Statistiques spatiales avec applications à l'écologie et à l'économie

Eric Marcon

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture Agriculture Alimentation Food Biologie Biology Environnement Environment Santé Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Ecologie

Unité Mixte de Recherche ECOlogie des FOrêts de Guyane
cnrs INRA AgroParisTech cirad

présentée et soutenue publiquement par

Eric MARCON

le 16 novembre 2010

Statistiques spatiales avec applications à l'écologie et à l'économie

Directeur de thèse : **Gabriel LANG**
Co-encadrement de la thèse : **Jean-Pierre PASCAL**

Jury

M. Pierre Couteron, Directeur de Recherches, IRD	Président
M. Claude MILLIER, Directeur de Recherches honoraire, AgroParisTech	Rapporteur
M. François HOULLIER, Ingénieur Général des Ponts, des Eaux et des Forêts, INRA	Rapporteur
M. Jean-Pierre PASCAL, Directeur de Recherches honoraire, CNRS	Directeur
M. Gabriel LANG, Ingénieur en Chef des Ponts, des Eaux et des Forêts, AgroParisTech	Directeur

AgroParisTech
UMR Ecologie des Forêts de Guyane
Campus agronomique - BP 316 - 97379 Kourou cedex

Résumé

Les statistiques spatiales sont pour les écologues un ensemble d'outils permettant de caractériser la structure d'un semis de points, par exemple une carte représentant des emplacements des arbres dans une parcelle de forêt. Cette structure est définie implicitement comme un écart à semis complètement aléatoire, résultat d'un processus de Poisson. L'hétérogénéité du processus ponctuel dont le semis de points est une réalisation et la non-indépendance des points en sont les causes indiscernables, ce qui amène généralement à supposer l'intensité du processus connu et nommer « concentration spatiale » ou « agrégation » (la régularité spatiale est possible, mais rare en pratique) la non-indépendance. Une revue de la littérature des processus ponctuels et des mesures de structures spatiales est fournie pour clarifier les concepts et les choix.

L'objectif de ce travail de thèse était de produire des améliorations méthodologiques. Ses résultats principaux sont :

- L'établissement d'un test pour la principale statistique utilisée, la fonction K de Ripley, permettant de s'affranchir de la méthode de Monte-Carlo utilisée dans la littérature pour rejeter l'hypothèse nulle d'un processus complètement aléatoire.
- L'extension de K aux processus hétérogènes, dans le cadre d'une typologie claire des statistiques (absolues, relatives, topographiques).

Lorsque la position exacte des objets n'est pas connue, mais que des effectifs par zone sont disponibles (par exemple des nombres d'arbres par parcelle), la théorie de l'information est utilisée pour définir un cadre général permettant de caractériser la structure spatiale (des espèces dans les parcelles) et la diversité (des parcelles, en termes d'espèces) comme deux aspects d'une même mesure d'inégalité. Ce cadre est appliqué à l'indice de biodiversité de Shannon pour définir clairement la mesure de diversité bêta, son calcul direct indépendamment de la différence entre diversités gamma et alpha, et fournir un test statistique de non nullité. La voie est ouverte pour l'application à d'autres mesures de diversité et de structure spatiale.

En conclusion, il semble clair que ces outils de caractérisation sont un premier pas pour traiter les questions écologiques, leur développement étant toujours du domaine de la recherche. Ils sont cependant très insuffisants pour répondre à des questions liées aux processus écologiques, assez éloignés des processus ponctuels qui ignorent l'aspect temporel de l'installation des objets.

Les opinions émises par les auteurs sont personnelles et n'engagent pas l'UMR EcoFoG ou ses tutelles.

STATISTIQUES SPATIALES AVEC APPLICATIONS À L'ÉCOLOGIE ET À L'ÉCONOMIE

Table des matières

INTRODUCTION

Pourquoi étudier les structures spatiales ?

L'analyse des structures spatiales peut être considérée tout d'abord comme un outil statistique élémentaire. Quand on observe une carte sur laquelle sont placés des objets, ils ne semblent en général pas répartis au hasard. On peut voir des variations de densité, sous la forme de gradients ou d'agrégats. Pour décrire un phénomène quantitatif à partir d'un certain nombre d'observations, disons le diamètre moyen des arbres d'un peuplement, on utilise classiquement l'outil fourni par la théorie des variables aléatoires : on mesure un certain nombre d'arbres, on considère que leur diamètre moyen est un estimateur de l'espérance de la variable aléatoire qu'est le diamètre de chaque arbre, et on calcule la variance de l'échantillon pour avoir une idée de la variabilité des diamètres. Si on s'intéresse maintenant à la répartition spatiale de ces arbres, la quantité d'information disponible est bien plus importante : la densité des arbres, leur diamètre moyen selon le lieu et les interactions éventuelles des arbres entre eux. Les outils mathématiques nécessaires sont les processus ponctuels, généralisation spatialisée (généralement à deux dimensions) des variables aléatoires.

La question fondamentale consiste à caractériser clairement ce qu'on observe. Prenons l'exemple d'un peuplement forestier dans lequel les arbres d'une espèce sont apparemment regroupés et forment des agrégats. Deux causes peuvent en être responsables : l'hétérogénéité du milieu favorise la présence de cette espèce localement, la densité des arbres est alors variable, ou bien les arbres sont regroupés par des processus d'agrégation (par exemple une faible dispersion des graines) indépendamment du milieu qu'on supposera dans ce cas homogène. Les mécanismes écologiques sont extrêmement différents dans les deux cas. Mais la distinction entre hétérogénéité et agrégation est formellement impossible sur un jeu de données unique : des hypothèses de travail sont indispensables. Cette difficulté est probablement la raison pour laquelle de nombreuses méthodes statistiques ont été mises en place pour détecter les structures spatiales. Une revue complète en sera faite (annexe 2), la famille d'outils la plus efficace sera étudiée en détail (chapitre 1) ainsi que ses fondements mathématiques (les processus ponctuels, annexe 1) et de nouveaux outils seront proposés pour repousser les limites actuelles.

La relation directe entre les processus ponctuels et les phénomènes écologiques sous-jacents est difficile, à cause de la simplification excessive mais surtout parce

que la réalisation d'un processus observée sous la forme d'un semis de points est instantanée et unique alors qu'un peuplement forestier est le résultat d'une histoire et de processus successifs. Quelques pistes seront données en conclusion, mais il ne s'agit pas du centre de ce travail, qui se concentre sur la méthodologie.

La théorie des processus ponctuels est indispensable pour comprendre les méthodes présentées ici. Une présentation en est faite en annexe 1 plutôt que dans le texte parce qu'il s'agit d'une compilation de la littérature et non de développements nouveaux.

Le premier chapitre présente l'état de l'art en termes d'analyse spatiale des processus ponctuels suivi de l'introduction de nouveaux outils permettant d'étendre son champ d'application. L'objectif est de permettre la caractérisation des structures spatiales dans un cadre théorique aussi vaste que possible et répondant aux exigences des données réelles, comme la prise en compte de l'hétérogénéité. L'approche utilisée est celle de Ripley (1976; 1977). Les résultats principaux sont :

- Le calcul de l'intervalle de confiance de la fonction K (Lang et Marcon, 2010) et son application sous la forme d'un test de l'hypothèse nulle d'une distribution complètement aléatoire (Marcon *et al.*, in prep.),
- L'extension de K aux processus hétérogènes (Marcon et Puech, 2009; 2010).

Les autres méthodes de caractérisation des semis de points sont passées en revue en annexe 2 : Méthodes alternatives en statistiques spatiales continues.

La deuxième partie traite la situation où les données ne sont disponibles que sous la forme d'effectifs dans des zones, sans précision sur la position exacte des objets. C'est le cas de loin le plus fréquent, les inventaires forestiers par exemple sont fournis sous la forme de nombre d'arbres par espèce et par parcelle. L'approche utilisée est celle de l'entropie, qui fournit un cadre théorique permettant des développements nouveaux. Les résultats principaux sont :

- L'unification des mesures de concentration spatiale et de diversité, comme des cas particuliers de mesure de l'écart entre une distribution observée et une distribution attendue,
- L'application de la méthode générale de la décomposition de l'entropie à l'indice de biodiversité de Shannon, permettant la formulation analytique de la diversité β (Marcon *et al.*, in prep).

De nombreux outils ont été développés pour caractériser la structure spatiale de données discrètes. Ceux qui ne contribuent pas directement aux avancées présen-

tées dans les deux parties principales sont passées en revue en Annexe 3 : Méthodes alternatives en statistiques spatiales discrètes.

Une approche pratique de la caractérisation des structures spatiales est proposée en conclusion, ainsi qu'une ouverture sur les possibilités de modélisation des processus écologiques. Enfin, les codes informatiques et les textes des publications sont fournis en annexes 4 et 5.

Notations

Statistiques spatiales continues, processus ponctuels

$\mathbf{1}(b)$: la fonction indicatrice, $\mathbf{1}(b) = 1$ si b est vrai, 0 sinon.

$\propto$: l'opérateur de proportionnalité entre fonctions réelles.

$\sim$: l'opérateur de voisinage (entre points), ou l'opérateur « est distribué comme » (entre fonctions).

$\setminus$: l'opérateur de soustraction entre semis de points.

α : le seuil de risque des tests statistiques, $\alpha = 5\%$ ou $\alpha = 1\%$.

A : l'aire d'étude, $A \subset \mathbb{R}^2$, et, selon le contexte, sa surface.

$b(x, r)$: le disque fermé (la boule fermée si le nombre de dimensions est supérieur à 2) de centre x et de rayon r.

$c(i, j, r)$: le facteur de correction des effets de bord pour le point de référence x_i et son voisin x_j dans le calcul d'une fonction de voisinage à la distance r.

d : la dimension du vecteur des distances $(r_1, \dots, r_i, \dots, r_d)$ choisies pour calculer les statistiques.

dx : la surface infinitésimale autour du point x. $dx = \lim_{r \to 0} \big(b(x, r)\big)$.

$\mathbb{E}(Z)$: l'espérance de la variable aléatoire Z.

$f(\Xi)$: une fonction de densité de probabilité par rapport au processus Ξ.

$h(\Xi)$: une fonction de densité de mesure non normalisée à 1 par rapport au processus Ξ, $h \propto f$.

L_{ir} : longueur de l'arc de cercle de rayon r autour du point x_i située à l'intérieur de l'aire d'étude. $2\pi r / L_{ir}$ est le facteur de correction des effets de bord selon Ripley.

$\lambda(x)$: l'intensité du processus au point x, appelée *fonction d'intensité*.

λ : l'intensité du processus quand elle est constante (processus homogène).

$\hat{\lambda}$: l'estimateur de l'intensité d'un processus homogène.

$\lambda_{\mathbf{2}}(x_1, x_2)$: densité du produit de second ordre du processus ponctuel.

$\lambda^*(X, x)$: l'intensité conditionnelle de Papangelou.

$\mu(S)$: l'intensité du processus sur la surface S, appelée *mesure d'intensité*.

$\mu_2(S_1, S_2)$: mesure du moment factoriel de second ordre du processus ponctuel.

N_f : l'ensemble des semis de points finis de A, $N_f = \{X \subset A, n(X) < \infty\}$

$\mathcal{N}_f$: la tribu (σ-algèbre) engendrée par N_f.

$N\big(b(x,r)\big)$: le nombre de points (aléatoire) dans le cercle (la boule) de centre x et de rayon r.

$n\big(b(x,r)\big)$: le nombre de points observés dans le cercle (la boule) de centre x et de rayon r.

$N(S)$: le nombre de points (aléatoire) dans la surface S.

$n(S)$: le nombre de points observés dans la surface S. $n(A)$ est le nombre de points dans l'aire d'étude. Le nombre de cas est $n_c(S)$ et le nombre de contrôles $n_0(S)$.

$P(\quad)$: une probabilité.

ϕ_n : la fonction d'interaction entre les n-uplets de points de X.

R : une distance maximale (définissant un voisinage ou encore une distance d'interaction).

r : :une distance. $(r_1, \dots, r_i, \dots, r_d)$ est le vecteur des distances choisies pour calculer les statistiques.

S : une surface dans A, $S \subseteq \mathrm{A}$.

$t_\alpha(d)$: la valeur de la fonction de Student au seuil de risque α, pour le nombre de degrés de libertés d. $t_{5\%}(d) \approx 1{,}96$ pour $d > 500$.

$\mathrm{Var}(Z)$: la variance de la variable aléatoire Z.

X : un semis de points, $X \subset \mathbb{R}^2$. X est une réalisation du processus ponctuel Ξ.

X_A : la partie observable d'un semis de points dans l'aire d'étude, $X_A = X \cap A$ donc $X_A \subset X$ et $X_A \subset A$.

$v_c(i,r)$: le nombre de cas voisins du point x_i.

$\bar{v}(r)$: le nombre moyen de voisins à la distance r.

$w(x)$: le poids attribué au point x, par exemple la surface terrière d'un arbre.

W : le poids cumulé de tous les points. $W = \sum_{i=1}^{n(A)} w(x_i)$.

W_c : le poids cumulé de tous les cas. $W_c = \sum_{i=1}^{n_c(A)} w(x_i^c)$.

x : un point, $x \in X$. x_i^1 est la notation pour le point de type *1* et d'indice i. Quand les points sont séparés en cas et contrôles, l'exposant est respectivement c et 0.

(y,m) : un point marqué, $(y,m) \in Y$.

Ξ (Xi) : un processus ponctuel, $\Xi: \mathbb{R}^2 \to \mathbb{R},\ X \mapsto P(X)$

Y : un semis de points marqués, $Y \subset \mathbb{R}^2 \times \mathbb{R}^+$ (l'espace des marques peut être différent selon le cas traité). Y est une réalisation du processus ponctuel Υ.

Y_n : un état d'une chaîne de Markov.

Υ (Upsilon) : un processus ponctuel marqué, $\Upsilon: \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R},\ Y \mapsto P(Y)$.

Z : un semis de points inclus dans X, $Z \subseteq X$.

$Z(x)$: une variable aléatoire définie en un point du plan, $Z(x)$: $\mathbb{R}^2 \rightarrow \mathbb{R}^+$, utilisée pour définir l'intensité locale d'un processus de Cox.

$\{Z(x)\}$: l'ensemble des variables $Z(x)$ pour tout $x \in \mathbb{R}^2$, appelé un champ aléatoire.

Entropie

I : le nombre de placettes.

i : l'indice des placettes.

n : le nombre d'individus échantillonnés.

p_s : la probabilité de tirer l'espèce s, ou, selon le contexte, sa fréquence observée.

p_{si} : la probabilité de tirer l'espèce s dans la placette i.

q_{si} : la fréquence observée de l'espèce s dans la placette i.

S : le nombre d'espèces.

s : l'indice des espèces.

$t_\alpha(d)$: la valeur de la fonction de Student au seuil de risque α, pour le nombre de degrés de libertés d. $t_{5\%}(d) \approx 1{,}96$ pour $d > 500$.

y_{si} : le nombre d'arbres de l'espèce s dans la placette i, y_{+i} est le nombre d'arbres de la placette i, toutes espèces confondues. y_{s+} est le nombre d'arbres total de l'espèce s.

CARACTÉRISATION DE LA STRUCTURE SPATIALE DES PROCESSUS PONCTUELS

Les travaux théoriques fondamentaux sur la structure des processus ponctuels (Ripley, 1976; 1977) ont mis en place un cadre théorique clair quoique limité :

- Les semis de points observés peuvent être considérés comme la réalisation de processus ponctuels décrits en détail dans l'annexe 1.
- La référence est un semis de points complètement aléatoire : tous les points sont distribués avec une probabilité égale partout, indépendamment les uns des autres.

La structure spatiale est l'écart à l'indépendance entre les points : la concentration, due à l'attraction entre les points, ou la dispersion, due à leur répulsion. Une présentation détaillée en est faite ici.

La significativité statistique des valeurs de concentration ou de répulsion observées est classiquement testée par la méthode de Monte-Carlo parce que la distribution de la statistique est inconnue. Un test analytique est introduit.

La possibilité de traiter des semis de points d'intensité variable a été fournie par Cuzick et Edwards (1990) dans le cadre d'un test non paramétrique (présenté page 171), mais le test ne permet que d'affirmer que deux distributions sont différentes. Pélissier et Goreaud (2001) ont proposé des méthodes permettant d'utiliser *K* dans le cas de discontinuités d'intensité du processus. Baddeley *et al.* (2000) ont généralisé la fonction *K* aux processus non stationnaires mais leur méthode n'est utilisable en pratique que lorsque l'intensité est connue autour des points, ce qui est rarement le cas. Pour faire reculer les limites empiriques, une nouvelle fonction permettant de traiter les processus d'intensité variable sans restriction est développée.

La fonction *K* de Ripley

La fonction *K* définie par Ripley (1976; 1977) est un bon indicateur de la structure spatiale (Besag, 1977 ; Diggle, 1983 ; Cressie, 1993).

