partial^2\lambda_Y^0}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon)\nu(dz)g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt)\\
&+ \frac{h^2}{2}\int_{\mathbb{R}^2} e^{-is't}\Big\{\frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\int_{B_{z,h}} u_{(2)}^2K(u)\nu(du)e^{-it'z}\frac{\partial^2\lambda_Y^0}{\partial {s_{(2)}}^2}(z-\epsilon)\nu(dz)g(\epsilon)\nu(d\epsilon)}{\int_{\mathbb{R}^2} e^{-it'\epsilon}g(\epsilon)\nu(d\epsilon)}\Big\}\nu(dt) + O(h)^3\Bigg\}.
\end{aligned}$$

In this expression, the terms depending on h such as $h\int_{B_{z,h}} u_{(1)}K(u)\nu(du)$ are asymptotically negligible. Indeed, for example, from (2),

$$\big|h\int_{B_{z,h}} u_{(1)}K(u)\nu(du)\big| < h\int_{\mathbb{R}^2} ||u|||K(u)|\nu(du) \to 0.$$

We realize that the ideal edge-correction term $p_h^*(s)$, leading to asymptotic unbiasedness, should be $\mathcal{F}^{-1}\Big(\frac{\mathcal{F}\Big((\lambda_Y^0 * g)\times p_h\Big)(t)}{\mathcal{F}(g)(t)}\Big)(s)/\lambda_Y(s)$ which is of course unknown.

If we use this formula for a constant intensity, i.e. $\forall s \in \mathbb{R}^2, \lambda_Y^0(s) = 1$, we obtain

$$\begin{aligned} p_h^*(s) &= \mathcal{F}^{-1}\Big(\frac{\mathcal{F}(p_h)(t)}{\mathcal{F}(g)(t)}\Big)(s) \\ &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} e^{is't}\frac{\int_{\mathbb{R}^2} e^{-it'z}\frac{1}{h^2}\int_D K\big(\frac{z-u}{h}\big)\nu(du)\nu(dz)}{\int_{\mathbb{R}^2} e^{-it'z}g(z)\nu(dz)}\nu(dt) \\ &= \int_D \frac{1}{h^2}\frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} e^{is't}\frac{\int_{\mathbb{R}^2} e^{-it'z}K\big(\frac{z-u}{h}\big)\nu(dz)}{\int_{\mathbb{R}^2} e^{-it'z}g(z)\nu(dz)}\nu(dt)\nu(du) \\ &= \int_D \frac{1}{h^2}K_h^*\Big(\frac{s-u}{h}\Big)\nu(du). \end{aligned}$$

This edge-correction term is finite and $p_h^*(s)$ reduces to $p_h(s)$ when g reduces to a Dirac function (no measurement error).

Finally, we conclude that it seems that no consistent estimator is available for this complex problem. That is why we choose to focus on

$$\lambda_{Y,h}^{**}(s) = \frac{\sum_{j=1}^n \frac{1}{h^2}K_h^*\Big(\frac{s-z_j}{h}\Big)}{\int_D \frac{1}{h^2}K_h^*\Big(\frac{s-u}{h}\Big)\nu(du)}, \forall s \in G_h',$$

where K_h^* is the so-called deconvoluting kernel introduced by Stefanski & Carroll (1990).

Indeed, this estimator is much more tractable than $\hat{\lambda}_{Y,h}(s)$ as it uses the Fourier transform of the kernel K which is explicit, and the use of a band-limited kernel K ensures its existence. Moreover it reduces to Diggle's estimator when there is no measurement error.

6.4.4. Asymptotic variance of the estimator $\lambda^{**}_{Y,h}$. — The first integral from expression (5) is, $\forall s \in G'_h$,

$$
\begin{aligned}
B &= \int_D \left\{ \frac{1}{(2\pi)^2 p^*_h(s)} \int_{\mathbb{R}^2} e^{is't} \Big\{ \int_{\mathbb{R}^2} e^{-it'z} \frac{1}{h^2} K\big(\frac{z-x}{h}\big)\nu(dz)/\mathcal{F}(g)(t) \Big\}\nu(dt) \right\}^2 \lambda^0_Z(x)\nu(dx) \\
&= \frac{1}{(2\pi)^4 {p^*_h}^2(s)} \int_D \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} e^{is't} e^{is'v} \Big\{ \int_{\mathbb{R}^2} e^{-it'z} \frac{1}{h^2} K\Big(\frac{z-x}{h}\Big)\nu(dz)/\mathcal{F}(g)(t) \Big\} \\
&\quad \Big\{ \int_{\mathbb{R}^2} e^{-iv'w} \frac{1}{h^2} K\Big(\frac{w-x}{h}\Big)\nu(dw)/\mathcal{F}(g)(v) \Big\}\nu(dt)\nu(dv)\lambda^0_Z(x)\nu(dx) \\
&= \frac{1}{(2\pi)^4 {p^*_h}^2(s) m_Z} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \frac{e^{is't}e^{is'v}}{\mathcal{F}(g)(t)\mathcal{F}(g)(v)} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} e^{-it'z} e^{-iv'w} \\
&\quad \Big\{ \int_D \frac{1}{h^4} K\Big(\frac{z-x}{h}\Big) K\Big(\frac{w-x}{h}\Big) \lambda_Z(x)\nu(dx) \Big\}\nu(dw)\nu(dz)\nu(dt)\nu(dv).
\end{aligned}
$$

Now

$$
\int_D K\big(\frac{z-x}{h}\big) K\big(\frac{w-x}{h}\big) \frac{\lambda_Z(x)}{h^4}\nu(dx) = \int_{\mathbb{R}^2}\int_D K\big(\frac{z-x}{h}\big) K\big(\frac{w-x}{h}\big) \frac{\lambda_Y(x-\epsilon)}{h^4}\nu(dx)g(\epsilon)\nu(d\epsilon)
$$

and

$$
\begin{aligned}
& \int_D \frac{1}{h^4} K\big(\frac{z-x}{h}\big) K\big(\frac{w-x}{h}\big) \lambda_Y(x-\epsilon)\nu(dx) \\
= \;& \frac{1}{h^2} \int_{B_{z,h}} K(u) K\big(u - \frac{z-w}{h}\big) \lambda_Y(z-uh-\epsilon)\nu(du) \\
= \;& \frac{\lambda_Y(z-\epsilon)}{h^2} \int_{B_{z,h}} K(u) K\big(u - \frac{z-w}{h}\big)\nu(du) - \frac{1}{h}\frac{\partial \lambda_Y}{\partial s_{(1)}}(z-\epsilon) \int_{B_{z,h}} u_{(1)} K(u) \\
& K\big(u - \frac{z-w}{h}\big)\nu(du) - \frac{1}{h}\frac{\partial \lambda_Y}{\partial s_{(2)}}(z-\epsilon) \int_{B_{z,h}} u_{(2)} K(u) K\big(u - \frac{z-w}{h}\big)\nu(du) + O(1) \\
\sim_{h\to 0} \; & \frac{\lambda_Y(z-\epsilon)}{h^2} \int_{B_{z,h}} K^2(u) \mathbb{1}(z=w)\nu(du) - \frac{1}{h}\frac{\partial \lambda_Y}{\partial s_{(1)}}(z-\epsilon) \\
& \int_{B_{z,h}} u_{(1)} K^2(u) \mathbb{1}(z=w)\nu(du) - \frac{1}{h}\frac{\partial \lambda_Y}{\partial s_{(2)}}(z-\epsilon) \int_{B_{z,h}} u_{(2)} K^2(u) \mathbb{1}(z=w)\nu(du)
\end{aligned}
$$

So we get

$$B \sim_{h\to 0} \frac{1}{h^2(2\pi)^4 p_h^{*2}(s)} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \frac{e^{is't}e^{is'v}}{\mathcal{F}(g)(t)\mathcal{F}(g)(v)} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} e^{-it'(z-\epsilon)}e^{-iv'(z-\epsilon)}\lambda_Y^0(z-\epsilon) \int_{B_{z,h}} K^2(u)\nu(du)\nu(dz)e^{-it'\epsilon}e^{-iv'\epsilon}g(\epsilon)\nu(d\epsilon)\nu(dt)\nu(dv).$$

Moreover we have $\int_{B_{z,h}} K^2(u)\nu(du) \xrightarrow[h\to 0]{} \begin{cases} \int_{\mathbb{R}^2} K^2(u)\nu(du) & \text{if } z \in D, \\ 0 & \text{otherwise.} \end{cases}$

Consequently,

$$B \sim_{h\to 0} \frac{\int_{\mathbb{R}^2} K^2(u)\nu(du)}{h^2(2\pi)^4 p_h^{*2}(s)} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \frac{e^{is't}e^{is'v}}{\mathcal{F}(g)(t)\mathcal{F}(g)(v)} \int_{\mathbb{R}^2}\int_{D} e^{-it'(z-\epsilon)}e^{-iv'(z-\epsilon)}\lambda_Y^0(z-\epsilon) \nu(dz)e^{-it'\epsilon}e^{-iv'\epsilon}g(\epsilon)\nu(d\epsilon)\nu(dt)\nu(dv).$$

Now, the second integral from expression (5) is

$$C = \left\{ \int_D \frac{1}{(2\pi)^2 p_h^*(s)} \int_{\mathbb{R}^2} e^{is't} \Big\{ \int_{\mathbb{R}^2} e^{-it'z} \frac{1}{h^2} K\big(\frac{z-x}{h}\big)\nu(dz)/\mathcal{F}(g)(t) \Big\} \nu(dt)\lambda_Z^0(x)\nu(dx) \right\}^2$$
$$= O(p_h^*(s)^{-2}).$$

So the asymptotic variance of $\lambda_{Y,h}^{**0}(s)$ is the product of $\dfrac{A(m_Z)}{h^2} \dfrac{\int_{\mathbb{R}^2} K^2(u)\nu(du)}{(2\pi)^4 p_h^{*2}(s)}$ by

$$\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \frac{e^{is't}e^{is'v}}{\mathcal{F}(g)(t)\mathcal{F}(g)(v)} \int_{\mathbb{R}^2}\int_{D} e^{-it'(z-\epsilon)}e^{-iv'(z-\epsilon)}\lambda_Y^0(z-\epsilon)\nu(dz)e^{-it'\epsilon}e^{-iv'\epsilon}g(\epsilon)\nu(d\epsilon)\nu(dt)\nu(dv).$$

6.4.5. The link with the classical deconvolution estimator. — As mentioned before, we could also let the observation domain D tend to $\mathbb{R}^2$ and thus use an asymptotic framework similar to Lahiri & *al* (1999) 's mixed framework or Fuentes (2002) 's "shrinking asymptotics" framework.

For example, let $D = \gamma D_0$ with $\gamma \to \infty$ and $\frac{m_Z}{\gamma} \to \infty$ and consider a product kernel $K(s) = K_0(s^{(1)})K_0(s^{(2)}), \forall s = (s^{(1)}, s^{(2)})' \in \mathbb{R}^2$. In that case, the results of 4.3 and 4.4 would be exactly the two-dimensional equivalent of those

obtained by the estimator introduced by Stefanski & Carroll (1990) in the unbounded-domain case. By "equivalent" we mean that the asymptotic bias is $\frac{h^2}{2}\int_{\mathbb{R}} x^2 K_0(x)dx\nabla^2\lambda_Y^0(s)$ and the asymptotic variance is $\frac{\mathbb{E}(1/N)}{h^2}K_h^*(t)^2\lambda_Y^0(s)$, the only difference being the factor $\frac{1}{n}$ in the asymptotic variance term replaced by $A(m_Z) = \mathbb{E}\Big(\frac{1}{N}\Big)$.

6.5. The bandwidth selection procedure

Let us look for the bandwidth h minimizing the mean integrated square error (MISE)

$$MISE(h) = \mathbb{E}\int_D \{\lambda_{Y,h}^{**0}(s) - \lambda_Y^0(s)\}^2\nu(ds).$$

Due to their complexity and their dependance on the domain D, we will not use the expressions of the asymptotic bias and variance to set up a bandwidth selection procedure. Instead, we will rely on the procedures described by Delaigle & Gijbels (2004) in the unbounded-domain framework and adapt them to the bidimensional case.

The expression of the asymptotic MISE obtained by Stefanski & Carroll (1990) comes directly from the asymptotic bias and variance expressed before and Parseval's identity. We get for dimension 2

$$AMISE(h) = \frac{\mathbb{E}(1/N)}{(2\pi h)^2}\int_{\mathbb{R}^2}\frac{\mathcal{F}(K)(t)^2}{|\mathcal{F}(g)(t/h)|^2}\nu(dt) + \frac{h^4}{4}\alpha^2\int_{\mathbb{R}^2}\big(\nabla^2\lambda_Y^0(s)\big)^2\nu(ds),$$

where $\alpha = \int_{\mathbb{R}} x^2 K_0(x)dx$.

In order to minimize this expression, we need to estimate the term

$\int_{\mathbb{R}^2}\big(\nabla^2\lambda_Y^0(s)\big)^2\nu(ds)$. We propose to use a normal-reference rule. Suppose λ_Y^0 is the density of a Gaussian distribution with variance matrix

$\Sigma = \begin{pmatrix} \sigma_{Y,1}^2 & \rho_Y \sigma_{Y,1}\sigma_{Y,2} \\ \rho_Y \sigma_{Y,1}\sigma_{Y,2} & \sigma_{Y,2}^2 \end{pmatrix}$, then we have $\int_{\mathbb{R}^2} \left(\nabla^2 \lambda_Y^0(s)\right)^2 \nu(ds)$

$$
\begin{aligned}
&= \frac{(\sigma_{Y,1}^2+\sigma_{Y,2}^2)^2}{4\pi\sigma_{Y,1}^5\sigma_{Y,2}^5(1-\rho_Y^2)^{5/2}} - \frac{\sigma_{Y,1}^2+\sigma_{Y,2}^2}{4\pi\sigma_{Y,1}^3\sigma_{Y,2}^5(1-\rho_Y^2)^{3/2}} + \frac{3}{16\pi\sigma_{Y,1}\sigma_{Y,2}^5\sqrt{1-\rho_Y^2}} \\
&- \frac{\sqrt{1-\rho_Y^2}\left(\sigma_{Y,1}^4\rho_Y^2+\sigma_{Y,1}^2\sigma_{Y,2}^2(1+\rho_Y^2)+\sigma_{Y,2}^4\right)}{4\pi\sigma_{Y,1}^5\sigma_{Y,2}^5(1-\rho_Y^2)^3} + \frac{(3\rho_Y^2\sigma_{Y,1}^2+\sigma_{Y,2}^2)\sqrt{1-\rho_Y^2}}{8\pi\sigma_{Y,1}^3\sigma_{Y,2}^5(1-\rho_Y^2)^2} \\
&+ \frac{3\sqrt{1-\rho_Y^2}(\rho_Y^2\sigma_{Y,1}^2+\sigma_{Y,2}^2)^2}{16\pi\sigma_{Y,1}^5\sigma_{Y,2}^5(1-\rho_Y^2)^3} \\
&= H(\sigma_{Y,1},\sigma_{Y,2},\rho_Y).
\end{aligned}
$$

Denote $\sigma_{Z,1}^2 = Var(z^{(1)})$, $\sigma_{Z,2}^2 = Var(z^{(2)})$ and $\rho_Z = \frac{Cov(z^{(1)},z^{(2)})}{\sqrt{Var(z^{(1)})}\sqrt{Var(z^{(2)})}}$, where z is distributed according to λ_Z^0.