On considérera ici seulement des processus homogènes et isotropes.

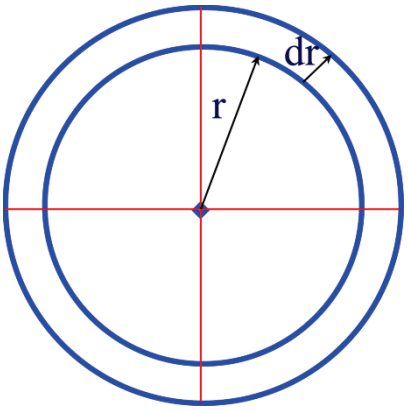
Introduction : Probabilité de trouver un voisin d'un point à une distance donnée

On définit les voisins du point x comme les points situés à une distance inférieure ou égale à une valeur donnée r. Le disque de rayon r autour du point x est noté $b(x,r)$. $g(\cdot)$ est la propriété de second ordre du processus ponctuel (équation (103), page 140). L'espérance du nombre de voisin est $\mathbb{E}\left(N\left(b(x,r)\right)\right)$. On démontre que :

$$\mathbb{E}\left(N\left(b(x,r)\right)\right) = \int_0^r g(\rho) 2\lambda\pi r d\rho \qquad \textbf{(1)}$$

Démonstration :

Considérons un cercle de rayon r autour du point x. Augmentons légèrement le rayon r, d'une valeur dr. La surface du cercle augmente d'une valeur dS, surface de la couronne, égale à $d(\pi r^2)$, soit :



$$dS = 2\pi r dr$$

La probabilité de trouver un point dans cette surface élémentaire est λdS (équation (99), page 139), soit :

$$P(N(dS) = 1) = 2\lambda\pi r dr$$

On considère maintenant un cercle identique, mais centré sur un point appartenant au semis.

On cherche la probabilité conditionnelle $P(N(dS) = 1 | N(dx) = 1)$ de trouver un point dans la couronne sachant qu'un point se situe au centre du cercle. On doit donc diviser la probabilité jointe $P\left((N(dS) = 1) \cap (N(dx) = 1)\right)$ de trouver à la fois un point au centre du cercle et dans la couronne par la probabilité qu'un point se trouve au centre du cercle, $P(N(dx) = 1)$. Cette probabilité conditionnelle peut s'écrire :

$$P(N(dS) = 1 | \{x\}) = \frac{P\left((N(dS) = 1) \cap (N(dx) = 1)\right)}{P(N(dx) = 1)}$$
$$= \frac{g(r)P(N(dS) = 1)P(N(dx) = 1)}{P(N(dx) = 1)}$$

$= g(r)2\lambda\pi r dr$

Ici encore, la probabilité de trouver 2 points dans la couronne est négligeable. L'espérance du nombre de points à l'intérieur du cercle de rayon *r* autour d'un point est donc :

$$\mathbb{E}\left(N\left(b(x,r)\right)\right) = \int_0^r g(\rho)2\lambda\pi\rho d\rho$$

■

Il vient immédiatement de ce qui précède :

$$\frac{\mathbb{E}\left(N\left(b(x,r)\right)\right)}{\lambda} = \int_0^r g(\rho)2\pi\rho d\rho \qquad \textbf{(2)}$$

Définition de la fonction K

Ripley (1977) a défini la fonction K :

$$\boxed{K(r) = \int_0^r g(\rho)2\pi\rho d\rho} \qquad \textbf{(3)}$$

(3) : Définition de la fonction *K* de Ripley

Si les points sont distribués indépendamment les uns des autres (processus de Poisson), $g(\rho)$ vaut 1 pour toutes les valeurs de ρ, et $K(r) = \pi r^2$. Cette valeur sert de référence :

- $K(r) > \pi r^2$ indique qu'en moyenne $g(\rho)$ est supérieur à 1. La probabilité de trouver un voisin à la distance ρ est donc supérieure à la probabilité de trouver un point dans un lieu quelconque du domaine d'étude : les points sont agglomérés.
- Inversement, $K(r) < \pi r^2$ indique que la densité de voisins autour des points est moins grande que la densité moyenne sur l'ensemble du domaine d'étude. Les points se repoussent.

$K(r)$ est égal au rapport du nombre de voisins sur l'intensité. Son estimateur est donc le rapport du nombre moyen de voisins réellement observé, $n\left(b(x,r)\right)$, sur l'estimateur de l'intensité :

$$\widehat{K}(r) = \frac{n\big(b(x,r)\big)}{\hat{\lambda}} \quad \textbf{(4)}$$

Un estimateur naturel de l'intensité est le rapport entre le nombre de points observé et la surface de la fenêtre d'étude :

$$\hat{\lambda} = \frac{n(A)}{A} \quad \textbf{(5)}$$

On peut expliciter le nombre moyen de voisins en définissant l'indicatrice $\mathbf{1}\big(\|x_i - x_j\| \leq r\big)$ qui vaut 1 si la distance entre deux points x_i et x_j est inférieure à r. Chaque point est pris comme centre du cercle tour à tour, et tous les autres points sont testés comme voisins potentiels. Si le deuxième point est voisin du premier, on compte 1. La moyenne du nombre de voisins est obtenue en divisant la somme par le nombre de points. Il est clair que pour les centres proches de la limite de la fenêtre, le nombre de voisins est sous-estimé parce qu'une partie du cercle est située en dehors de la fenêtre, où la position des points n'est pas connue. Un facteur de *correction des effets de bord*, dépendant dans le cas le plus compliqué du point central, du voisin et de r, doit être utilisé. Il est noté $c(i,j,r)$. Sa forme exacte sera discutée plus loin, mais il est forcément supérieur ou égal à 1. Finalement :

$$n\big(b(x,r)\big) = \frac{1}{n(A)} \sum_{i=1}^{n(A)} \sum_{j=1, i\neq j}^{n(A)} \mathbf{1}\big(\|x_i - x_j\| \leq r\big)\, c(i,j,r) \quad \textbf{(6)}$$

En combinant les équations précédentes, on obtient l'estimateur de Ripley :

$$\boxed{\widehat{K}(r) = \frac{A}{\big(\mathrm{n}(A)\big)^2} \sum_{i=1}^{\mathrm{n}(A)} \sum_{j=1, i\neq j}^{\mathrm{n}(A)} \mathbf{1}\big(\|x_i - x_j\| \leq r\big)\, c(i,j,r)} \quad \textbf{(7)}$$

(7) : Estimation de la fonction *K* de Ripley

La fonction *L* de Besag

La fonction de Ripley n'est pas très pratique à utiliser. Comparer une valeur calculée à sa référence, πr^2, nécessite de nouveaux calculs, et la représentation graphique hyperbolique n'est pas très parlante.

Besag (1977) a proposé une normalisation de la fonction, dont la valeur de référence est 0 :

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r \qquad (8)$$

(8) : La fonction *L* de Besag

Dans la littérature, une forme alternative de L existe : $L(r) = \sqrt{\frac{K(r)}{\pi}}$. La confusion n'est pas possible : dans ce dernier cas, $L(r) = r$ pour un processus de Poisson. Illian *et al.* (2008) par exemple préfèrent utiliser cette forme pour montrer le caractère cumulatif de la fonction.

Significativité

Intervalle de confiance local

Les valeurs estimées des fonctions K et L sont comparées à leurs valeurs de référence correspondant à une distribution complètement aléatoire. Pour tester si une valeur de $\hat{L}(r)$ est significativement différente de 0, la technique la plus commune consiste à utiliser la méthode de Monte-Carlo :

- On génère un grand nombre N de jeux de données aléatoires correspondant à l'hypothèse nulle à tester.
- On se fixe un seuil d'incertitude α.
- Pour chaque valeur de r, on classe les valeurs obtenues de $\hat{L}(r)$ par ordre croissant. On note $\hat{L}_n(r)$ la nième valeur.
- On élimine les valeurs extrêmes. On retient comme bornes de l'intervalle de confiance de l'hypothèse nulle les valeurs $\hat{L}_{N\alpha/2+1}(r)$ et $\hat{L}_{N(1-\alpha/2)}(r)$. Pour N=1000 et $\alpha = 5\%$, on retient la 26ème et la 975ème valeur.
- On considère que $\hat{L}(r)$ est significativement différent de 0 si sa valeur observée est hors de l'intervalle de confiance de l'hypothèse nulle : $\left[\hat{L}_{N\alpha/2+1}(r); \hat{L}_{N(1-\alpha/2)}(r)\right]$.

Des tentatives de calcul de l'intervalle de confiance de l'hypothèse nulle existent dans la littérature. Ripley (1979) propose respectivement $\pm 1{,}42\frac{\sqrt{A}}{n(A)-1}$ et $\pm 1{,}68\frac{\sqrt{A}}{n(A)-1}$ comme approximation des intervalles aux seuils de 5% et 1%. Ces valeurs proviennent de simulations et sont des constantes. Faute de base théorique suffisante, ces valeurs sont très rarement employées (un des rares exemples d'application est donné par Szwagrzyk et Czerwczak, 1993) et les auteurs utilisent la méthode de Monte-Carlo. À titre d'exemple, on peut se reporter à la Figure 9, page 30 : les bornes de l'intervalle de confiance au seuil de 5% calculées par la méthode de Monte-Carlo sur 1000 simulations sont effectivement à

peu près constantes, de l'ordre de $\pm 0{,}1$ contre une valeur prévue par Ripley de l'ordre de $\pm 0{,}15$.

Les valeurs de Ripley ont été remises en cause par Koen (1991) mais avec des erreurs de calcul, corrigées par Chiu (2007), toujours à partir de simulations.

Intervalle de confiance global

La méthode précédente donne pour chaque valeur de L un intervalle de confiance pour r fixé, dit « local » (Goreaud, 2000 ; Duranton et Overman, 2005), même si le terme n'est pas le mieux choisi pour une fonction cumulative comme K ou L : si la valeur se trouve hors de l'intervalle de confiance, le risque qu'elle soit issue d'une distribution correspondant à l'hypothèse nulle est limité à α. En supposant que les valeurs soient distribuées indépendamment, pour un seuil α de 5% et une fonction calculée sur 100 pas, on attend 5 points hors de l'intervalle de confiance en restant dans le cadre de l'hypothèse nulle : si la courbe réellement observée sort de l'intervalle de confiance en un nombre réduit de points, il n'est donc pas possible de rejeter l'hypothèse nulle. En réalité, les valeurs de L sont très corrélées, ce qui limite fortement le risque de mauvaise interprétation, sans permettre de le quantifier.

Ce problème est considéré comme critique par Duranton et Overman (2005), mais ignoré par toute la littérature qui n'utilise à notre connaissance que des intervalles de confiance locaux. Goreaud (2000, page 56) observe sur un exemple que près de 10% des courbes générées par simulation de l'hypothèse nulle sortent au moins une fois de intervalle de confiance local calculé à 1% à partir de leurs valeurs. Il conclut qu'étant donné la complexité de la construction d'un intervalle de confiance global, l'utilisation d'un intervalle local à 1% est suffisante.

La définition d'un intervalle de confiance global nécessite de trouver deux courbes telles que le risque qu'une courbe L d'un semis de point respectant l'hypothèse nulle sorte de cet intervalle au moins une fois soit α, ceci sans privilégier une distance r particulière. Un choix assez naturel consiste à calculer des intervalles de confiance locaux à un seuil donné et à compter le nombre de courbes simulées sortant au moins une fois de ces limites pour connaître le risque global (comme Goreaud, 2000), mais en augmentant progressivement le seuil de risque local jusqu'à ce que le risque global atteigne le seuil choisi (Duranton et Overman, 2005 ; Marcon et Puech, 2010) :

- On génère un grand nombre N de jeux de données aléatoires correspondant à l'hypothèse nulle à tester.
- On élimine pour chaque distance r les l valeurs extrêmes (les $l/2$ plus grandes et les $l/2$ plus petites) de $\hat{L}(r)$. On compte le nombre g de courbes

auxquelles ces valeurs appartiennent. On a ainsi défini un intervalle de confiance local au seuil l/N, correspondant à un risque global g/N.

- On augmente progressivement la valeur de l (2, 4, …) jusqu'à ce que g/N atteigne le seuil choisi. L'intervalle de confiance global au seuil g/N est donc défini par l'intervalle local au seuil l/N. Il peut être nécessaire d'interpoler les valeurs de deux intervalles locaux si deux valeurs successives de l/N encadrent la valeur choisie pour le risque global sans l'égaler.

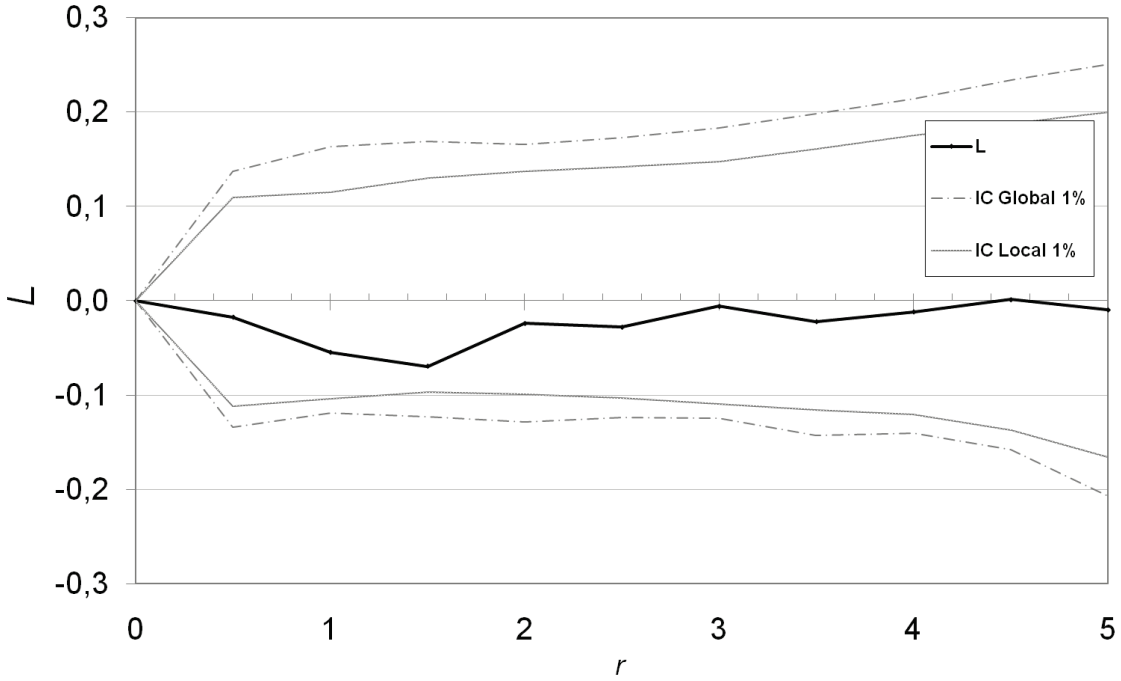


Figure 1 : Fonction *L* pour un processus de Poisson homogène
Les courbes légendées IC sont les intervalles de confiance de l'hypothèse nulle

Les résultats sont présentés Figure 1 pour un semis de points poissonnien (100 points dans un domaine carré de 10 de côté). Les intervalles de confiance local et global sont calculés au seuil de 1% sur 10 000 simulations. Le pas de calcul est de 0,5. Le seuil de risque local correspondant au risque global de 1% est 0,11%. A titre de comparaison avec Goreaud, le seuil local correspondant au risque global de 10% est pour cet exemple 0,97%.

Le problème de cette méthode est que ses résultats varient avec le nombre de pas du calcul. La Figure 2 traite les mêmes données que la Figure 1 mais avec un pas de calcul plus serré : 0,1. Le seuil de risque local correspondant au risque global de 1% est 0,06%, et 0,27% pour 10%, loin du 1% du calcul précédent. Les intervalles locaux sont en revanche peu sensibles à ce problème.

Le choix du pas de calcul est arbitraire, la méthode est donc peu satisfaisante sur le plan théorique. En pratique, la comparaison de la Figure 1 et de la Figure 2 montre que les valeurs des intervalles de confiance varient relativement peu.

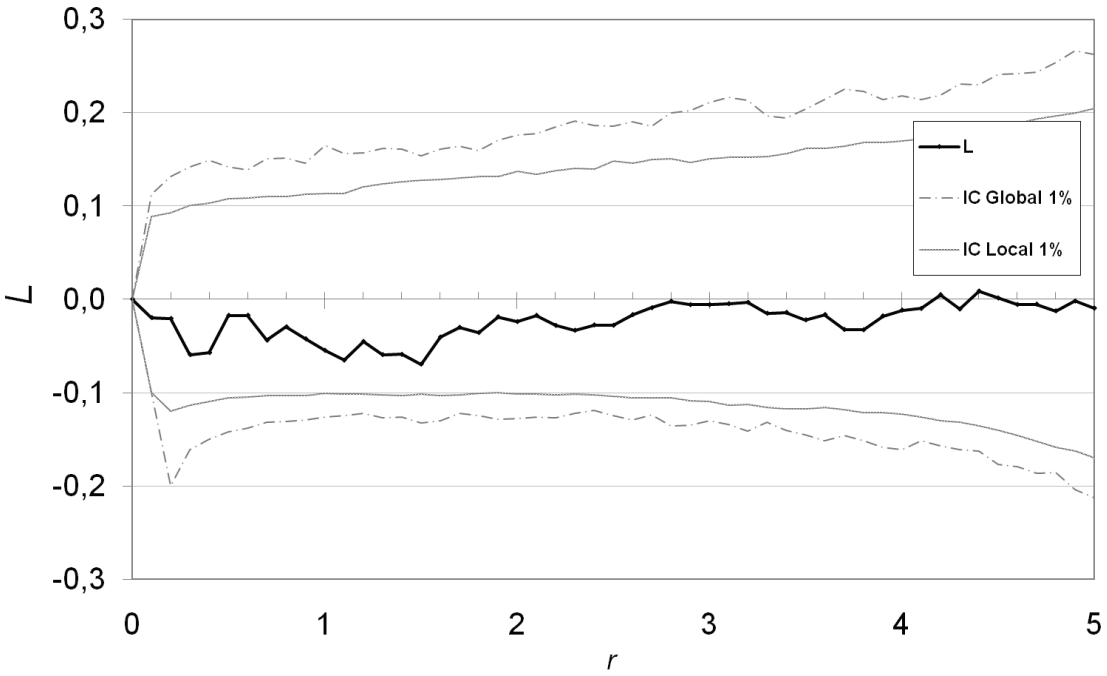


Figure 2 : Fonction *L* calculée avec un pas fin.
Le processus est le Poisson homogène de la Figure 1, la fonction est calculée avec un pas de 0 ,1 au lieu de 0,5.

Par la suite, nous calculerons systématiquement les intervalles de confiance globaux et locaux.

Correction des effets de bord

Les points situés près de la limite de la zone d'étude posent problème car une partie du cercle dans lequel on compte les voisins se trouve hors de la zone.

Négliger ces effets de bord entraine une sous-estimation de la fonction K. La Figure 3 correspond à un processus de Poisson de 100 points semblable à celui de la Figure 1, pour lequel aucune correction des effets de bord n'a été appliquée. Depuis la publication de la fonction K, la correction des effets de bord a fait l'objet de solutions variées.

La fonction de correction $c(i,j,r)$ de l'équation (7) ne dépend en fait jamais des trois variables : il existe des corrections globales où $c(r)$ corrige en une fois la valeur de $\widehat{K}(r)$, des corrections applicables à chaque point de référence, $c(i,r)$, et une correction dépendant des paires de points $c(i,j)$.

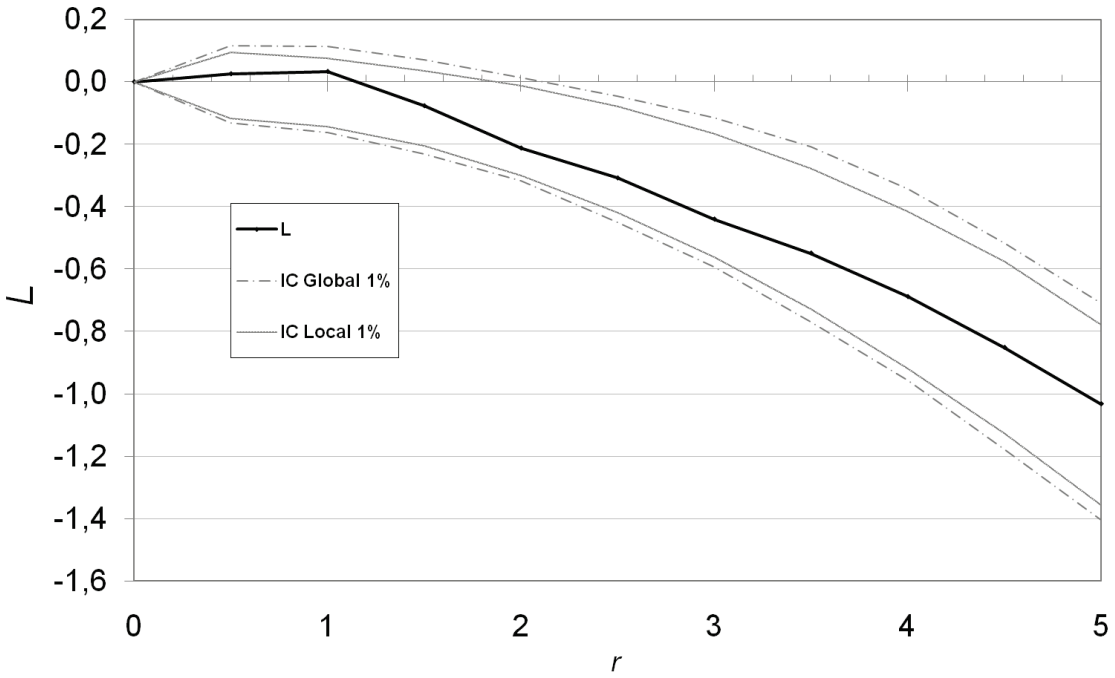


Figure 3 : Fonction *L* sans correction des effets de bord

La correction de Ripley

Ripley (1977) propose une forme de correction des effets de bord $c(i,r)$.

Si toute la couronne de largeur dr dans laquelle se trouve le voisin ne se trouve pas dans la zone d'étude, on attribue au point un poids égal au rapport entre la surface totale de la couronne et la partie de cette surface se trouvant à l'intérieur de la zone, en supposant que la partie de la couronne hors de la zone aurait pu contenir la même densité de voisins que la partie dans la zone.

Si on note L_{ir} la partie du périmètre du cercle de rayon r centré sur le point i située à l'intérieur de la zone d'étude, étant donné que dr est très petit, on obtient immédiatement la correction à apporter au poids de chaque point :

$$\widehat{K}(r) = \frac{A}{\left(n(A)\right)^2} \sum_{i=1}^{n(A)} \frac{2\pi r}{L_{ir}} \sum_{j=1, i\neq j}^{n(A)} \mathbf{1}\left(\|x_i - x_j\| \leq r\right) \quad \textbf{(9)}$$

La correction ne dépend que des centres de cercles x_i, pas des voisins x_j. Si le cercle est tout entier dans le domaine, $L_{ir} = 2\pi r$ et la correction disparaît.

La correction de Besag

Besag (1977), dans sa discussion de l'article de Ripley, estime que cette correction donne une importance excessive aux voisins les plus lointains. En effet, plus le

rayon augmente, plus L_{ir} est petit, et plus la correction est forte. Il propose une solution alternative, qui consiste à corriger les effets de bord non plus pour chaque voisin mais pour tous de la même manière.

On note A_{ir} la partie de la surface du cercle de rayon r centré sur le point x_i située à l'intérieur de la zone d'étude. On compte le nombre de voisins à l'intérieur du cercle et on corrige ce nombre par le rapport entre la surface du disque et la partie de cette surface à l'intérieur de la zone d'étude, en supposant que la partie du disque située hors de la zone aurait pu contenir la même densité de voisins que la partie dans la zone. Finalement :

$$\widehat{K}(r) = \frac{A}{\big(n(A)\big)^2} \sum_{i=1}^{n(A)} \frac{\pi r^2}{A_{ir}} \sum_{j=1, i \neq j}^{n(A)} \mathbf{1}\big(\|x_i - x_j\| \leq r\big) \qquad \textbf{(10)}$$

Ces deux méthodes de correction des effets de bord sont utilisées alternativement, avec une prédominance de la méthode de Ripley (voir par exemple Goreaud, 2000).

La correction de Ward et Ferrandino (1999)

Ward et Ferrandino (1999) rejettent la correction de Ripley « pour des raisons à la fois philosophiques et mathématiques ». Leur argumentation est fondée sur l'importance excessive accordée aux points proches des limites, dont le nombre de voisins est augmenté artificiellement par la correction des effets de bord. La position stochastique de ces points est responsable d'une augmentation de la variance de l'estimateur de *K*. Les auteurs proposent une correction globale des effets de bord valable pour un processus de Poisson homogène : pour une valeur de r donnée l'espérance du nombre de points proches de la bordure du domaine est calculée. Pour chacun de ces points, une correction est apportée par la méthode de Besag (proportionnellement à la surface du cercle hors du domaine). La sous-estimation du nombre de paires de points à la distance r est donc estimée en fonction de la seule géométrie du domaine d'étude. Pour un domaine rectangulaire de longueur L et des valeurs de r inférieures à la largeur l du rectangle, la correction globale est :

$$c(r) = 1 - \frac{4}{3\pi}\left(\frac{r}{L} - \frac{r}{l}\right) + \left(\frac{11}{3\pi} - 1\right)\left(\frac{r^2}{lL}\right) \qquad \textbf{(11)}$$

Appliquée à *K*, cette correction des effets de bord donne un nouvel estimateur de la fonction de Ripley pour un processus de Poisson homogène, appelée *K* analytique par les auteurs :

$$\widehat{K}_A(r) = \frac{A}{n(A)(n(A)-1)c(r)} \sum_{i=1}^{n(A)} \sum_{j=1, i\neq j}^{n(A)} \mathbf{1}\left(\|x_i - x_j\| \leq r\right) \quad \textbf{(12)}$$

(12) : Estimateur K_A de Ward et Ferrandino (1999)

Les notations sont celles de l'équation (7), page 18. Les auteurs ne justifient pas le facteur $(n(A)-1)$ présent au dénominateur à la place du facteur $n(A)$ de l'estimateur de Ripley de l'équation (7). Nous le ferons page 49.

Les auteurs assument sans la justifier la normalité de la distribution de l'estimateur. Ils calculent sa variance et donnent donc son intervalle de confiance. Au seuil de risque α, l'intervalle de confiance de $\widehat{K}_A(r)$ est :

$$CI\left(\widehat{K}_A(r)\right) = \pm t_\alpha \sqrt{\frac{A}{2\pi n(A)(n(A)-1)c(r)}\left(1 - \frac{A}{8\pi n(A)(\mathrm{n}(A)-1)r^2 c(r)}\right)} \quad \textbf{(13)}$$

L'intérêt de cette méthode de correction des effets de bord est double :

- Elle permet de calculer analytiquement l'intervalle de confiance de K ou L pour un processus de Poisson, et donc d'éviter les simulations de jeux de données nécessaires à la méthode de Monte-Carlo, très coûteuses en temps de calcul.
- La variance de la valeur de K pour l'hypothèse nulle est plus faible qu'avec la correction habituelle : on détecte plus facilement les écarts significatifs à l'hypothèse nulle, la puissance de K est augmentée.

Bizarrement, le travail de Ward et Ferrandino (1999) est resté confidentiel. Aucun article ultérieur n'a apparemment utilisé cette formulation de la fonction K et il n'est pas cité dans la bibliographie pourtant vaste de Goreaud (2000). Ses avantages sont pourtant considérables : le test d'écart à l'hypothèse nulle est plus puissant, le temps de calcul est très largement réduit, et les auteurs montrent par ailleurs que leur estimateur de L est meilleur (moins biaisé) que celui de Ripley. Une explication possible est que les formules explicites de correction des effets de bord ne sont données que pour un domaine d'étude rectangulaire et une distance inférieure à sa largeur, alors que les outils existants (Goreaud et Pélissier, 2000) permettent de calculer K dans un cadre bien plus général.

Le calcul de la variance de $\widehat{K}_A(r)$ est en réalité faux (Lang et Marcon, 2010). Il est réalisé en considérant que les paires de points sont indépendantes dans un processus de Poisson, ce qui n'est pas le cas : les points sont bien indépendants, mais le déplacement d'un point influe sur plusieurs paires. L'approximation est acceptable si l'intensité du processus de Poisson est faible. L'idée de Ward et Ferrandi-

no est reprise de façon plus rigoureuse plus loin (pages 31 et suivantes) : le calcul de la variance y est exact, et la normalité (asymptotique) démontrée.

Autres méthodes de correction

Les autres méthodes de correction sont plus anecdotiques. La plus simple de toutes consiste à faire en sorte de ne jamais devoir appliquer de correction en entourant le domaine d'étude d'une zone tampon dont la largeur est au moins égale à la valeur maximale de r. Cette zone tampon devant être cartographiée de la même façon que la zone d'étude, la tentation est grande de l'y inclure : cette méthode est rarement utilisée, seulement par Szwagrzyk et Czerwczak (1993) et Pancer-Koteja *et al.* (1998) à notre connaissance. Le deuxième article est une bonne illustration de la limite de la méthode : les placettes étudiées sont des cercles de 3m de rayon, dont 1,5 m de zone tampon ; la zone centrale ne constitue que 25% de la surface totale.

La correction toroïdale consiste à traiter le domaine d'étude comme un tore, c'est-à-dire à mettre en continuité ses bords opposés, à condition bien sûr que sa forme le permette. Une illustration en est donnée par Haase (1995 fig. 3, p. 578). Cette solution est intuitivement peu satisfaisante car elle considère comme très proche les points les plus éloignés.

Møller et Waagepetersen (2004, p. 37) proposent une correction variable pour chaque paire de points. Définissons $A_{x_j - x_i}$ la translation du domaine d'étude A par le vecteur $x_j - x_i$. La correction par $c(i,j) = \frac{1}{A \cap A_{x_j - x_i}}$ fournit un estimateur non biaisé de K. Elle n'a apparemment jamais été utilisée dans la littérature empirique.

La correction des effets de bord peut poser des problèmes pratiques importants. Getis et Franklin (1987) donnent les formules de correction pour une aire d'étude rectangulaire, Diggle (1983 p. 72), pour une aire d'étude rectangulaire ou circulaire. Haase (1995) réalise un comparatif des méthodes de correction de Ripley, de la zone tampon, et toroïdale. Il relève et corrige une erreur dans les formules de Diggle amenant à une sous-estimation assez importante de K et une petite erreur menant à une légère surestimation dans les formules de Getis et Franklin.

Goreaud et Pélissier (1999) ont développé des algorithmes pour l'étude de domaines plus complexes, implémentés dans le logiciel ADE (Thioulouse *et al.*, 1997).

Wiegand et Moloney (2004) utilisent une technique alternative qu'ils appellent approche numérique (Figure 4) : ils pixellisent la carte des points pour obtenir des données discrètes. Le nombre de points situé dans le cercle est divisé par le nombre de pixels du cercle situé dans la zone d'étude et par la surface unitaire des pixels. La correction est donc celle de Besag, mais le calcul éventuellement complexe de l'intersection du cercle et du domaine est inutile. En contrepartie, l'introduction de la grille dégrade légèrement la précision des données et surtout nécessite le choix arbitraire de sa maille.

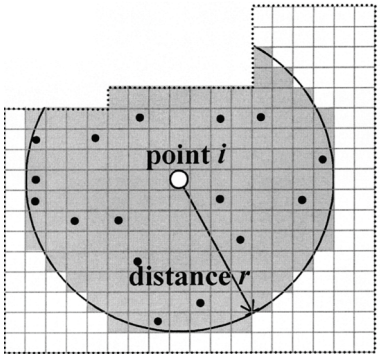


Figure 4 : Approche numérique de Wiegand et Moloney (2004).
La figure est extraite de l'article, Fig. 1, page 211

Le traitement des limites d'une zone géographique complexe comme les frontières d'un pays est donc possible, mais n'a jamais été appliqué dans la littérature : la zone étudiée est toujours un polygone (Rowlingson et Diggle, 1993 fig. 5, p. 634 ; Sweeney et Feser, 1998 fig. 1, p. 52 ; Marcon et Puech, 2002 p. 12).

Exemples

Dans toutes les applications empiriques, on calculera les estimateurs $\widehat{K}(r)$ et $\hat{L}(r)$ à partir des données.

Distribution agrégée

Un processus de Thomas est utilisé pour générer, dans un domaine d'étude carré de 10x10, 10 agrégats contenant 100 points. Le paramètre de dispersion est $\sigma = 0{,}2$. La carte des points est présentée sur la Figure 5.