Denote $\sigma_{\epsilon,1}^2 = Var(\epsilon^{(1)})$, $\sigma_{\epsilon,2}^2 = Var(\epsilon^{(2)})$ and $\rho_\epsilon = \frac{Cov(\epsilon^{(1)},\epsilon^{(2)})}{\sqrt{Var(\epsilon^{(1)})}\sqrt{Var(\epsilon^{(2)})}}$, where ϵ is distributed according to g.

Now we get $\sigma_{Z,1}^2 = \sigma_{Y,1}^2 + \sigma_{\epsilon,1}^2$, $\sigma_{Z,2}^2 = \sigma_{Y,2}^2 + \sigma_{\epsilon,2}^2$ and $\rho_Z = \frac{\rho_Y\sigma_{Y,1}\sigma_{Y,2}+\rho_\epsilon\sigma_{\epsilon,1}\sigma_{\epsilon,2}}{\sigma_{Z,1}\sigma_{Z,2}}$.

$\sigma_{Z,1}^2$ can be estimated by $\hat{\sigma}_{Z,1}^2 = \frac{1}{n}\sum_{i=1}^n (z_i^{(1)} - z^{\bar{(1)}})^2$, where $z^{\bar{(1)}} = \frac{1}{n}\sum_{i=1}^n z_i^{(1)}$.

$\sigma_{Z,2}^2$ can be estimated by $\hat{\sigma}_{Z,2}^2 = \frac{1}{n}\sum_{i=1}^n (z_i^{(2)} - z^{\bar{(2)}})^2$, where $z^{\bar{(2)}} = \frac{1}{n}\sum_{i=1}^n z_i^{(2)}$.

ρ_Z can be estimated by $\hat{\rho}_Z = \frac{\sum_{i=1}^n (z_i^{(1)}-z^{\bar{(1)}})(z_i^{(2)}-z^{\bar{(2)}})}{\sqrt{\sum_{i=1}^n (z_i^{(1)}-z^{\bar{(1)}})^2}\sqrt{\sum_{i=1}^n (z_i^{(2)}-z^{\bar{(2)}})^2}}$.

And finally an estimator of $\int_{\mathbb{R}^2} \left(\nabla^2 \lambda_Y^0(s)\right)^2 \nu(ds)$ is

$$
H\Big(\hat{\sigma}_{Z,1}^2 - \sigma_{\epsilon,1}^2, \hat{\sigma}_{Z,2}^2 - \sigma_{\epsilon,2}^2, \frac{\hat{\rho}_Z\hat{\sigma}_{Z,1}\hat{\sigma}_{Z,2} - \rho_\epsilon\sigma_{\epsilon,1}\sigma_{\epsilon,2}}{\sqrt{(\hat{\sigma}_{Z,1}^2 - \sigma_{\epsilon,1}^2)(\hat{\sigma}_{Z,2}^2 - \sigma_{\epsilon,2})^2}}\Big).
$$

On the other hand, $\mathbb{E}(1/N)$ will be estimated by $1/n$.

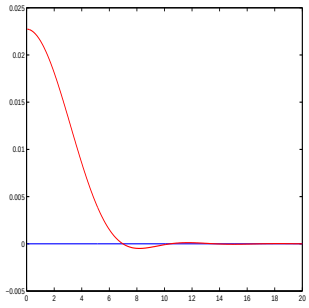

FIGURE 1. Profile of the kernel K_0

6.6. Computation of the estimator

6.6.1. A band-limited kernel. — As already said, the choice of the kernel is of secondary importance for the quality of our estimator. Here, for practical purpose, we choose a bidimensional kernel whose Fourier transform has compact support. The chosen kernel is a product kernel $K(x, y) = K_0(x)K_0(y)$, where

$$K_0(t) = \frac{48}{\pi} \frac{t^3 \cos(t) - 6t^2 \sin(t) + 15 \sin(t) - 15t \cos(t)}{t^7}$$

is a one-dimensional band-limited kernel also used by Delaigle & Gijbels (2004). Figure 1 gives its profile.

We notice that it is very similar to the triangular kernel. It can lead to negative values for $\lambda_Z(s)$ but a nonnegative kernel may also lead to negative values for $\lambda_Y(s)$ due to the deconvolution method.

6.6.2. The Fourier transforms. — The Fourier transform of the chosen kernel is

$$\mathcal{F}(K)(t) = (1 - t_1^2)^3 (1 - t_2^2)^3 \mathbb{1}_{[-1,1]^2}(t).$$

The Fourier transform of the density function of the errors g can usually be calculated analytically. For example, if the locational errors are normally distributed with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance matrix $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, then we have $\mathcal{F}(g)(t) = e^{-\frac{\sigma^2}{2}|t|^2}$. If the marginal locational errors are independent Laplace random variables with mean 0 and variance σ^2, we have $\mathcal{F}(g)(t) = \frac{1}{1+\sigma^2 t_1^2} \frac{1}{1+\sigma^2 t_2^2}$.

As in Stefanski & Carroll (1990), the inverse Fourier transforms are evaluated by a numerical Simpson procedure, slower but more accurate than the FFT procedure.

6.7. A simulation study

An inhomogeneous Poisson process is simulated in $[0,1]^2$ enlarged by a guard area with intensity

$$\lambda_Y(s) = C\big[1 + 0.7\cos\big(2\pi(||s|| - 0.5)\big)\big],$$

where C is a constant chosen such that the expected number of events in $[0,1]^2$ is 100. This is done by an acceptation-rejection method (Gentle, 2002).

The location errors $\{\epsilon_i,\ i = 1, \cdots, n\}$ are then simulated and added to the simulated locations :

$$z_i = y_i + \epsilon_i.$$

Only the observations z_i in $[0,1]^2$ will be used to estimate the intensity.

From the simulated sample, we compute the estimates $\hat{\lambda}_{Z,h_{opt}}$, λ^*_{Y,h^*} and λ^{**}_{Y,h^*}, where h_{opt} is the bandwidth obtained by the classical cross-validation procedure (Silverman, 1986) and h^* is the bandwidth obtained via the procedure described in section 5.

Denote $ISE = \int_{[0,1]^2} \big(\hat{\lambda}_{Z,h_{opt}} - \lambda_Y(s)\big)^2 \nu(ds)$,

$ISE^* = \int_{[0,1]^2} \big(\lambda^*_{Y,h^*}(s) - \lambda_Y(s)\big)^2 \nu(ds)$, $ISE^{**} = \int_{[0,1]^2} \big(\lambda^{**}_{Y,h^*}(s) - \lambda_Y(s)\big)^2 \nu(ds)$.

This procedure is repeated m times and we compute the empirical quartiles of ISE, ISE^* and ISE^{**}. Tables 1, 2 and 3 give the results when ϵ follows a Gaussian distribution with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance matrix $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, and the number m of realizations is equal to 10.

Tables 4, 5 and 6 give the results when ϵ follows a Laplace distribution with same mean and variance matrix, and the number m of realizations is equal to 10.

TABLE 1. Gaussian error, σ=0.02

	ISE	ISE^*	ISE^{**}
1st quartile ($*10^3$)	1.0600	1.6745	0.9038
median ($*10^3$)	1.3939	1.9613	1.0279
3rd quartile ($*10^3$)	1.5899	2.2432	1.3158

TABLE 2. Gaussian error, σ=0.05

	ISE	ISE^*	ISE^{**}
1st quartile ($*10^3$)	0.8185	1.4153	0.6655
median ($*10^3$)	1.2474	1.7199	0.9298
3rd quartile ($*10^3$)	1.5281	1.8908	1.2138

TABLE 3. Gaussian error, σ=0.1

	ISE	ISE^*	ISE^{**}
1st quartile ($*10^3$)	0.7669	1.2194	0.7223
median ($*10^3$)	0.8854	1.4123	0.8733
3rd quartile ($*10^3$)	1.4305	1.6451	1.2544

TABLE 4. Laplace error, σ=0.02

	ISE	ISE^*	ISE^{**}
1st quartile ($*10^3$)	1.0444	1.4676	0.8274
median ($*10^3$)	1.4129	1.7275	1.0025
3rd quartile ($*10^3$)	2.1357	1.9753	1.2334

TABLE 5. Laplace error, σ=0.05

	ISE	ISE^*	ISE^{**}
1st quartile ($*10^3$)	0.7869	1.1814	0.7689
median ($*10^3$)	1.4859	1.4223	1.1308
3rd quartile ($*10^3$)	2.0375	1.5114	1.4210

In each case, the estimator λ^{**}_{Y,h^*} gives the best results. The results of the estimator λ^{*}_{Y,h^*} are not better, or even worse, than the ones obtained by the classical Diggle estimator $\hat{\lambda}_{Z,h_{opt}}$, suggesting that deconvolution and edge-correction should both be considered when dealing with perturbed locations in a bounded domain.

TABLE 6. Laplace error, σ=0.1

	ISE	ISE^*	ISE^{**}
1st quartile ($*10^3$)	1.4211	1.2435	1.2350
median ($*10^3$)	1.7803	1.7003	1.4141
3rd quartile ($*10^3$)	2.1798	1.9612	1.6842

To get a better understanding of the use of the deconvolution kernel estimator, Figure 3 shows the contours of the true intensity and of the mean values of the three estimators when ϵ follows a Gaussian distribution with $\sigma = 0.05$.

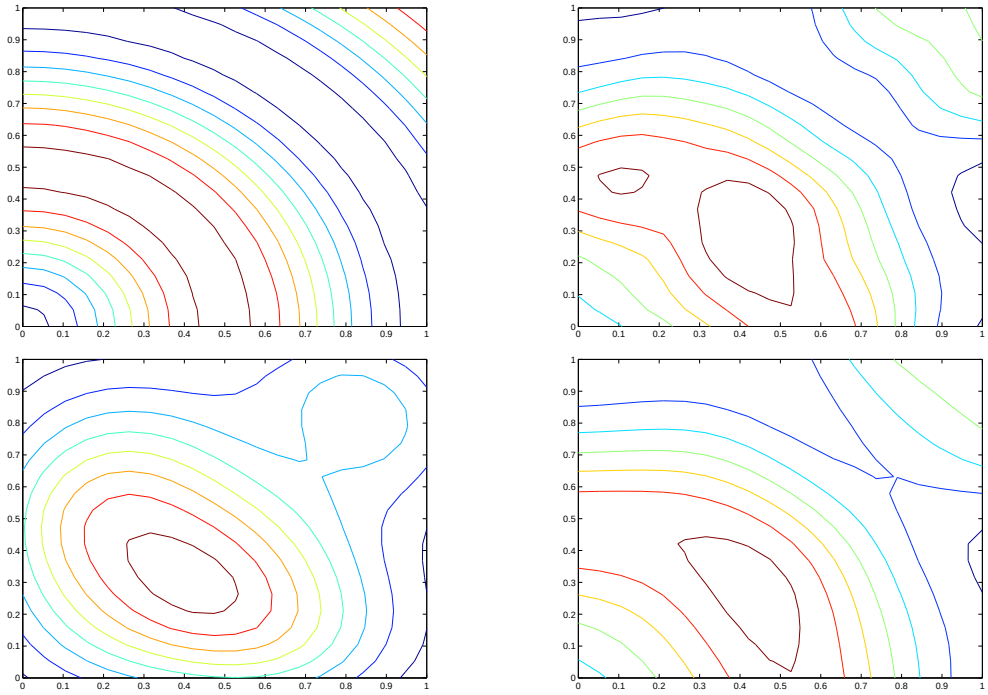

Figure 3 : Up-left figure : Contours of λ_Y. Up-right figure : Contours of $\hat{\lambda}_{Z,h_{opt}}$. Down-left figure : Contours of λ^*_{Y,h^*}. Down-right figure : Contours of λ^{**}_{Y,h^*}

It appears that the values taken by λ^*_{Y,h^*} close to the boundary of the square are too low, due to the absence of edge-correction. At the same time, the deconvolution technique used to get λ^{**}_{Y,h^*} leads to a better recognition of the peaks and troughs than the classical estimator $\hat{\lambda}_{Z,h_{opt}}$.

Finally, we consider how to handle uniform locational errors. Indeed, in this case, condition (3) is not satisfied and there is no appropriate deconvoluting intensity estimator. A solution can be to use the equivalent estimator for another error distribution. To illustrate this, Table 7 shows the results obtained when using the convoluting kernel estimator adapted to Laplace (index L) or Gaussian (index G) errors to uniform errors. The simulation procedure remains the same.

TABLE 7. uniform error, σ=0.05

	ISE	ISE_L^*	ISE_L^{**}	ISE_G^*	ISE_G^{**}
1st quartile ($*10^3$)	0.6939	1.3107	0.6804	1.3125	0.6823
median ($*10^3$)	1.0755	1.6191	1.0158	1.6181	1.0167
3rd quartile ($*10^3$)	1.1079	1.7944	1.1942	1.7955	1.2008

It appears that, even when the error distribution is misspecified, the deconvoluting kernel estimator remains useful. This goes along with the results of Hesse (1999) in the deconvoluting kernel density estimation framework asserting that the important point to specify is the error variance more than the error distribution.

6.8. An application to real data

In this section we illustrate our method on the spatial distributions of trees observed at Paracou site, which are data provided by the Forest department of CIRAD (Gourlet-Fleury & *al.*, 2004). This experimental station is located in the coastal part of French Guyana. It is composed of 14 experimental permanent sample plots of 6.25 ha each and one of 16 ha. In 1984, on each plot, all trees of diameter at breast height greater than 10 cm were localized by cartesian coordinates and botanically identified, when possible. The station is used for various ecological studies.

The trees were located in the following way : each plot was squared (12.5m $\times$ 12.5m) with ropes placed at the edge of the plot with decametre and compass. The coordinates of a tree were then measured with respect to the nearest origin (of the system of ropes axis) with decametre and compass (to keep the orthogonality). It can be noted that GPS is not well working around the equator and is not at all precise under canopy. Thus the trees were approximately localized independently of each other, with the same error that is a sum of the

metrology error, a bad estimation of the center of a tree whose trunk could be deformed (that is not circular) in tropical context, plus various entry errors (on the field, the coordinates were called out by the measurer to someone else who recorded the values). Finally, the localization errors are suspected to follow approximately a gaussian distribution with standard deviation equal to 4m.

Figure 4 presents the results obtained when applying both the classical Diggle estimator (on the left) and the deconvoluting kernel estimator (on the right) to one of the data sets from Paracou, representing the spatial distribution of a tree species called Dicorynia. The estimated standard deviation of the location errors is quite important here so that the strong aggregation exhibited by Diggle estimator becomes less obvious when applying the deconvolution estimator. This could also come from the different bandwidth selection procedure adapted to each estimator : a larger bandwidth leads to a smoother estimation.