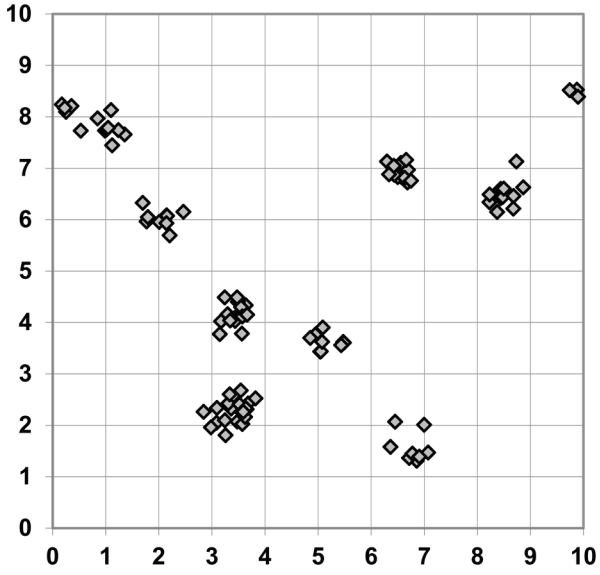


Figure 5 : Semis agrégé, Carte des points

L'estimation de L est donnée sur la Figure 6. L'intervalle de confiance de l'hypothèse nulle est calculé au risque de 1%, à partir de 10 000 simulations.

La courbe $\hat{L}$ présente un pic caractéristique pour une valeur de r correspondant approximativement au rayon des agrégats (Goreaud, 2000), c'est-à-dire à 0,5. Le pic secondaire qu'on observe à une distance de 2,5 correspond à l'espacement moyen entre les agrégats, qui constituent en quelque sorte des agrégats d'agrégats, visibles sur la carte.

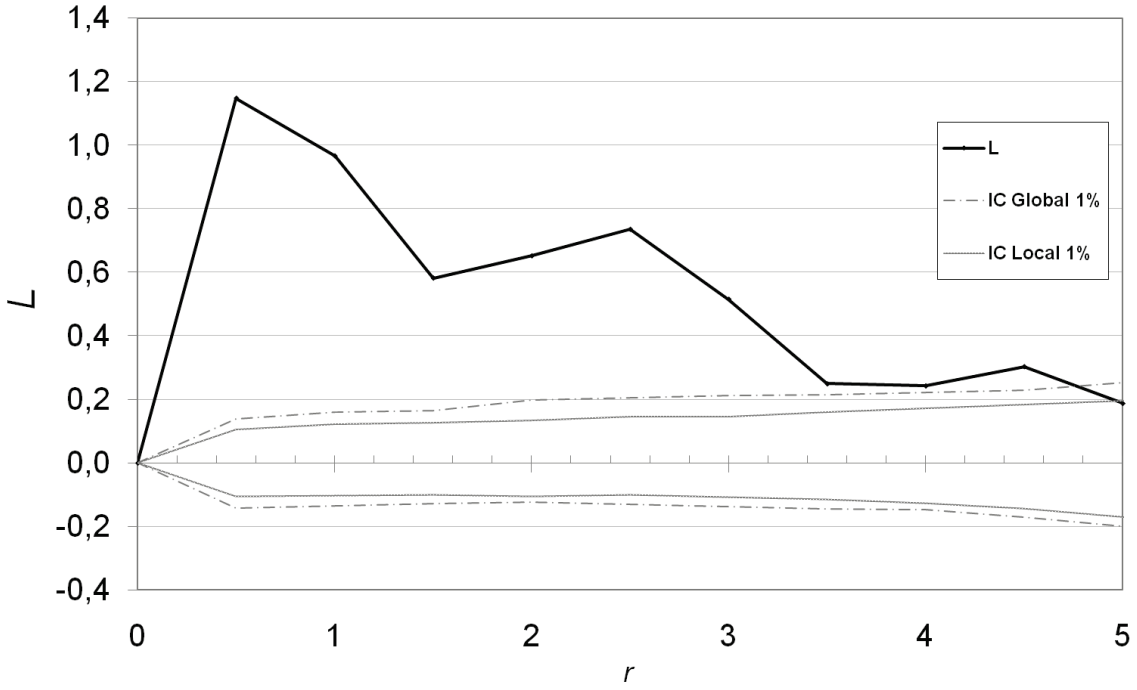


Figure 6 : Fonction *L* pour un semis de point agrégé

On peut faire quelques observations sur la fonction $\hat{L}(r)$:

- Sa valeur est homogène à une distance. $\hat{L}(0{,}5)$ est à peu près égal à 1,2, ce qui signifie qu'on trouve en moyenne dans un cercle de rayon 0,5 autour de chaque point autant de voisins qu'on en compterait dans un cercle de rayon 1,2+0,5=1,7 si la distribution était homogène.
- La fonction L intégrant le nombre de voisins de 0 à r, elle reste hors de l'intervalle de confiance de l'hypothèse nulle bien au delà de la taille des agrégats. Sa décroissance maximum est facilement calculable : si on ne rencontre plus de voisins à partir d'un certain rayon, $\mathrm{d}\hat{L}(r)/\mathrm{d}r = -r$, la pente de la courbe est donc égale à -1.

Pour les valeurs supérieures de r, la courbe L rentre dans l'intervalle de confiance : la correction des effets de bord devient de plus en plus importante et fait tendre la fonction L vers sa valeur de référence, 0.

Distribution régulière

La distribution régulière est constituée de 100 points disposés sur une maille carré de 1x1, dans le même domaine d'étude (carte en Figure 7).

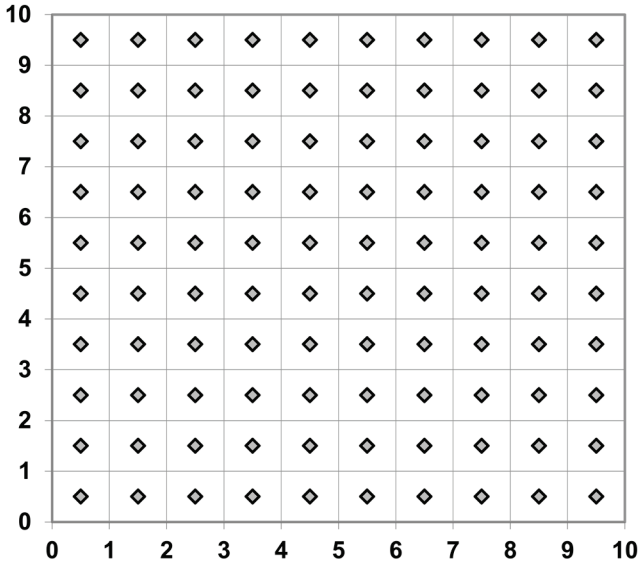


Figure 7 : Semis régulier, Carte des points

La courbe L est en Figure 8. L'intervalle de confiance est calculé au seuil de 1% sur 10 000 simulations.

La courbe présente des pics négatifs au niveau du dernier pas avant la taille de la maille (ici, 0,75 car la courbe est calculée avec un pas de 0,25. Si le pas était de 0,01, le pic négatif se trouverait à 0,99). Le pic correspond à la distance pour laquelle le nombre de voisins est minimum. A la taille de la maille, la courbe remonte brutalement (en toute rigueur, elle est discontinue), et peut même être significativement positive quand la densité de points est faible comme dans l'exemple.

Ce comportement se répète pour les distances correspondant à la diagonale de la maille ($\sqrt{2} \approx 1{,}44$), puis deux fois la maille (2), puis la diagonale du rectangle de longueur 2 et de largeur 1 ($\sqrt{5} \approx 2{,}24$).

La détection de la régularité est très sensible au pas du calcul ε :

- Avant $r = 1$, $L(r) = -r$. Le premier pic négatif correspond à $L(1-\varepsilon) = 1 + \varepsilon$. Le pas de 0,25 fait apparaître un pic pour $L(0{,}75) = -0{,}75$ et sous-estime la dispersion.
- un pic négatif existe entre les valeurs 2 et 2,24, mais il n'apparaît pas dans la courbe à cause du pas de 0,25.

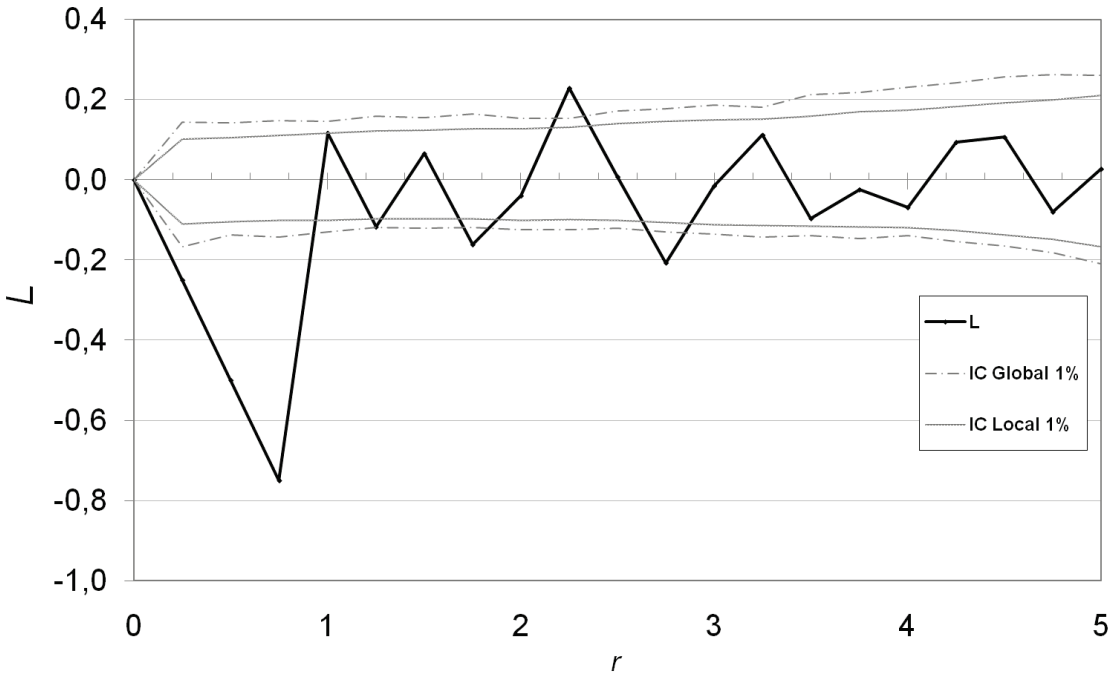


Figure 8 : Fonction *L* pour un semis de point régulier

La courbe *L* recalculée avec un pas de 0,1 est en Figure 9.

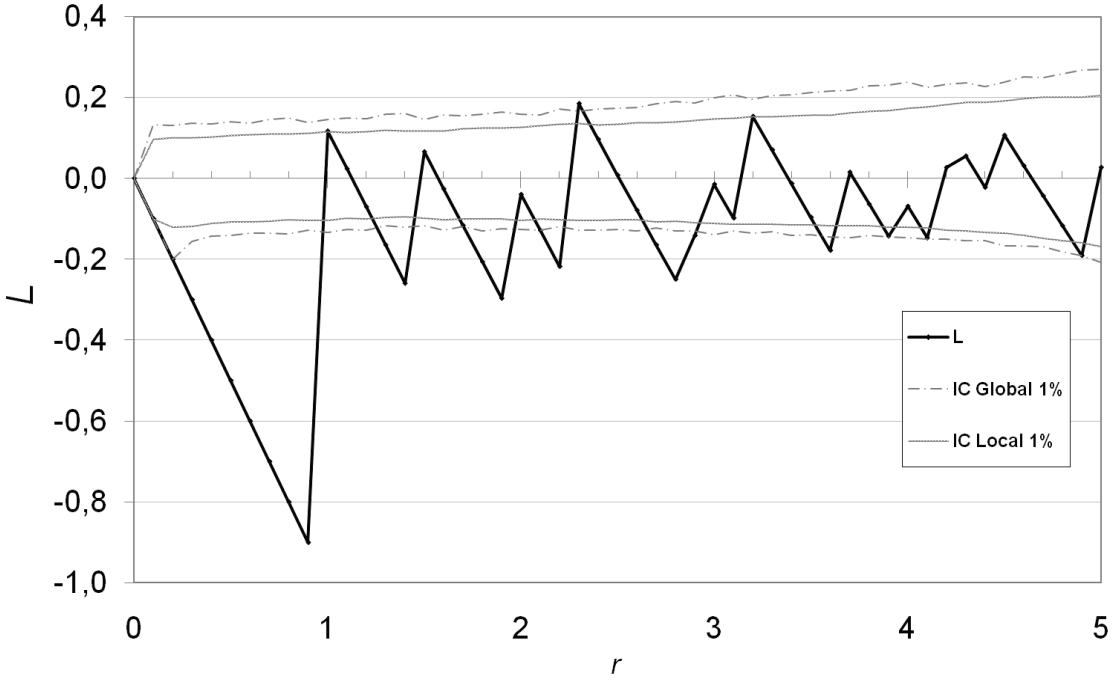


Figure 9 : Fonction *L* affinée pour un semis de point régulier

Un pas de calcul trop grand peut occulter des structures régulières. Les structures agrégées sont beaucoup plus faciles à détecter puisque la pente décroissante de la courbe *L* ne peut pas dépasser –1.

Intervalle de confiance asymptotique de la fonction K de Ripley

Le calcul des intervalles de confiance (pour chaque valeur de r) par la méthode de Monte-Carlo présente deux inconvénients :

- Sur le plan théorique, la méthode utilisée par Duranton et Overman pour exclure les courbes n'a pas de fondement robuste. Elle est sensible notamment au nombre de valeurs de r utilisées : en diminuant le pas de calcul de K, on augmente la probabilité d'éliminer des courbes non détectées avec un pas plus grand. Ce problème est assez limité en pratique parce que les courbes se croisent peu.
- Sur le plan pratique, le temps de calcul étant à peu près proportionnel au carré du nombre de points (pour le calcul des distances entre les paires de points) multiplié par le nombre de simulations, la méthode devient inutilisable au-delà de quelques milliers de points.

Il est possible de calculer analytiquement l'intervalle de confiance de K, au moins dans des cas simples. La stratégie présentée ici consiste à calculer la matrice de variance-covariance de K pour un processus de Poisson homogène, et à l'utiliser pour établir un intervalle de confiance des valeurs de K. La méthode nécessite de calculer l'espérance des effets de bord, ce qui la limite à des formes de domaines relativement simples. Les calculs se trouvent dans Lang et Marcon (2010).

Cadre et notations

Le domaine carré a un côté de longueur n et est noté A_n. Les résultats asymptotiques seront obtenus en faisant tendre n vers l'infini.

Deux cas doivent être pris en compte :

- l'intensité λ du processus peut être connue. L'estimateur de K est noté $\widehat{K}_{1,A_n}(r)$.
- en situation réelle, λ est inconnu et doit être estimé à partir du nombre de points observés $n(A_n)$. Les estimateurs retenus sont $\hat{\lambda} = n(A_n)/n^2$ et $\widehat{\lambda^2} = \sqrt{n(A_n)(n(A_n)-1)}/n^2$: l'estimateur de λ^2 sera justifié page 50. L'estimateur de K est noté $\widehat{K}_{2,A_n}(r)$.

Les valeurs de r forment le vecteur $(r_1, \dots, r_i, \dots, r_d)$ de dimension d. Lorsque deux valeurs de r différentes sont utilisées pour le calcul des covariances, elles sont notées r et r' pour alléger les écritures.

Les estimateurs $\widehat{K}_{1,A_n}(r)$ et $\widehat{K}_{2,A_n}(r)$ sont calculés sans correction des effets de bord. Enfin, la notation $\widehat{K}_{A_n}(r)$ signifie que les résultats sont valables pour les deux estimateurs.

Biais

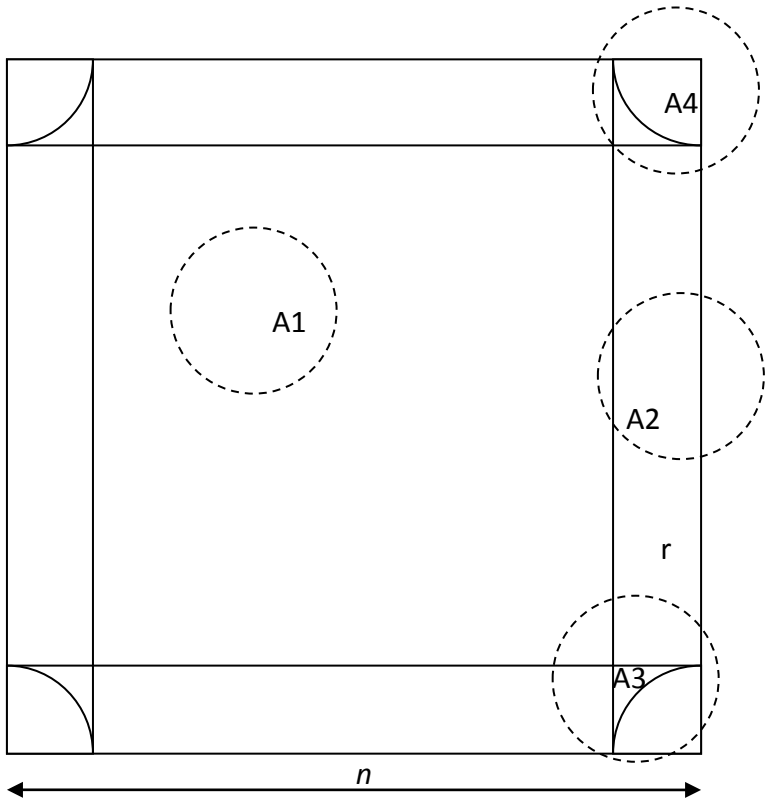


Figure 10 : Zonage du carré pour le calcul des effets de bord

L'objectif est de calculer l'espérance du biais de $\widehat{K}_{A_n}(r)$ dû aux effets de bord. Pour un processus de Poisson homogène, il suffit de calculer l'espérance de la surface des cercles de rayon r située hors du domaine d'étude. La Figure 10 montre les 4 cas à traiter, selon la position du centre du cercle :

- Dans la zone centrale du carré (A1), tout le cercle se trouve dans le carré, il n'y a pas de biais.
- Sur les bords du carré (A2), quand le centre du cercle se trouve à moins de r du bord, une partie du cercle sort du carré,
- Dans la zone A3, à moins de r de deux bords mais à plus de r du coin du carré, le cercle sort deux fois, mais le coin du carré se trouve à l'extérieur du cercle,
- Dans la zone A4, à moins de r du coin du carré, celui-ci se trouve dans le cercle.