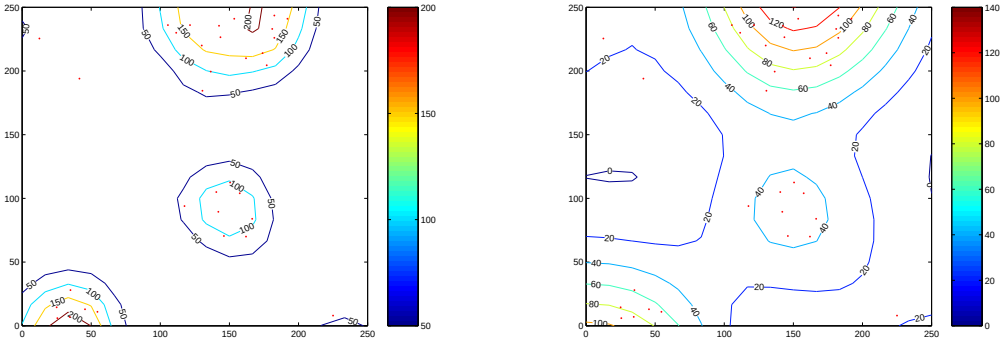

Figure 4 : Left figure : Contours of $\hat{\lambda}_{Z,h_{opt}}$. Right figure : Contours of λ^{**}_{Y,h^*}

Appendix

Denote

$$J = \int_D \frac{1}{h^2} K\big(\frac{z-x}{h}\big)\lambda_Z(x)\nu(dx) = \int_{\mathbb{R}^2}\int_D \frac{1}{h^2} K\big(\frac{z-x}{h}\big)\lambda_Y(x-\epsilon)\nu(dx)g(\epsilon)\nu(d\epsilon)$$
$$= \int_{\mathbb{R}^2}\int_{B_{z,h}} K(u)\lambda_Y(z-uh-\epsilon)\nu(du)g(\epsilon)\nu(d\epsilon),$$

where $B_{z,h} = \{\frac{z-x}{h} : x \in D\}$, as illustrated in Figure 2.

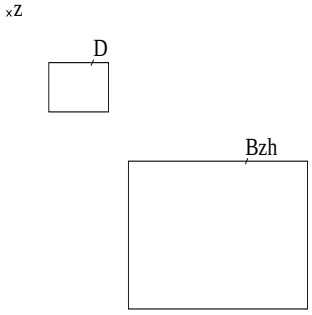

FIGURE 2. Illustration of the different sets

Then $\lambda_Y(z-\epsilon-uh) = \lambda_Y(z-\epsilon) - h\big(u_{(1)}\frac{\partial\lambda_Y}{\partial s_{(1)}}(z-\epsilon) + u_{(2)}\frac{\partial\lambda_Y}{\partial s_{(2)}}(z-\epsilon)\big) + \frac{h^2}{2}\big(u_{(1)}^2\frac{\partial^2\lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon) + u_{(1)}u_{(2)}\frac{\partial^2\lambda_Y}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon)\big) + u_{(2)}^2\frac{\partial^2\lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon)\big) + O(h^3)$.

So

$$
\begin{aligned}
J = \int_{\mathbb{R}^2} \Big\{ & \lambda_Y(z-\epsilon)\int_{B_{z,h}} K(u)\nu(du) - h\frac{\partial\lambda_Y}{\partial s_{(1)}}(z-\epsilon)\int_{B_{z,h}} u_{(1)}K(u)\nu(du) \\
& -h\frac{\partial\lambda_Y}{\partial s_{(2)}}(z-\epsilon)\int_{B_{z,h}} u_{(2)}K(u)\nu(du) + \frac{h^2}{2}\frac{\partial^2\lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon)\int_{B_{z,h}} u_{(1)}^2K(u)\nu(du) \\
& +h^2\frac{\partial^2\lambda_Y}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon)\int_{B_{z,h}} u_{(1)}u_{(2)}K(u)\nu(du) \\
& +\frac{h^2}{2}\frac{\partial^2\lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon)\int_{B_{z,h}} u_{(2)}^2K(u)\nu(du) + O(h^3)\Big\} g(\epsilon)\nu(d\epsilon).
\end{aligned}
$$

And

$$
\begin{aligned}
I &= \int_{G_h} \frac{e^{-it'z}}{p_h(z)} \int_D \frac{1}{h^2} K\big(\frac{z-x}{h}\big) \lambda_Z(x)\nu(dx)\nu(dz) \\
&= \int_{G_h} \frac{e^{-it'z}}{p_h(z)} \Bigg\{ \int_{\mathbb{R}^2} \Big\{ \lambda_Y(z-\epsilon) \int_{B_{z,h}} K(u)\nu(du) - h\frac{\partial \lambda_Y}{\partial s_{(1)}}(z-\epsilon) \int_{B_{z,h}} u_{(1)} K(u)\nu(du) \\
&- h\frac{\partial \lambda_Y}{\partial s_{(2)}}(z-\epsilon) \int_{B_{z,h}} u_{(2)} K(u)\nu(du) + \frac{h^2}{2}\frac{\partial^2 \lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon) \int_{B_{z,h}} u_{(1)}^2 K(u)\nu(du) \\
&+ h^2 \frac{\partial^2 \lambda_Y}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon) \int_{B_{z,h}} u_{(1)}u_{(2)} K(u)\nu(du) + \frac{h^2}{2}\frac{\partial^2 \lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon) \int_{B_{z,h}} u_{(2)}^2 K(u)\nu(du) \\
&+ O(h^3)\Big\} g(\epsilon)\nu(d\epsilon) \Bigg\} \nu(dz) \\
&= \int_{G_h} \frac{e^{-it'z}}{\int_{B_{z,h}} K(u)\nu(du)} \Bigg\{ \int_{\mathbb{R}^2} \Big\{ \lambda_Y(z-\epsilon) \int_{B_{z,h}} K(u)\nu(du) \\
&- h\frac{\partial \lambda_Y}{\partial s_{(1)}}(z-\epsilon) \int_{B_{z,h}} u_{(1)} K(u)\nu(du) - h\frac{\partial \lambda_Y}{\partial s_{(2)}}(z-\epsilon) \int_{B_{z,h}} u_{(2)} K(u)\nu(du) \\
&+ \frac{h^2}{2}\frac{\partial^2 \lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon) \int_{B_{z,h}} u_{(1)}^2 K(u)\nu(du) + h^2 \frac{\partial^2 \lambda_Y}{\partial s_{(1)}\partial s_{(2)}}(z-\epsilon) \int_{B_{z,h}} u_{(1)}u_{(2)} K(u)\nu(du) \\
&+ \frac{h^2}{2}\frac{\partial^2 \lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon) \int_{B_{z,h}} u_{(2)}^2 K(u)\nu(du) + O(h^3)\Big\} g(\epsilon)\nu(d\epsilon) \Bigg\} \nu(dz) \\
&= \int_{\mathbb{R}^2} \int_{G_h} e^{-it'(z-\epsilon)} \lambda_Y(z-\epsilon)\nu(dz) e^{-it'\epsilon} g(\epsilon)\nu(d\epsilon) \\
&- \int_{\mathbb{R}^2} \int_{G_h} \frac{h\int_{B_{z,h}} u_{(1)} K(u)\nu(du)}{\int_{B_{z,h}} K(u)\nu(du)} e^{-it'(z-\epsilon)} \frac{\partial \lambda_Y}{\partial s_{(1)}}(z-\epsilon)\nu(dz) e^{-it'\epsilon} g(\epsilon)\nu(d\epsilon) \\
&- \int_{\mathbb{R}^2} \int_{G_h} \frac{h\int_{B_{z,h}} u_{(2)} K(u)\nu(du)}{\int_{B_{z,h}} K(u)\nu(du)} e^{-it'(z-\epsilon)} \frac{\partial \lambda_Y}{\partial s_{(2)}}(z-\epsilon)\nu(dz) e^{-it'\epsilon} g(\epsilon)\nu(d\epsilon) \\
&+ \int_{\mathbb{R}^2} \int_{G_h} \frac{h^2\int_{B_{z,h}} u_{(1)}^2 K(u)\nu(du)}{2\int_{B_{z,h}} K(u)\nu(du)} e^{-it'(z-\epsilon)} \frac{\partial^2 \lambda_Y}{\partial {s_{(1)}}^2}(z-\epsilon)\nu(dz) e^{-it'\epsilon} g(\epsilon)\nu(d\epsilon) \\
&+ \int_{\mathbb{R}^2} \int_{G_h} \frac{h^2\int_{B_{z,h}} u_{(1)}u_{(2)} K(u)\nu(du)}{\int_{B_{z,h}} K(u)\nu(du)} e^{-it'(z-\epsilon)} \frac{\partial^2 \lambda_Y}{\partial s_{(1)}\partial s_{(1)}}(z-\epsilon)\nu(dz) e^{-it'\epsilon} g(\epsilon)\nu(d\epsilon) \\
&+ \int_{\mathbb{R}^2} \int_{G_h} \frac{h^2\int_{B_{z,h}} u_{(2)}^2 K(u)\nu(du)}{2\int_{B_{z,h}} K(u)\nu(du)} e^{-it'(z-\epsilon)} \frac{\partial^2 \lambda_Y}{\partial {s_{(2)}}^2}(z-\epsilon)\nu(dz) e^{-it'\epsilon} g(\epsilon)\nu(d\epsilon) + O(h^3).
\end{aligned}
$$