Pour les cas A2 à A4, la surface du cercle se trouvant à l'extérieur du carré est calculée pour une position quelconque du centre du cercle. Cette valeur doit ensuite être intégrée sur la surface de A2, A3 ou A4 pour calculer l'espérance de la valeur de l'intersection. Les calculs sont assez lourds pour les zones A3 et A4. Les calculs pourront être étendus à des domaines rectangulaires ou même polygonaux. Pour des domaines de forme complexe, les intégrales seront souvent incalculables.

Le biais est au total :

$$\begin{aligned}\mathbb{E}\widehat{K}_{1,A_n}(r) - K(r) &= r^2\left(-\frac{8r}{n^3}+\frac{r^2}{2n^2}\right)\\ \mathbb{E}\widehat{K}_{2,A_n}(r) - K(r) &= r^2\left(-\frac{8r}{n^3}+\frac{r^2}{2n^2}\right) - r^2e^{-\lambda n^2}\left(\pi-\frac{8r}{3n}+\frac{r^2}{2n^2}\right)\left(1+\lambda n^2e^{-\lambda n^2}\right)\end{aligned} \quad \textbf{(14)}$$

Le terme supplémentaire du biais quand λ est inconnu est très petit, toujours négligeable en pratique. Quand λ est inconnu, il est estimé par $n(A_n)/n^2$.

Variance et covariance

La variance de $\widehat{K}_{A_n}(r)$ peut être calculée de façon exacte quelle que soit la valeur de n. La forme analytique de la covariance entre $\widehat{K}_{A_n}(r)$ et $\widehat{K}_{A_n}(r')$ contient une intégrale double qui ne peut pas être résolue mais peut être calculée numériquement sans difficulté.

Variance

Intensité connue :

$$\mathrm{Var}\left(\widehat{K}_{1,A_n}(r)\right) = \frac{2e_{r,n}}{\lambda^2}+\frac{4n^2e_{r,n}^2}{\lambda}+\frac{4n^2}{\lambda}\mathbb{E}\left(h_1^2(U,r)\right) \quad \textbf{(15)}$$

Intensité inconnue :

$$\begin{aligned}\mathrm{Var}\left(\widehat{K}_{2,A_n}(r)\right) &= 2n^4\mathbb{E}\left(\frac{\mathbf{1}(N(A_n)>1)}{N(A_n)(N(A_n)-1)}\right)\left(e_{r,n}-e_{r,n}^2\right)\\ &+ 4n^4\mathbb{E}\left(\frac{\mathbf{1}(N(A_n)>1)(N(A_n)-2)}{N(A_n)(N(A_n)-1)}\right)\mathbb{E}\left(h_1^2(U,r)\right)\\ &+ n^4e^{-\lambda n^2}(1+\lambda n^2)\left(1-e^{-\lambda n^2}-\lambda n^2e^{-\lambda n^2}\right)e_{r,n}^2\end{aligned} \quad \textbf{(16)}$$

Où $e_{r,n}$ est l'espérance du rapport entre la surface du cercle de rayon r incluse dans le carré de côté n et celle du carré, c'est-à-dire la probabilité qu'un point soit voisin d'un autre, les effets de bord étant pris en compte :

$$e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8r^3}{3n^3} + \frac{r^4}{2n^4} \quad \textbf{(17)}$$

La fonction $h_1(U,r)$ est un intermédiaire de calcul tel que :

$$\mathbb{E}\left(h_1^2(U,r)\right) = \frac{r^5}{n^5}\left(\frac{8}{3}\pi - \frac{256}{45}\right) + \frac{r^6}{n^6}\left(\frac{11}{48}\pi - \frac{56}{9}\right) + \frac{8r^7}{3n^7} - \frac{r^8}{4n^8} \quad \textbf{(18)}$$

$\mathbb{E}\left(\frac{\mathbf{1}(N(A_n)>1)}{N(A_n)(N(A_n)-1)}\right)$ et $\mathbb{E}\left(\frac{\mathbf{1}(N(A_n)>1)(N(A_n)-2)}{N(A_n)(N(A_n)-1)}\right)$ sont estimés par $\frac{1}{n(A_n)(n(A_n)-1)}$ et $\frac{(n(A_n)-2)}{n(A_n)(n(A_n)-1)}$ parce que $N(A_n)$ suit une loi de Poisson de paramètre élevé donc peu dispersée.

Covariances

Le calcul des covariances est nettement plus laborieux.

Intensité connue :

$$\mathrm{cov}\left(\widehat{K}_{1,A_n}(r), \widehat{K}_{1,A_n}(r')\right) = \frac{2e_{r,n}}{\lambda^2} + \frac{4n^2 e_{r,n} e_{r',n}}{\lambda} + \frac{4n^2}{\lambda}\mathrm{cov}\left(h_1(U,r)h_1(U,r')\right) \quad \textbf{(19)}$$

Intensité inconnue :

$$\begin{aligned}\mathrm{cov}&\left(\widehat{K}_{2,A_n}(r), \widehat{K}_{2,A_n}(r')\right) \\ &= 2n^4\mathbb{E}\left(\frac{\mathbf{1}(N(A_n)>1)}{N(A_n)(N(A_n)-1)}\right)\left(e_{r,n} - e_{r,n}e_{r',n}\right) \\ &+ 4n^4\mathbb{E}\left(\frac{\mathbf{1}(N(A_n)>1)(N(A_n)-2)}{N(A_n)(N(A_n)-1)}\right)\mathrm{cov}\left(h_1(U,r)h_1(U,r')\right) \\ &+ n^4 e^{-\lambda n^2}(1+\lambda n^2)\left(1 - e^{-\lambda n^2} - \lambda n^2 e^{-\lambda n^2}\right)e_{r,n}e_{r',n}\end{aligned} \quad \textbf{(20)}$$

Où :

$$
\begin{aligned}
&\operatorname{cov}\big(h_1(U,r)h_1(U,r')\big) \\
&\quad = \left(1-\frac{2r'}{n}\right)\frac{r^2r'^2}{n^4}b_{r,n}b_{r',n} \\
&\quad + 4\left(1-\frac{2r'}{n}\right)\frac{r^2r'^3}{n^5}b_{r,n}\int_{r/r'}^{1}\Big(b_{r',n}-g(x'_1)\Big)\,dx'_1 \\
&\quad + 4\left(1-\frac{2r'}{n}\right)\frac{r^3r'^2}{n^5}b_{r,n}\int_{0}^{1}\left(b_{r',n}-g\left(\frac{rx_1}{r'}\right)\right)\Big(b_{r,n}-g(x_1)\Big)\,dx_1 \\
&\quad + 4\frac{r^2r'^4}{n^6}\int_0^1\int_0^1\left[h_{A1}\left(\frac{r'x'}{r},r\right)+h_{A2}\left(\frac{r'x'}{r},r\right)+h_{A3}\left(\frac{r'x'}{r},r\right)\right. \\
&\quad \left. + h_{A4}\left(\frac{r'x'}{r},r\right)\Big(h_{A3}(x',r')+h_{A4}(x',r)\Big)\right]dx'_1dx'_2
\end{aligned}
\tag{21}
$$

Et :

$$
b_{r,n}=\pi-\frac{n^2}{r^2}e_{r,n}=\frac{8r}{3n}-\frac{r^2}{2n^2} \tag{22}
$$

Et

$$
g(x)=\mathbf{1}(x<1)\left(\cos^{-1}x+x\sqrt{1-x^2}\right) \tag{23}
$$

Et :

$$
\begin{aligned}
h_{A1}(x,r) &= b_{r,n}\mathbf{1}(x_1\geq 1)\mathbf{1}(x_2\geq 1) \\
h_{A2}(x,r) &= \Big(b_{r,n}-g(x_2)\Big)\mathbf{1}(x_1\geq 1)\mathbf{1}(x_2<1) \\
&\quad + \Big(b_{r,n}-g(x_1)\Big)\mathbf{1}(x_2\geq 1)\mathbf{1}(x_1<1) \\
h_{A3}(x,r) &= \Big(b_{r,n}-g(x_1)-g(x_2)\Big)\mathbf{1}(x_1<1)\mathbf{1}(x_2<1)\mathbf{1}(x_1^2+x_2^2\geq 1) \\
h_{A4}(x,r) &= \left(b_{r,n}-\frac{\pi}{4}+x_1x_2-\frac{g(x_1)+g(x_2)}{2}\right)\mathbf{1}(x_1^2+x_2^2\leq 1)
\end{aligned}
\tag{24}
$$

Les intégrales sont calculées numériquement car les fonctions intégrands sont très régulières.

Le détail des calculs se trouve dans Lang et Marcon (2010).

Théorème central limite

Un vecteur de $\widehat{K}_{A_n}(r)$ pour différentes valeurs de r (notées r_i, $i\in\{1,2,\dots,d\}$ tend vers un vecteur gaussien quand $n\to\infty$:

$$\widetilde{K}_{A_n} = n\sqrt{\rho}\left[\begin{pmatrix}\widehat{K}_{A_n}(r_1)\\ \dots \\ \widehat{K}_{A_n}(r_d)\end{pmatrix} - \pi\begin{pmatrix}{r_1}^2\\ \dots \\ {r_d}^2\end{pmatrix}\right] \to \mathcal{N}(0,\Sigma) \quad \textbf{(25)}$$

Σ est la matrice de variance covariance d'une loi normale de dimension d. Ses éléments valent :

- Si λ est connu :

$$\Sigma_{s,t} = \frac{2\pi}{\lambda}\min({r_s}^2, {r_t}^2) + 4\pi {r_s}^2 {r_t}^2 \quad \textbf{(26)}$$

- Si λ est inconnu :

$$\Sigma_{s,t} = \frac{2\pi}{\lambda}\min({r_s}^2, {r_t}^2) \quad \textbf{(27)}$$

Dans ce dernier cas, λ est estimé par $\hat{\lambda} = \frac{n(A_n)}{n^2}$.

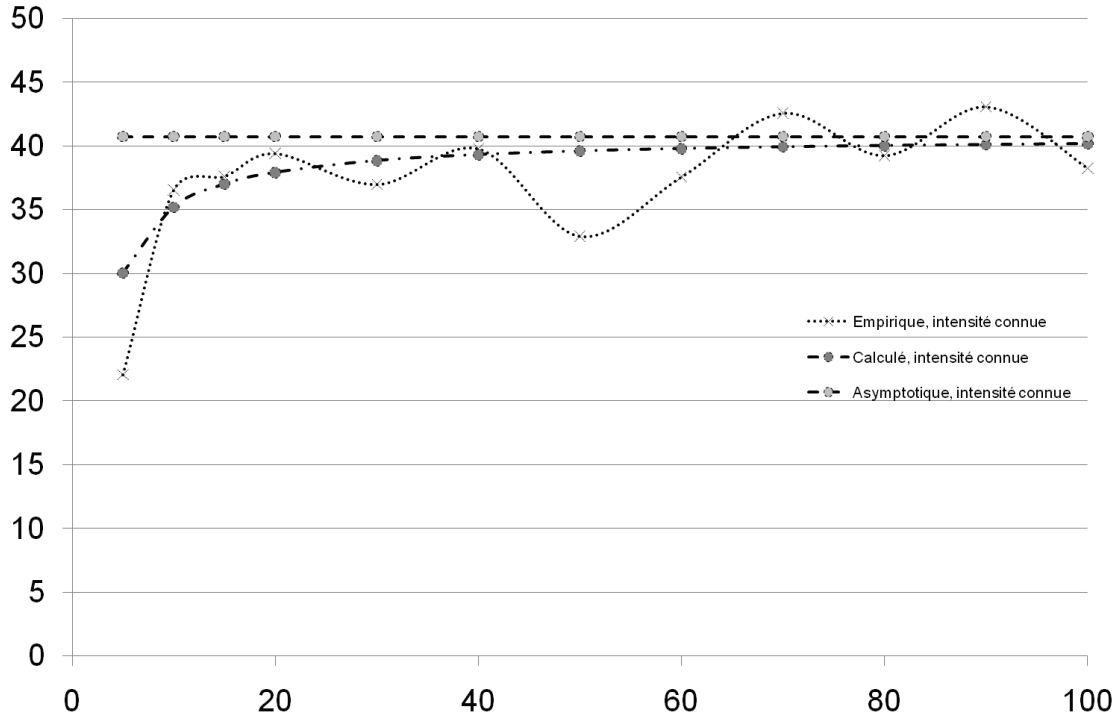


Figure 11 : Convergence de la variance de $\widehat{K}_{1,A_n}(1)$, intensité connue.

Les tests montrent que les valeurs limites sont atteintes pour des nombres de points très supérieurs à ceux utilisés en pratique. Les variances asymptotiques n'ont donc pas d'intérêt pratique.

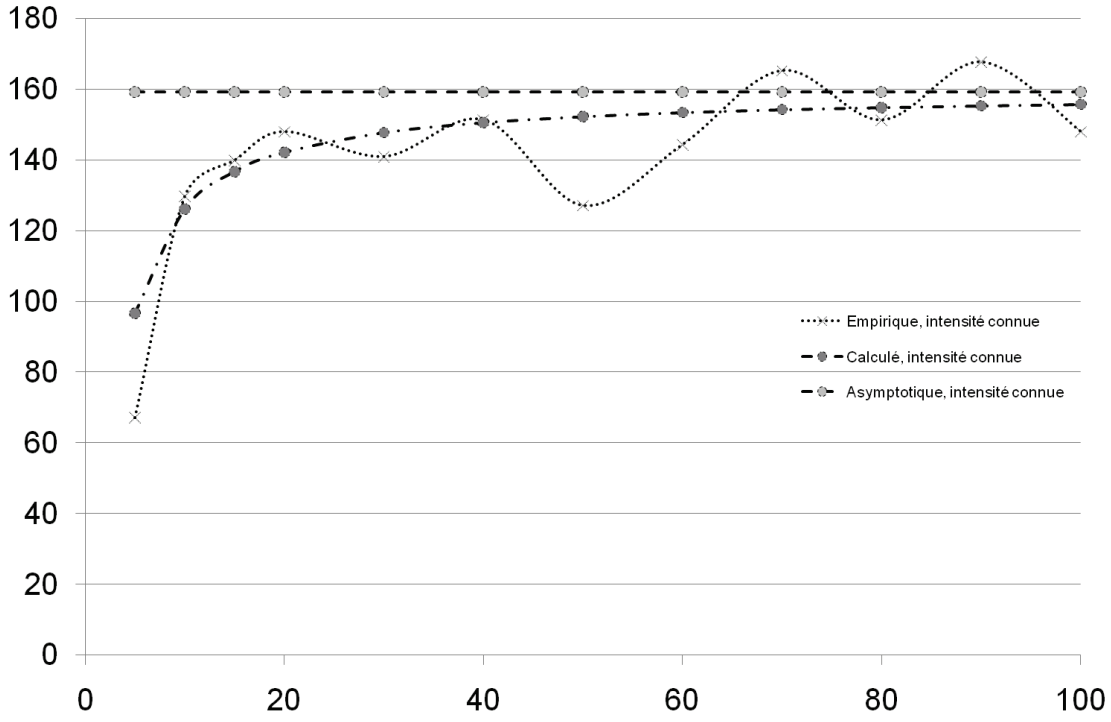


Figure 12 : Convergence de la covariance de $\widehat{K}_{1,A_n}(1)$ et $\widehat{K}_{1,A_n}(2)$, intensité connue.

La Figure 11 et la Figure 13 montrent les valeurs de la variance de $\widehat{K}_{A_n}(1)$ pour des valeurs croissantes de n. Les courbes empiriques sont calculées à partir de 1 000 simulations d'un processus de Poisson d'intensité 5 tiré dans un carré de côté n. L'espérance du nombre de points varie de 125 à 50 000. Les variances décroissent à la vitesse λn^2 : elles sont donc toutes multipliées par λn^2 pour permettre de les comparer. Quand λ est inconnu, λn^2 est estimé par $\widehat{\lambda n^2} = n(A_n)$.

Les estimateurs $\widehat{K}_{1,A_n}(1)$ et $\widehat{K}_{2,A_n}(1)$ sont calculés pour chaque tirage. Leur variance empirique varie beaucoup, malgré le nombre de simulations important. Quand l'intensité est connue, la variance (normalisée) augmente avec le nombre de points, mais le comportement est inversé quand l'intensité est inconnue, parce que l'incertitude due à son estimation est plus faible.

Quand l'intensité est connue, la variance de $\widehat{K}_{1,A_n}(1)$ est calculée indépendamment des données. Quand elle ne l'est pas, elle doit être estimée à partir des données. À chaque tirage du processus correspond donc une valeur de la variance calculée de $\widehat{K}_{2,A_n}(1)$, qui varie selon le nombre de points observé. La Figure 13 et la Figure 14 présentent pour chaque valeur de n la plus petite et la plus grande de ces valeurs sur les 1 000 tirages.

Même pour un très grand nombre de points ($n = 100$, $\lambda = 5$, 50 000 points attendus), la variance (empirique ou calculée) est encore loin de la variance asymptotique.

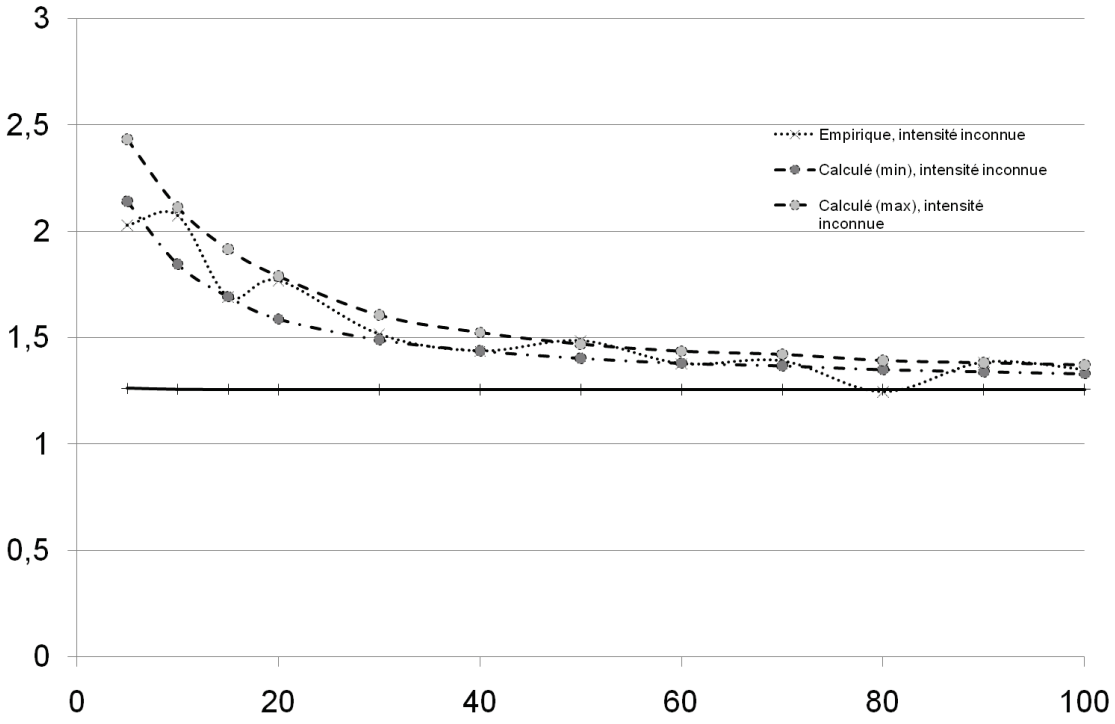


Figure 13 : Convergence de la variance de $\widehat{K}_{2,A_n}(1)$, intensité inconnue.

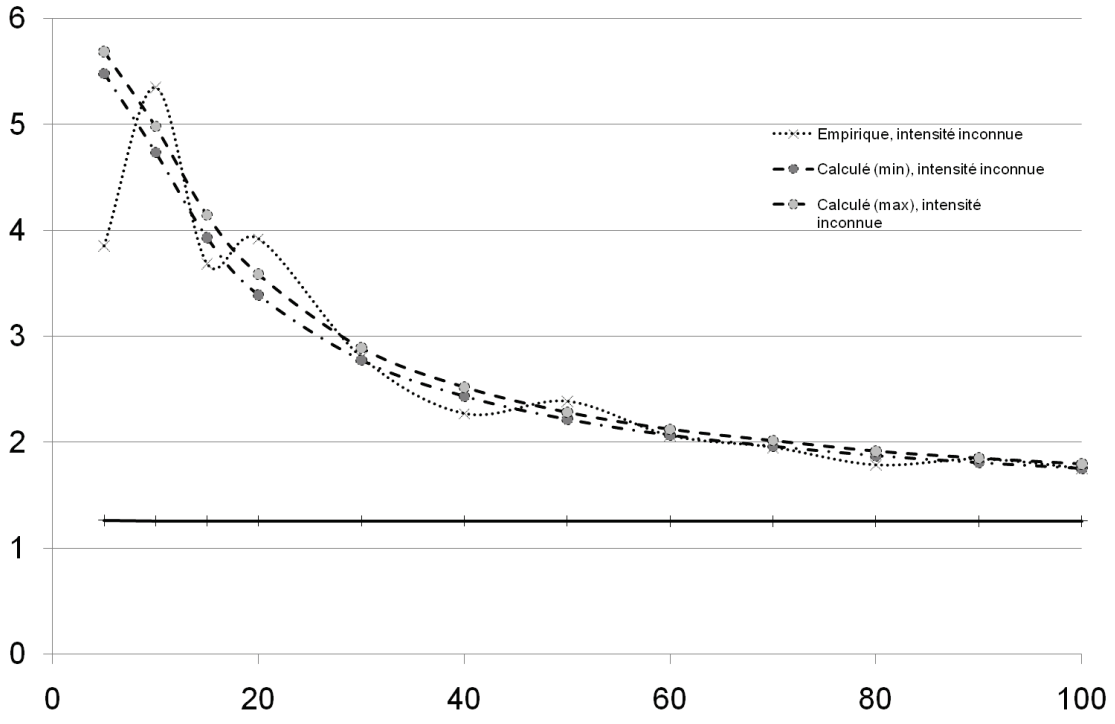


Figure 14 : Convergence de la covariance de $\widehat{K}_{2,A_n}(1)$ et $\widehat{K}_{2,A_n}(2)$, intensité inconnue.

La même approche est appliquée à la covariance entre $\widehat{K}_{A_n}(1)$ et $\widehat{K}_{A_n}(2)$. Les résultats (Figure 12 et Figure 14) sont sensiblement les mêmes.

En conclusion, il est illusoire de chercher à utiliser la variance asymptotique dans les cas pratiques. La matrice de variance-covariance devra être calculée pour les valeurs réelles (éventuellement estimées) de n et λ.

Test

$\mathrm{T} = \Sigma^{-1/2} \begin{pmatrix} \widehat{K}_{A_n}(r_1) \\ \dots \\ \widehat{K}_{A_n}(r_d) \end{pmatrix}$ est distribué asymptotiquement comme une loi normale centrée réduite de dimension d. Le test statistique consiste donc à calculer la norme de T au carré, qui suit une loi de χ^2_α à d degrés de libertés. Si $\|T\|^2$ dépasse le quantile de $\chi^2_\alpha(d)$ au seuil de risque α, l'hypothèse selon laquelle le semis de points est une réalisation d'un processus de Poisson homogène peut être rejetée.

Pour vérifier la méthode, un processus de Poisson pris comme hypothèse nulle est tiré 10 000 fois. Au seuil de risque de 5%, 500 tirages doivent être rejetés comme n'étant pas des Poisson. Le test peut être lui-même testé : le rejet d'un tirage suit une loi de Bernoulli, et les tirages sont indépendants. L'espérance de la proportion de rejets est 0,05 et sa variance théorique 0,0475. Comme le test est répété 10 000 fois, l'approximation normale de la distribution des résultats est valide. Au seuil de risque de 5%, la proportion de rejets devrait être $5\% \pm t_{5\%}(10000)\sqrt{\frac{0{,}0475}{10000}}$, c'est-à-dire entre 4,56% et 5,44%.

Les calculs ont été réalisés dans R (R Development Core Team, 2010) : la simulation des processus avec la fonction `rpoispp` du module Spatstat, l'intégration numérique simple avec `integrate` et l'intégration double avec la fonction `adapt` du module du même nom.

Différentes méthodes de calcul de la matrice Σ sont testées :

- Empirique : 10 000 tirages supplémentaires du même processus de Poisson sont effectués, $\widehat{K}_{1,A_n}$ est calculé pour toutes les valeurs de r et la matrice de variance-covariance est calculée à partir de ces valeurs.
- Empirique (auto) : la matrice de variance-covariance est calculée à partir des tirages testés eux-mêmes.
- Intensité connue ou inconnue : Σ est calculée. Si l'intensité est connue, Σ ne dépend pas des données : seules les valeurs de $\widehat{K}_{1,A_n}$ sont calculées pour chaque tirage. Si l'intensité est inconnue, Σ est aussi calculée à chaque tirage en fonction du nombre de points observés.

Méthode de calcul de Σ	*Empirique*	*Empirique (auto, intensité connue)*	*Empirique (auto, intensité inconnue)*	*Intensité connue, calcul exact*	*Intensité inconnue, calcul exact*
Poisson, $n = 30$, $\lambda = 1$	*5,40*	*5,04*	*5,01*	*5,20*	*5,10*
Poisson, $n = 10$, $\lambda = 5$	5,61	5,40	5.38	5,19	5,37
Poisson, $n = 10$, $\lambda = 5$ $r \in \{1; 2; \ldots; 10\}$	5,28	5,32	6,67	6,08	5,84
Poisson, $n = 10$, $\lambda = 1$	5,67	5,86	5,30	5,81	5,25
Poisson, $n = 10$, $\lambda = 0,5$	5,52	5,73	5,60	5,52	4,91
Poisson, $n = 10$, $\lambda = 0,2$	6,40	6,84	6,59	6,59	5,22
Thomas ($\kappa = 1$, $\sigma = 3$, $\mu = 5$)					71,63
Thomas ($\kappa = 0,5$, $\sigma = 0,5$, $\mu = 10$)					100
Strauss ($\beta = 10$, $\gamma = 0,95$, $r = 1$)					21,31

Tableau 1 : Pourcentage de rejets de tirages de processus divers (1 000 tirages). Les distances traitées sont $r \in \{1; 2; 5\}$ sauf indication contraire.

Le Tableau 1 présente les résultats en pourcentages de tirages rejetés. Chaque ligne du tableau correspond à un processus différent :

- Un processus de Poisson d'intensité 1 tiré dans un carré de 30x30 (espérance du nombre de points : 900), $\widehat{K}_{A_{30}}$ est calculé pour $r \in \{1; 2; 5\}$. Ce cas ne présente aucune difficulté : peu d'effets de bord, beaucoup de points.

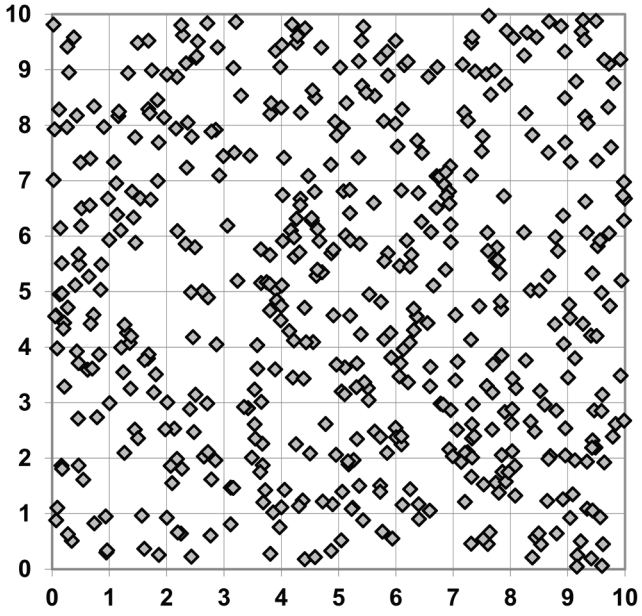


Figure 15 : Processus de Poisson

- Un processus de Poisson d'intensité 5 tiré dans un carré de 10x10 (espérance du nombre de points : 500, voir un tirage Figure 15), $\widehat{K}_{A_{10}}$ est calculé pour $r \in \{1; 2; 5\}$. Cet exemple correspond à un cas concret assez courant. Les effets de bord sont importants : pour $r = 5$, la zone centrale du carré de la Figure 10 disparaît totalement.
- Un processus de Poisson d'intensité 1 tiré dans un carré de 30x30 (espérance du nombre de points : 900), $\widehat{K}_{A_{10}}$ est calculé pour $r \in \{1; 2; 3; \dots; 10\}$. Ce cas pose le problème de l'inversion d'une matrice Σ de grande taille (10x10).
- Trois processus de Poisson d'intensité 1, 0,5 puis 0,2 tirés dans un carré de 10x10. $\widehat{K}_{A_{10}}$ est calculé pour $r \in \{1; 2; 5\}$. L'objectif est de tester les performances de la méthode pour un nombre de points aussi faible que possible (espérance de 100, 50 puis 20 points).

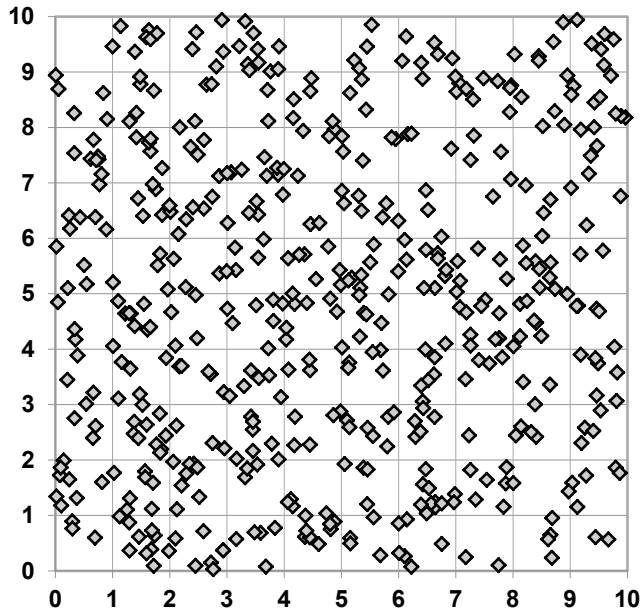


Figure 16 : Processus de Thomas
Voir la définition en annexe 1, page 144

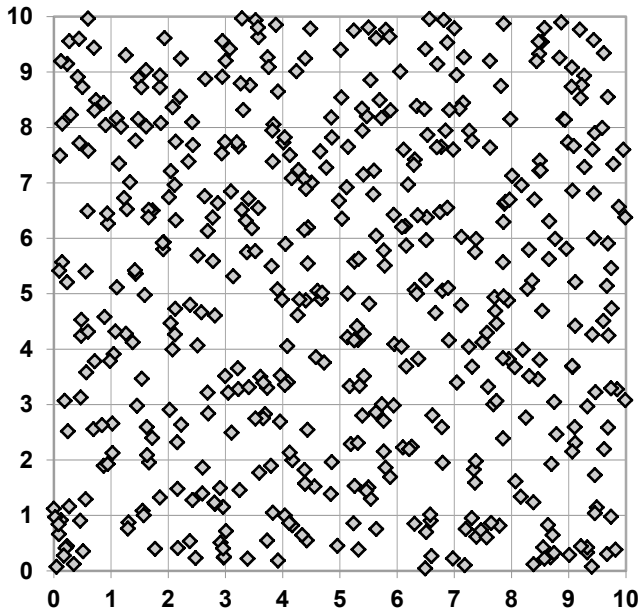


Figure 17 : Processus de Strauss
Voir la définition en annexe 1, page 147

- Un processus de Thomas très légèrement agrégé (voir un tirage Figure 16), dans un carré de 10x10. L'intensité des centres $\kappa = 1$, l'écart type de la distance entre les points et les centres $\sigma = 3$ et l'espérance du nombre de points par agrégat $\mu = 5$ (espérance du nombre total de points : 500) font que l'agrégation est difficile à détecter parce que les agrégats de points sont grands et se superposent largement. Le même processus avec deux fois moins d'agrégats, plus petits ($\sigma = 0{,}5$) et contenant deux fois plus de points, devrait être plus facile à détecter.
- Un processus de Strauss (Figure 17), très légèrement répulsif, de paramètres $\beta = 10$, $\gamma = 0{,}95$ et $r = 1$ dans un carré de 10x10. Ce processus est

indiscernable du Poisson et du Thomas des figures précédentes. L'objectif est de tester la puissance du test.