CONCLUSION

Dans cette thèse, nous nous intéressons à deux grandes questions des processus ponctuels spatiaux : les aspects distributionnels des espacements et l'estimation de l'intensité d'un processus bruité.

L'étude de la première question a tout d'abord abouti à la mise en place de tests de l'hypothèse d'homogénéité spatiale. Cette approche ayant été abondamment traitée dans le cadre unidimensionnel (Pyke, 1965), nous nous appuyons sur les techniques utilisées dans ce cadre et les étendons au cas spatial. La notion d'espacement bidimensionnel est tout d'abord introduite : notons que l'avantage d'utiliser les espacements ainsi définis aux distances entre évènements est que les premiers cités s'adaptent à toute inhomogénéité. Puis le deuxième lemme de Le Cam (1958), prouvant la normalité asymptotique de sommes de fonctions d'espacements uniformes, est étendu à $\mathbb{R}^2$. Enfin, des tests d'homogénéité spatiale sont construits et leurs performances comparées à de nombreux tests existants. De nombreuses perspectives se présentent. Dans un premier temps, la loi asymptotique des statistiques utilisées étant approchée uniquement pour des échantillons de très grande taille, nous envisageons d'obtenir une approximation de la distribution exacte de ces statistiques par l'intermédiaire d'une approximation point-selle (Gatto & Jammalamadaka, 1999). Ensuite, il existe, dans le cadre unidimensionnel, des transformations des espacements uniformes, ne modifiant aucunement la distribution mais permettant parfois d'augmenter la puissance des tests basés sur ces espacements (D'Agostino & Stephens, 1986). L'utilisation de telles transformations sur les espacements bidimensionnels est envisagée. Notons également que ces tests peuvent être étendus à n'importe quelle dimension.

Une extension naturelle de ce travail, développée dans une deuxième partie, est la recherche des zones de plus forte intensité des processus, appelées agrégats. En effet, alors que l'étude collective des espacements permet la création de tests globaux d'homogénéité spatiale, leur étude individuelle amène à tester l'existence locale d'agrégats. Cette démarche originale est dans un premier temps exposée dans le cadre unidimensionnel et l'estimateur d'agrégats est comparé à des estimateurs existants. Ensuite, nous exposons une technique de transformation d'un processus de Poisson homogène spatial en processus de Poisson homogène sur $\mathbb{R}$, qui nous permet d'étendre l'utilisation de l'estimateur d'agrégats à tout processus ponctuel spatial. Là encore se profilent de nombreuses perspectives à ce travail et notamment l'extension de ces techniques au cadre spatio-temporel, des jeux de données de ce type étant de plus en plus souvent disponibles.

Enfin, la deuxième question, traitée dans la dernière partie, est celle de l'estimation de l'intensité d'un processus bruité. En nous appuyant sur la méthode de déconvolution introduite en estimation de densité dans le cadre bruité (Stefanski & Carroll, 1990), nous construisons un estimateur à noyau de l'intensité tenant compte à la fois de la distribution des erreurs de localisation et du domaine d'observation borné. Les propriétés asymptotiques de cet estimateur sont étudiées puis nous évaluons ses performances sur des jeux de données réels ou simulés. La principale voie d'amélioration concerne la recherche de la largeur de bande optimale, problème souvent délicat lorsque l'on utilise un noyau. Pour le moment, le choix s'effectue par l'intermédiaire d'une approximation gaussienne et la mise en place de solutions plus exactes, inspirées de ce qui se fait en estimation de densité unidimensionnelle (Delaigle & Gijbels, 2004b), apparaît envisageable.

BIBLIOGRAPHIE

[1] Alam K., Abernathy R. and Williams, C. (1993). Multivariate goodness-of-fit tests based on statistically equivalent blocks. *Communications in Statistics. Theory and Methods*, **22**, 1515-1533.

[2] Baddeley, A., Møller, J. and Waagepetersen, R. (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54, 329-350.

[3] Baddeley, A., Turner, R., Møller J. and Hazelton M. (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society, Series B*, **67**, 1-35.

[4] Barber J. and Fuentes M. (2002) Nonstationary Spatial Process Modeling of Atmospheric Pollution Data. *Submitted to JABES*.

[5] Bar-Hen A., Chadœuf J., Dessard, H. and Monestiez, P. (2005). Estimating distance functions of point processes with known independent noise. *Preprint*.

[6] Bartlett, M.S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika*, **51**, 299-311.

[7] Barton, D.E. and David, F.N. (1956). Some notes on ordered intervals. *Journal of the Royal Statistical Society, Series B*, **18**, 79-94.

[8] Beirlant, J., Janssen, P. and Veraverbeke, N. (1991). On the asymptotic normality of functions of uniform spacings. *The Canadian Journal of Statistics*, **19**, 93-101.

[9] Bel L. (2002) Non parametric variogram estimation. Application to air pollution data. Fourth European Conference on Geostatistics for environmental applications, Barcelona.

[10] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

[11] Besag, J. and Diggle, P. (1977) Simple Monte-Carlo tests for spatial pattern. *Applied Statistics*, 26, 327-333.

[12] Biau, G. and Cadre, B. (2004) Nonparametric spatial prediction. *Statistical Inference and Stochastic Processes*, 7, 327-349.

[13] Bjørnstad O. and Falck, W. (2001) Nonparametric spatial covariance functions : estimation and testing. *Environmental and Ecological Statistics*, 8, 53-70.

[14] Bosq D. (1998) Nonparametric statistics for stochastic processes. Estimation and prediction. Second edition. *Lecture Notes in Statistics, 110.* Springer-Verlag.

[15] Brandt, M. (1995). Approximations for the distribution function of the sum of iid random variables with compact support in $\mathbb{R}^+$. *Preprint.*

[16] Burrows, P.M. (1979). Selected percentage points of Greenwood's statistic. *Journal of the Royal Statistical Society, Series A*, **142**, 256-258.

[17] Byers, S.D. and Raftery, A.E. (2002). Bayesian Estimation and Segmentation of Spatial Point Processes using Voronoi Tilings. In *Spatial Cluster Modelling* (Lawson and Denison, eds.), Chapman and Hall, London.

[18] Carroll, R., Maca, J. and Ruppert D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541-554.

[19] Chiu, S.N. (2003). Spatial point patterns analysis by using Voronoi diagrams and Delaunay tessellations-A comparative study. *Biometrical Journal*, **45**, 367-376.

[20] Choi, E. and Hall, P. (2000) On the estimation of poles in intensity functions. *Biometrika*, 87, 251-263.

[21] Cleveland W. and Devlin, S. (1988) Locally weighted regression : an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

[22] Cook, J.R. and Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314-1328.

[23] Cressie, N. (1976). On the logarithms of high-order spacings. *Biometrika*, **63**, 343-355.

[24] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

[25] Cressie, N. and Kornak J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical science*, **18**, 436-456.

[26] Cucala, L. (2005). Le Cam spacings theorem in dimension two. *Annales de l'Institut de statistiques de l'université de Paris*, **49**, 2-3, 41-55.

[27] Cucala, L. (2006). Intensity estimation for spatial point processes observed with noise. *Submitted*.

[28] Cucala, L. and Thomas-Agnan (2006a). Spacings-based tests for spatial randomness and coordinate-invariant procedures. *Annales de l'Institut de statistiques de l'université de Paris*, **50**, 1-2, 31-45.

[29] Cucala, L. and Thomas-Agnan, C. (2006b). Données spatiales. *Approches non-paramétriques en régression* (Droesbeke and Saporta, *editors*). Editions Technip, Paris.

[30] D'Agostino, R. and Stephens M. (1986). *Goodness-of-fit techniques*. Dekker, New York.

[31] Darling, D.A. (1953). On a class of problems relating to the random division of an interval. *The Annals of Mathematical Statistics*, **24**, 239-253.

[32] David, H.A. (1981). *Order statistics*. Second edition. Wiley, New York.

[33] Deheuvels, P. (1983a). Spacings and applications. *Probability and Statistical Decision Theory. Volume A* (F. Konecny, J. Mogyoródi and W. Wertz, *editors*). Reidel, Dordrecht, 1-30.

[34] Deheuvels, P. (1983b). Strong bounds for multidimensional spacings. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **64**, 411-424.

[35] Delaigle, A. and Gijbels I. (2004a). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society, Series B*, **64**, 869-886.

[36] Delaigle, A. and Gijbels I. (2004b). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, **45**, 249-267.

[37] Dematteï, C., Molinari, N. and Daurès J.P. (2005). Arbitrarily shaped multiple spatial cluster detection for case event data. A paraître dans *Computational Statistics and Data Analysis*.

[38] Diggle, P.J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, **35**, 87-101.

[39] Diggle, P. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.

[40] Diggle, P.J. (1985). A Kernel Method for Smoothing Point Process Data. *Applied Statistics*, **34**, 138-147.

[41] Diggle, P., Gates, D. and Stibbard, A. (1987) A nonparametric estimator for pairwise-interaction point processes. *Biometrika*, 74, 763-770.

[42] Diggle, P.J. and Hall P. (1993). A Fourier Approach to Nonparametric Deconvolution of a Density Estimate. *Journal of the Royal Statistical Society Series B*, **55**, 523-531.

[43] Diggle, P. and Marron, J. (1988) Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association*, 83, 793-800.

[44] Diggle, P.J. and Matérn, B. (1980). On sampling designs for the estimation of point-event nearest neighbour distributions. *Scandinavian Journal of Statistics*, **7**, 80-84.

[45] Does, R.J.M.M. and Klaassen, C.A.J. (1984). The Berry-Esseen theorem for functions of uniform spacings. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **65**, 461-472.

[46] Donnelly, K. (1978). Simulations to determine the variance and edge-effect of total nearest-neighbour distance. *Simulation Studies in Archaeology* (I. Hodder, *editor*). Cambridge University Press, London, 91-95.

[47] Duchon J. (1976) Fonctions splines et espérances conditionnelles de champs gaussiens. *Annales Scientifiques de l'Université de Clermont-Ferrand II. Mathématiques*, 14, 19-27.

[48] Ellis, S. (1991) Density estimation for point processes. *Stochastic Processes and their Applications*, 39, 345-358.

[49] Elogne, S. Perrin, O. and Thomas-Agnan C., (2006) Non parametric estimation of smooth stationary covariance functions by interpolation methods. *Preprint*.

[50] Fan, J. and Truong Y.K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, **21**, 1900-1925.

[51] Fiksel, T. (1988) Edge-corrected density estimators for point processes. *Statistics*, 19, 67-75.

[52] Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197-210.

[53] Fuller, W.A. (1987). *Measurement error models*. Wiley, New York.

[54] Gasser T., Müller H.G., (1979). Kernel estimation of regression functions. *Smoothing Techniques for curve estimation* (Gasser and Rosenblatt, *editors*). Heidelberg, Springer-Verlag.

[55] Gatto, R. and Rao Jammalamadaka, S. (1999). A conditional saddlepoint approximation for testing problems. *Journal of the American Statistical Association*, **94**, 533-541.

[56] Gentle, J. (2002). *Elements of computational statistics*. Springer.

[57] Genton M.G. and Gorsich D.J. (2002). Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices, *Computational Statistics and Data Analysis* 41, 47-57.

[58] Gourlet-Fleury S., Ferry B., Molino J.-F., Petronelli P. and Schmitt L. (2004). Experimental Plots : Key Features. *Ecology and Management of a Neotropical Rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana* (S. Gourlet-Fleury, J.-M. Guehl and O. Laroussinie, *editors*). Elsevier, Paris, 3-60.

[59] Greenwood, M. (1946). The statistical study of infectious diseases. *Journal of the Royal Statistical Society, Series A*, **109**, 85-110.

[60] Guillot G., Senoussi R. and Monestiez P. (2000). A positive definite estimator of the covariance of non stationary random fields. in *Proceedings of Geoenv 2000, 3rd international conference on geostatistics for environmental applications*. Kluwer Academic pub., 333-344.

[61] Guttorp, P. and Lockhart, R. (1988). On the asymptotic distribution of quadratic forms in uniform order statistics. *The Annals of Statistics*, **16**, 433-449.

[62] Hall P., Fischer N. and Hoffmann B. (1994). On The Nonparametric Estimation of Covariances Functions. *Annals of Statistics*, 22(4), 2115-2134.

[63] Hall P. and Patil P. (1994). Properties of Nonparametric Estimators of Autocovariance for Stationary Randoms Fields. *Probability Theory and Related Fields*, 99, 399-424.

[64] Hanisch, K. (1984) Some remarks on estimators of the distribution function of nearest neighbor distance in stationary spatial point processes. *Mathematische Operationsforschung und Statistik, Series Statistics*, 15, 409-412.

[65] Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, **25**, 435-450.

[66] Hesse, C. (1999). Data-driven deconvolution. *Journal of Nonparametric Statistics*, **10**, 343-373.

[67] Huntington, R.J. and Naus, J.I. (1975). A simpler expression for kth nearest neighbor coincidence probabilities. *Annals of Probability*, **5**, 894-896.

[68] Janson, S. (1987). Maximal spacings in several dimensions. *Annals of Probability*, **15**, 274-280.

[69] Justel, A., Peña D. and Zamar R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Probability Letters*, **35**, 251-259.

[70] Kelsall, J.E. and Diggle P.J. (1995a). Kernel estimation of relative risk. *Bernoulli*, **1**, 3-16.