Les résultats sont proches de 5% pour les processus de Poisson, ce qui montre que le test asymptotique fonctionne sur des cas réels. Bien que la variance asymptotique soit loin d'être atteinte, la normalité est approximativement atteinte rapidement, ce qui permet d'appliquer le test à condition de calculer la variance précisément. Les processus agrégés sont bien détectés (71% de rejet pour le Thomas très lâche et 100% pour le cas plus facile). Le processus très légèrement répulsif est détecté dans 21% des cas.

L'estimation empirique de la variance à partir de simulations présente l'avantage d'être très simple mais en pratique, le calcul de Σ nécessite de connaître l'intensité du processus, ce qui limite son intérêt.

L'utilisation de la variance calculée est donc plus appropriée. En pratique, l'intensité du processus est très rarement connue. Elle doit être estimée à partir du seul tirage disponible, le jeu de points étudié.

Interprétation graphique

La Figure 18 montre la corrélation entre les valeurs de $\widehat{K}_{A_n}(r) - \pi r^2$ pour deux valeurs de r différentes. Chaque point représente un tirage d'un processus de Poisson. On peut imaginer le même graphique en un nombre de dimensions correspondant au nombre de valeurs de r.

Comme la fonction $\widehat{K}_{A_n}$ est cumulative, ses valeurs sont très autocorrélées. La figure présente les résultats de tirages de processus de Poisson. Certains tirages sont légèrement agrégés (valeurs positives de $\widehat{K}_{A_n}$), d'autres légèrement dispersés (valeurs négatives).

La multiplication par la matrice de variance à la puissance $-1/2$ rend les valeurs indépendantes, centrées, de variance 1. Les valeurs de T pour les mêmes données sont présentées dans la Figure 19.

Asymptotiquement, les valeurs sont normales, ce qui permet de pratiquer le test. En pratique, la normalité est atteinte assez rapidement pour que le test fonctionne avec des effectifs de points assez faibles, à condition de calculer les variances et covariances pour le nombre de points observé : la variance asymptotique n'est jamais atteinte.

Les points les plus éloignés de l'origine du repère sont rejetés. Le cercle a un rayon égal à la racine de la valeur critique d'un $\chi^2_{5\%}(2)$.

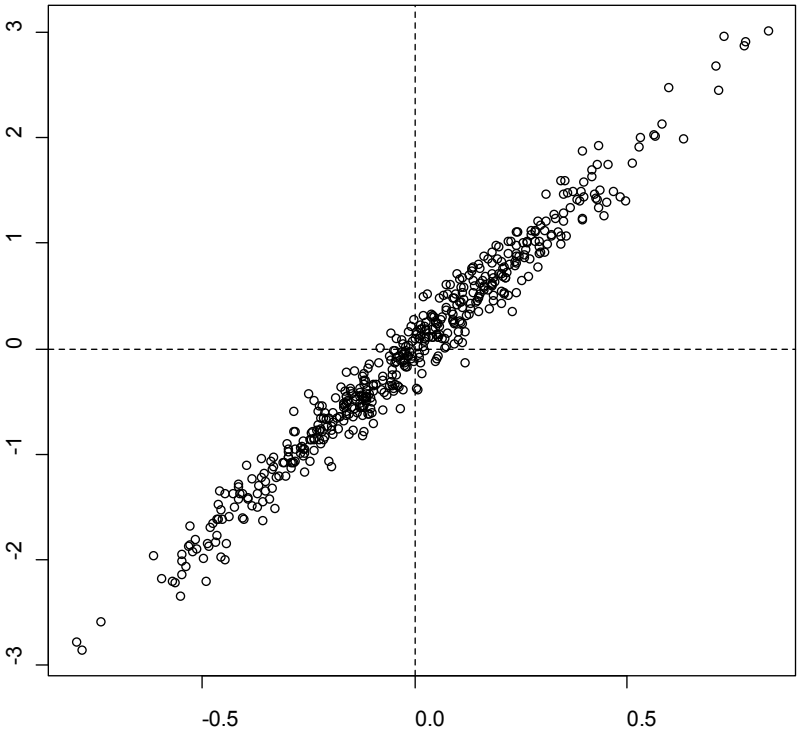


Figure 18 : Comparaison des valeurs de $\widehat{K}_{1,A_n}(2)$ et $\widehat{K}_{1,A_n}(5)$ centrées sur leur espérance pour 500 tirages d'une processus de Poisson, $\rho = 5$, $n = 10$.

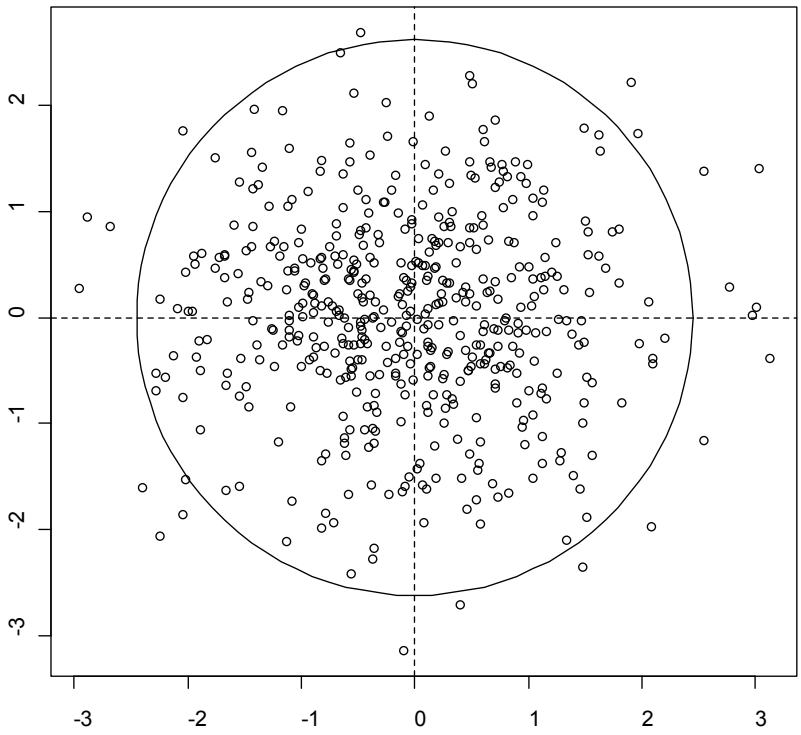


Figure 19 : Comparaison des valeurs de $T(2)$ et $T(5)$, après transformation des valeurs de $\widehat{K}_{1,A_n}(2)$ et $\widehat{K}_{1,A_n}(5)$ de la Figure 18.

Extension aux domaines rectangulaires

La démarche précédente a permis de montrer qu'un test global de la fonction de Ripley était possible, y compris pour des petits semis de points. Sa limite est que la démonstration a été faite pour un domaine carré, dans le but de montrer les propriétés asymptotiques du test en augmentant la taille du carré. Pour une application pratique, les calculs doivent être repris pour un domaine rectangulaire, noté A_{l_1,l_2}, dans lequel les tailles des côtés seront notées l_1 et l_2. Comme l'objectif est l'application empirique, seul l'estimateur $\widehat{K}_{2,A_{l_1,l_2}}(r)$ présente un intérêt : l'intensité du processus n'est jamais connue. La notation sera résumée à $\widehat{K}(r)$.

Les étapes du raisonnement sont exactement les mêmes que pour le carré. λ est estimée par $n(A_{l_1 l_2})/(l_1 l_2)$

La valeur du biais est :

$$\mathbb{E}\left(\widehat{K}(r) - K(r)\right) = \frac{4r^3(l_1+l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2} \quad \textbf{(28)}$$

La variance vaut :

$$\begin{aligned}\mathrm{Var}\left(\widehat{K}(r)\right) = {} & 2l_1^2 l_2^2 \mathbb{E}\left(\frac{\mathbf{1}\left(N(A_{l_1 l_2}) > 1\right)}{N(A_{l_1 l_2})\left(N(A_{l_1 l_2}) - 1\right)}\right)\left(e_{r,l_1,l_2} - e_{r,l_1,l_2}^2\right) \\ & + 4l_1^2 l_2^2 \mathbb{E}\left(\frac{\mathbf{1}\left(N(A_{l_1 l_2}) > 1\right)\left(N(A_{l_1 l_2}) - 2\right)}{N(A_{l_1 l_2})\left(N(A_{l_1 l_2}) - 1\right)}\right)\mathbb{E}\left(h_1^2(U,r)\right) \\ & + l_1^2 l_2^2 e^{-\lambda l_1 l_2}(1+\lambda l_1 l_2)\left(1 - e^{-\lambda l_1 l_2} - \lambda l_1 l_2 e^{-\lambda l_1 l_2}\right) e_{r,l_1,l_2}^2\end{aligned} \quad \textbf{(29)}$$

Où :

$$e_{r,l_1,l_2} = \frac{\pi r^2}{l_1, l_2} - \frac{4r^3(l_1+l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2} \quad \textbf{(30)}$$

Et :

$$\begin{aligned}\mathbb{E}\left(h_1^2(U,r)\right) = {} & \frac{r^5(l_1+l_2)}{l_1^3 l_2^3}\left(\frac{8}{3}\pi - \frac{256}{45}\right) + \frac{r^6}{l_1^3 l_2^3}\left(\frac{11}{48}\pi - \frac{8}{9} - \frac{16(l_1+l_2)^2}{l_1 l_2}\right) \\ & + \frac{4r^7(l_1+l_2)}{3l_1^4 l_2^4} - \frac{r^8}{4l_1^4 l_2^4}\end{aligned} \quad \textbf{(31)}$$

La covariance vaut :

$$
\begin{aligned}
&\operatorname{cov}\left(\widehat{K}(r),\widehat{K}(r')\right)\\
&\quad= 2l_1^2l_2^2\mathbb{E}\left(\frac{\mathbf{1}\left(N\left(A_{l_1l_2}\right)>1\right)}{N\left(A_{l_1l_2}\right)\left(N\left(A_{l_1l_2}\right)-1\right)}\right)\left(e_{r,l_1,l_2}-e_{r,l_1,l_2}e_{r\prime,l_1,l_2}\right)\\
&\quad+4l_1^2l_2^2\mathbb{E}\left(\frac{\mathbf{1}\left(N\left(A_{l_1l_2}\right)>1\right)\left(N\left(A_{l_1l_2}\right)-2\right)}{N\left(A_{l_1l_2}\right)\left(N\left(A_{l_1l_2}\right)-1\right)}\right)\operatorname{cov}\left(h_1(U,r),h_1(U,r')\right)\\
&\quad+l_1^2l_2^2e^{-\lambda l_1l_2}(1+\lambda l_1l_2)\left(1-e^{-\lambda l_1l_2}-\lambda l_1l_2e^{-\lambda l_1l_2}\right)e_{r,l_1,l_2}e_{r\prime,l_1,l_2}
\end{aligned}
\tag{32}
$$

Où :

$$
\begin{aligned}
&\operatorname{cov}\left(h_1(U,r),h_1(U,r')\right)\\
&\quad=(l_1-2r')(l_2-2r')\frac{r^2r'^2}{l_1^3l_2^3}b_{r,l_1,l_2}b_{r\prime,l_1,l_2}\\
&\quad+2(l_1+l_2-4r')\frac{r^2r'^3}{l_1^3l_2^3}b_{r,l_1,l_2}\int_{r/r'}^{1}\left(b_{r\prime,l_1,l_2}-g(x'_1)\right)dx'_1\\
&\quad+2(l_1+l_2\\
&\quad-4r')\frac{r^3r'^2}{l_1^3l_2^3}b_{r,l_1,l_2}\int_0^1\left(b_{r\prime,l_1,l_2}-g\left(\frac{rx_1}{r'}\right)\right)\left(b_{r,l_1,l_2}-g(x_1)\right)dx_1\\
&\quad+4\frac{r^2r'^4}{l_1^3l_2^3}\int_0^1\int_0^1\left[h_{A1}\left(\frac{r'x'}{r},r\right)+h_{A2}\left(\frac{r'x'}{r},r\right)+h_{A3}\left(\frac{r'x'}{r},r\right)\right.\\
&\quad\left.+h_{A4}\left(\frac{r'x'}{r},r\right)\left(h_{A3}(x',r')+h_{A4}(x',r)\right)\right]dx'_1dx'_2
\end{aligned}
\tag{33}
$$

Et :

$$
b_{r,l_1,l_2}=\pi-\frac{l_1l_2}{r^2}e_{r,l_1,l_2}=-\frac{4r(l_1+l_2)}{3l_1l_2}+\frac{r^2}{2l_1l_2}
\tag{34}
$$

Et :

$$
g(x)=I(x<1)\left(\arccos x+x\sqrt{1-x^2}\right)
\tag{35}
$$

Et :

$$
\begin{aligned}
h_{A1}(x,r)&=b_{r,l_1,l_2}\mathbf{1}(x_1\geq1)\mathbf{1}(x_2\geq1)\\
h_{A2}(x,r)&=\left(b_{r,l_1,l_2}-g(x_2)\right)\mathbf{1}(x_1\geq1)\mathbf{1}(x_2<1)\\
&\quad+\left(b_{r,l_1,l_2}-g(x_1)\right)\mathbf{1}(x_2\geq1)\mathbf{1}(x_1<1)\\
h_{A3}(x,r)&=\left(b_{r,l_1,l_2}-g(x_1)-g(x_2)\right)\mathbf{1}(x_1<1)\mathbf{1}(x_2<1)\mathbf{1}(x_1^2+x_2^2\geq1)\\
h_{A4}(x,r)&=\left(b_{r,l_1,l_2}-\frac{\pi}{4}+x_1x_2-\frac{g(x_1)+g(x_2)}{2}\right)\mathbf{1}(x_1^2+x_2^2\leq1)
\end{aligned}
\tag{36}
$$

Le détail des calculs se trouve dans Marcon *et al.* (in prep.).

Application du test à un semis de points

L'objectif du test est de rejeter l'hypothèse nulle H_0 : le semis de points observé est un Poisson homogène.

En pratique, la technique de test est la suivante :

- Estimer l'intensité du processus à partir du nombre de points : $\hat{\lambda} = \frac{\hat{n}(A)}{A}$,
- Calculer les valeurs de $\widehat{K}(r)$ pour toutes les valeurs de r choisies,

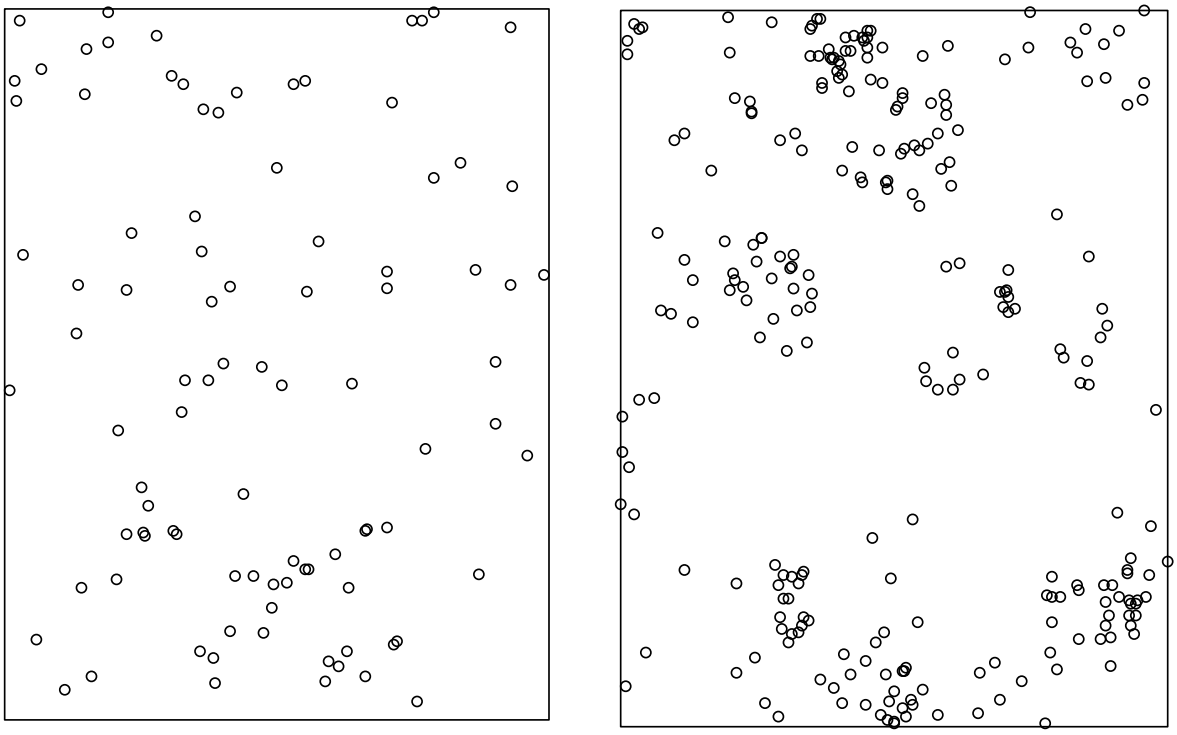


Figure 20 : Carte des *Tachigali melinonii* (94 arbres, à gauche) et *Dicorynia guianensis* (254 arbres, à droite). Les données proviennent du bloc sud du dispositif de Paracou.