[71] Kelsall, J.E. and Diggle P.J. (1995b). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medecine*, **14**, 2235-2342.

[72] Kent J.T. and Mardia K.V. (1994) Link between kriging and thin plate splines. In Festschrift Volume to P. Whittle : *Probability, Statistics and Optimisation,* ed Kelly FP. Wiley 324-339.

[73] Kim H. and Boos, D. (2004) Variance estimation in spatial regression using a non-parametric semivariogram based on residuals. *Scandinavian Journal of Statistics*, 31, 387-401.

[74] Kimeldorf G.S. and Wahba G. (1970a) Spline functions and stochastic processes. *Sankhya A*, 32, 173-180.

[75] Kimeldorf G.S. and Wahba G. (1970b) A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41, 495-502.

[76] Klamer D.M. and Masry E. (1982) Polynomial interpolation of randomly sampled bandlimited functions and processes, *SIAM Journal on Applied Mathematics*, 42, 1004-1019.

[77] Knox, G. (1959). Secular pattern of congenital oesophageal atresia. *British Journal of Preventive Social Medecine*, **13**, 222-226.

[78] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics. Theory and Methods*, **6**, 1481-1496.

[79] Kutoyants Y. (1998). Statistical inference for spatial Poisson processes. *Lecture Notes in Statistics, 134,* Springer.

[80] Lahiri, S.N., Kaiser, M.S., Cressie, N. and Hsu, N. (1999). Prediction of spatial cumulative distribution functions using subsampling. With comments and a rejoinder by the authors. *Journal of the American Statistical Association*, **94**, 86-110.

[81] Lawson, A.B. (1977). On tests for spatial trend in a non-homogeneous Poisson process. *Journal of Applied Statistics*, **15**, 225-234.

[82] Lawson, A.B. (1993). On the analysis of mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society, Series A*, **156**, 363-377.

[83] Lawson, A.B. (2001). *Statistical methods in spatial epidemiology*. Wiley, Chichester.

[84] 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*, **7**, 7-16.

[85] Lee T. and Berman, M. (1997) Nonparametric estimation and simulation of two-dimensional gaussian image textures. *Graphical models and image processing*, 59, 434-445.

[86] Liebetrau, A.M. (1977). Tests of randomness in two dimensions. *Communications in Statistics : Theory and Methods*, **6**, 1367-1383.

[87] Lund, J. and Rudemo M. (2000). Models for point processes observed with noise. *Biometrika*, **87**, 235-249.

[88] Masry E.(1983). Nonparametric covariance estimation from irregularly-spaced data, *Adv. in Appl. Probab.*, 15, 113-132.

[89] Matérn, B. (1960). Spatial variation : stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden Fran Statens Skogsforskningsinstitut, Band 49, Nr. 5, Stockholm*.

[90] Matheron G. (1981). Splines and Kriging : their formal equivalence. *Computer applications in the earth sciences : an update of the 70's* (D.F. Merriam, *editor*). Plenum Press, New-York.

[91] Molinari, N., Bonaldi C. and Daurès J.P. (2001). Multiple temporal cluster detection. *Biometrics*, **57**, 577-583.

[92] Moller, J. and Waagepetersen, R.P. (2004). *Statistical inference and simulation for spatial point processes*. Chapman & Hall, Boca Raton.

[93] Moran, P.A.P. (1947). The random division of an interval. *Journal of the Royal Statistical Society Series B*, **9**, 92-98.

[94] Mugglestone, M.A. and Renshaw, E. (1996). A practical guide to the spectral analysis of spatial point processes. *Computational Statistics and Data Analysis*, **21**, 43-65.

[95] Mugglestone, M.A. and Renshaw, E. (2001). Spectral tests of randomness for spatial point patterns. *Environmental and Ecological Statistics*, **8**, 237-251.

[96] Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in Medecine*, **15**, 845-850.

[97] Naus, J.I. (1965). he distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, **61**, 532-538.

[98] Ogata, Y. and Katsura K. (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics*, **40**, 29-39.

[99] Opsomer J., Ruppert D. and Wand, M. (1999) Kriging with nonparametric variance function estimation. *Biometrics*, 55, 704-710.

[100] Parzen E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, 3, 167-190.

[101] Proschan, F. and Pyke, R. (1964). Asymptotic normality of certain test statistics of exponentiality. *Biometrika*, **51**, 253-256.

[102] Pyke, R. (1965). Spacings (with discussion). *Journal of the Royal Statistical Society, Series B*, **27**, 395-449.

[103] Rao Jammalamadaka, S. and Goria, M.N. (2004). A test of goodness-of-fit based on Gini's index of spacings. *Statistics and Probability Letters*, **68**, 177-187.

[104] Ripley, B.D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, **13**, 255-266.

[105] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathemetical Statistics*, **23**, 470-472.

[106] Rudin W. (1990). Fourier analysis on groups, Wiley.

[107] Sampson P. D. and Guttorp P. (1992). Nonparametric representation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87, 108-119.

[108] Scott, D.W. (1992). *Multivariate density estimation. Theory, practice, and visualization.*. Wiley, New York.

[109] Selezn jev O. (2000). Spline approximation of random processes and design problems. *Journal of statistical planning and inference*, 84, 249-262.

[110] Shapiro A. and Botha J.D. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics and Data Analysis* 11, 87-96.

[111] Sherman, B. (1950). A random variable related to the spacing of sample values. *Annals of Mathematical Statistics*, **21**, 339-361.

[112] Sibson R. (1981) A brief description of natural neighbor interpolation. In *Interpreting multivariate data*, V. Barnett eds., Wiley, 21-36.

[113] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

[114] Simes86 Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751-754.

[115] Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169-184.

[116] Stoyan, D. and Stoyan, H. (1994). *Fractals, random shapes and point fields*. Wiley, New York.

[117] Strauss, D.J. (1975). A model for clustering. *Biometrika*, **62**, 467-475.

[118] Tango, T. (1984). The detection of disease clustering in time. *Biometrics*, **40**, 15-26.

[119] Thomas-Agnan C. (1991). Spline functions and stochastic filtering. *Annals of Statistics*, 19, 1512-1527.

[120] Tukey J. (1977) Exploratory data analysis. Addison-Wesley.

[121] Van Der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[122] Wahba G. (1990) Spline models for observational data. *CBMS 59, SIAM,* Philadelphia.

[123] Weba M. (1992). Simulation and approximation of Stochastic processes by splines functions. *SIAM Journal on Scientific and Statistical Computing*, 13, 1085-1096.

[124] Wolpert, R. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85, 251-267.

[125] Xu, C., Dowd P.A., Mardia K.V. and Fowell, R.J. (2003). Stochastic Approaches to Fracture Modelling. *Proceedings of IAMG.*

[126] Yaglom, A.M. (1957). Some classes of random fields in n-dimensional space, related to stationary random processes. *Theory of Probability and its Applications* 2, 273-320.

[127] Youness G. and Saporta G. (2004). Une méthodologie pour la comparaison de partitions . *Revue de Statistique Appliquée*, 52, 97-120.

[128] Zheng, P., Durr P. and Diggle P. (2004). Edge-correction for Spatial Kernel Smoothing - When Is It Necessary ? *Proceedings of the GisVet Conference 2004, University of Guelph, Ontario, Canada, June 2004.*

[129] Zimmerman, D.L. (1993). A Bivariate Cramer-Von Mises Type of Test for Spatial Randomness. *Applied Statistics*, **42**, 43-54.

[130] Zimmerman, D.L. (1994). On the limiting distribution of and critical values for an origin-invariant bivariate Cramer-Von Mises-type statistic. *Statistics and Probability Letters*, **20**, 187-195.