- Calculer la matrice Σ selon les équations (29) et (32),
- Calculer $\mathrm{T} = \Sigma^{-1/2} \begin{pmatrix} \widehat{K}(r_1) \\ \dots \\ \widehat{K}(r_d) \end{pmatrix}$,

Au choix, comparer $\|T\|^2$ à la valeur critique de $\chi^2_\alpha(d)$ au seuil de risque α, ou bien calculer le seuil de risque correspondant à $\chi^2_\alpha(d) = \|T\|^2$, c'est-à-dire la probabilité de rejeter l'hypothèse nulle par erreur. Le code informatique nécessaire, sous R, est fourni en annexe 4.

Exemples

Données

Les données traitées sont les distributions de deux espèces d'arbres dans le Bloc Sud du dispositif forestier de Paracou (Gourlet-Fleury *et al.*, 2004 en Guyane française. Tous les arbres de plus de 10 cm de diamètre à hauteur de poitrine sont cartographiés dans un rectangle de 400,6 par 522,3 m de côté. Les cartes sont en Figure 20. Les angéliques (*Dicorynia guianensis*) sont l'espèce la plus étudiée en Guyane, leur structure spatiale a été caractérisée depuis longtemps (Goreaud *et al.*, 1997) : une inspection visuelle de la carte permet de vérifier que l'espèce est agrégative. Les tachigalis (*Tachigali melinonii*) ont été étudiés précédemment pour leurs caractéristiques biomécaniques (Jaouen *et al.*, 2010) ou la plasticité de leurs traits foliaires (Coste *et al.*, 2009). La strucutre spatiale de ses juvéniles a été caractérisée par Flores *et al.* (2006) mais pas celle des adultes.

Nous appliquons le test à ces deux jeux de points. Une courbe classique de L calculée tous les 5 mètres de 0 à 250 m est fournie en complément.

Structure de *Dicorynia guianensis*

L'agrégation ne laisse aucun doute sur la Figure 21. Le test appliqué sur le vecteur de distances (10, 20, 30,…, 150) retourne un niveau de risque de rejeter l'hypothèse nulle par erreur (*p-value*) égal à 0 : le quantile de la distribution de χ^2 à 15 degrés de liberté pour $\left\|\Sigma^{-\frac{1}{2}}\boldsymbol{K}\right\|$ est si faible que R retourne 0.

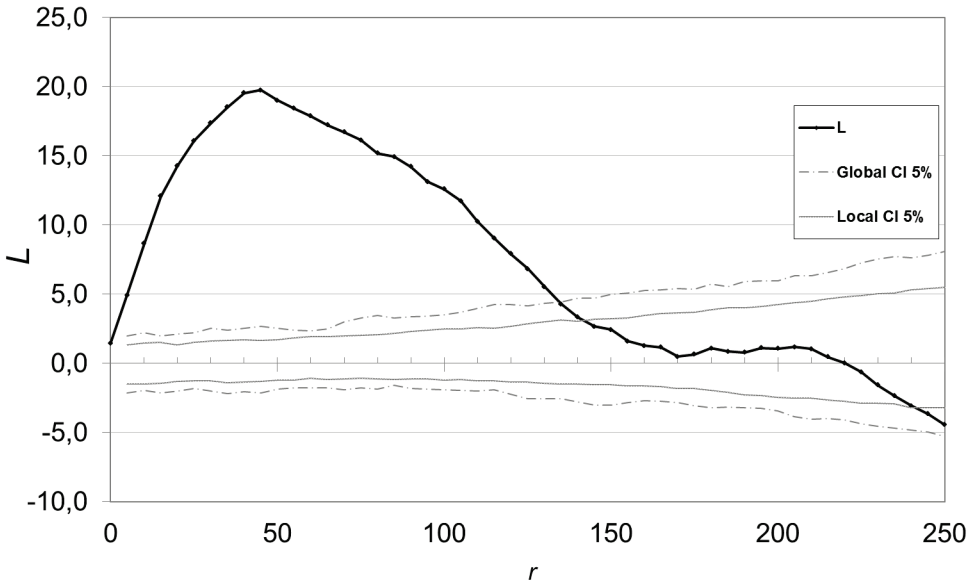


Figure 21 : Valeurs de *L* pour *Dicorynia guianensis*.

Structure de *Tachigali melinonii*

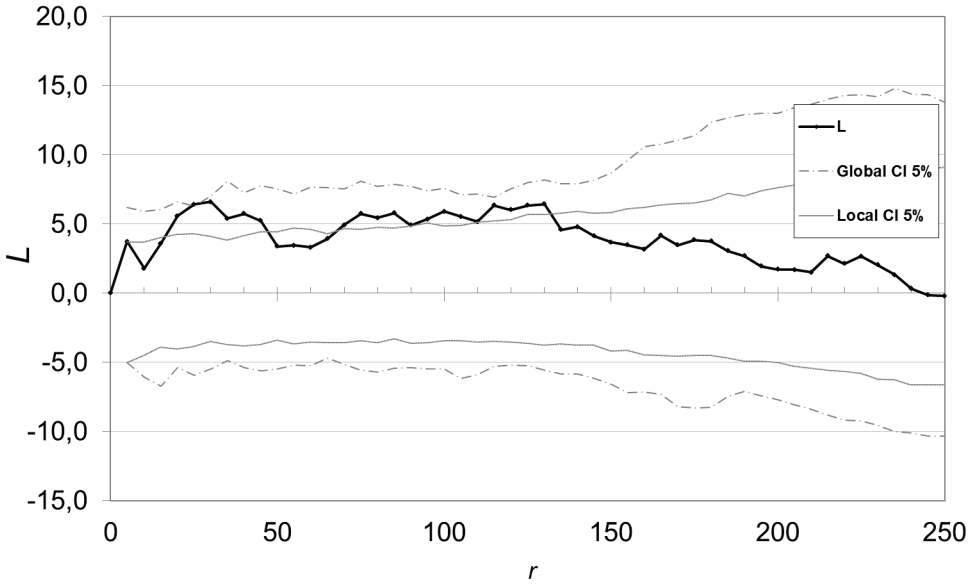


Figure 22 : Valeurs de *L* pour *Tachigali melinonii*.

La Figure 22 montre une structure moins claire pour les tachigalis. La courbe quitte l'intervalle de confiance local de l'hypothèse nulle plusieurs fois, mais pas l'intervalle global. Le test appliqué dans les mêmes conditions retourne une p-value de 2,5%, ce qui permet d'affirmer que l'agrégation est significative.

Analyse et perfectionnement de la fonction K de Ripley

L'objectif de ce paragraphe est de préparer la généralisation de la fonction K aux processus hétérogènes.

Deux méthodes de correction des effets de bord sont envisageables. Nous utiliserons celle de Besag (1977).

L'estimateur habituel de K est légèrement biaisé. Nous en donnons une version sans biais.

La fonction K peut être envisagée comme le rapport de la densité locale de voisins à sa densité moyenne, ce qui permet de l'interpréter simplement.

Correction des effets de bord

Choix de la correction de Besag

La correction des effets de bord par la méthode de Ripley, équation (9), présente l'inconvénient de ne pas être utilisable si une seule valeur de L_{ir} est nulle, c'est-à-dire si un cercle centré sur un des points se trouve entièrement hors de la zone d'étude.

Pour une zone d'étude rectangulaire, le calcul de K est traditionnellement limité à la moitié de la largeur du rectangle (Diggle, 1983). Goreaud et Pélissier (1999) ont amélioré la méthode pour étendre la correction à la moitié de la longueur du rectangle.

Nous utiliserons par la suite la méthode de Besag, équation (10), qui n'est pas limitée.

Formules de calcul explicites

Les formules de calcul des effets de bord ne se trouvent dans la littérature que pour la correction de Ripley (Goreaud et Pélissier, 1999). Nous verrons ici les formules de correction pour la méthode de Besag.

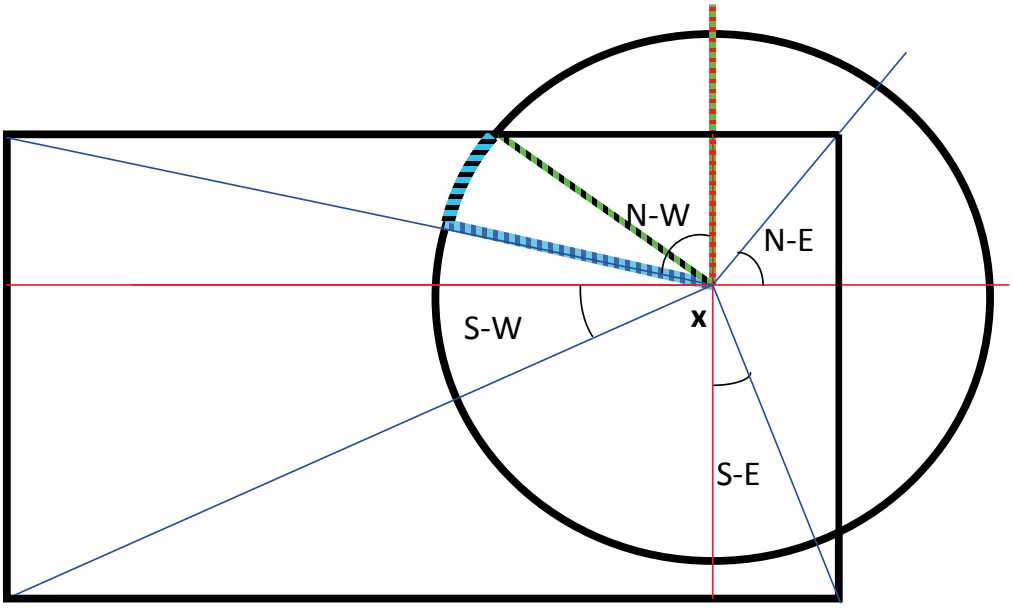


Figure 23 : Correction des effets de bord

La Figure 23 permet d'illustrer l'algorithme dans le cas d'un domaine d'étude rectangulaire. Le cercle est découpé en huit secteurs délimités par les rayons reliant le centre du cercle x aux angles du rectangle et à ses projections sur les côtés.

La méthode est applicable à tout polygone convexe. La technique consiste à traiter successivement chaque côté du polygone. On trace tout d'abord les deux rayons du cercle passant par les deux extrémités du côté (en bleu sur la figure) puis les hauteurs du triangle obtenu (en rouge). On obtient de cette façon deux triangles rectangles pour chaque côté du polygone. Chaque triangle est ensuite traité indépendamment : on note Y l'hypoténuse et H la hauteur partant de x (représentés pour le triangle W-S-W sur la figure) ; r est le rayon du cercle. Trois cas sont possibles :

- $H < Y < r$: le triangle est entièrement contenu dans le cercle (triangle N-N-E par exemple). La surface est $HY/2$.
- $r < H < Y$: le triangle contient entièrement le secteur de disque (triangle S-S-E par exemple) d'angle α. La surface est $r\arccos\frac{H}{Y}$.
- $H < r < Y$: dans ce cas intermédiaire, la surface comprend un triangle (N-N-W par exemple) de surface $H\sqrt{r^2 - H^2}/2$ et un secteur de disque (N-W) de surface $r\left(\arccos\frac{H}{Y} - \arccos\frac{H}{r}\right)$.

Correction du biais de K

Calculons $\widehat{K}(r)$ selon l'équation (10) pour une valeur de r grande, telle que la partie du cercle contenue dans l'aire d'étude soit l'aire d'étude entière : $A_{ir} = A$ pour

tous les points x_i. Tous les points sont à une distance inférieure à r de tous les autres points : $\mathbf{1}(\|x_i - x_j\| \leq r)$ vaut toujours 1. On peut donc calculer $\widehat{K}(r)$:

$$\widehat{K}(r) = \frac{A}{\big(n(A)\big)^2} \sum_{i=1}^{n(A)} \frac{\pi r^2}{A} \sum_{j=1, i \neq j}^{n(A)} 1 = \pi r^2 \frac{n(A) - 1}{n(A)} \quad \textbf{(37)}$$

Ce résultat pose problème : à grande distance, le semis de point a une structure spatiale homogène et K devrait tendre vers πr^2. Ce problème est très rarement évoqué dans la littérature parce que la correction des effets de bord par la méthode de Ripley limite la distance à une fraction de la taille de la zone d'étude.

Nous justifions ici la correction de ce biais, équation (38), mais des expressions équivalentes peuvent être trouvées dans la littérature. Getis (1984) indique que le nombre de paires de points, donc de distances, est $n(A)(n(A) - 1)$, et que l'estimateur non biaisé du carré de l'intensité est $n(A)(n(A) - 1)/A^2$. Getis et Franklin (1987) reprennent cette valeur sans explication. Diggle et Chetwynd (1991) évoquent indirectement le problème en donnant une formulation différente de K « pour obtenir un estimateur non biaisé », sans en expliquer la raison. Sweeney et Feser (1998) utilisent la méthode de Diggle et Chetwynd (1991) et donc l'estimateur non biaisé. Moeur (1993) indique que l'estimateur de K est biaisé mais que le biais est peu important, et l'utilise donc tel quel. Jones *et al.* (1996) utilisent une version non biaisée en la justifiant par la perte d'un degré de liberté. Enfin, Jones *et al.* (1999) montrent par un exemple que l'estimateur de L est biaisé quand le nombre de points est petit. Ils expliquent ce biais par la non linéarité de la transformation de K à L qui fait que si la moyenne de la valeur de K pour plusieurs réalisations d'un même processus est bien πr^2, la moyenne de la valeur de L n'est pas 0. Parallèlement, ils introduisent un estimateur de L conforme à l'équation (38), mais sans justification, et constatent sur les exemples qu'il n'est pas biaisé, sans explorer plus loin la cause de l'amélioration (changement de forme ou suppression du problème de non linéarité).

La cause du problème est à chercher dans l'estimateur du carré de l'intensité $\widehat{\lambda^2}$. Stoyan et Stoyan (2000) ont montré que l'estimateur non biaisé pour un processus de Poisson homogène est $\widehat{\lambda^2} = \sqrt{n(A_n)(n(A_n) - 1)}/n^2$. De façon plus intuitive, on compte le nombre de voisins de chaque point dans un cercle de rayon r. La densité de points utilisée dans l'équation (1) n'est pas le nombre total de points divisé par la surface ($\hat{n}(A)/A$) parce qu'un des points est obligatoirement au centre du cercle et ne peut pas se trouver dans la couronne. L'estimateur non biaisé de l'intensité est $(n(A) - 1)/A$. On peut donc définir un estimateur de K non biaisé :

$$\widehat{K}(r) = \frac{A}{n(A_n)(n(A_n)-1)} \sum_{i=1}^{n(A)} \frac{\pi r^2}{A_{ir}} \sum_{j=1, i\neq j}^{n(A)} \mathbf{1}(\|x_i - x_j\| \leq r) \quad \textbf{(38)}$$

Approche alternative

Enfin, nous pouvons réarranger l'équation (38) :

$$\boxed{\frac{\widehat{K}(r)}{\pi r^2} = \frac{\sum_{i=1}^{n(A)} \frac{\sum_{j=1, i\neq j}^{n(A)} \mathbf{1}(\|x_i - x_j\| \leq r)}{A_{ir}}}{n(A)} \Big/ \frac{n(A)-1}{A}} \quad \textbf{(39)}$$

(39) : Estimateur non biaisé de la fonction *K* de Ripley

Cette formulation de la fonction de Ripley, sans dimension, présente l'avantage de donner des valeurs plus simples à interpréter et permet une nouvelle approche de la définition de *K*.

On note λ_r l'intensité d'un processus de Poisson homogène faisant apparaître des voisins dans les cercles de rayon r autour des points. $\widehat{\lambda_r}(x_i) = \sum_{j=1, i\neq j}^{n(A)} \mathbf{1}(\|x_i - x_j\| \leq r)/A_{ir}$ en est l'estimateur autour du point x_i. Sa valeur moyenne autour de tous les points est un estimateur de λ_r. L'intensité des voisins sur tout le domaine, sans limitation à la distance r autour des points, notée λ_A, est quant à elle estimée par $(n(A)-1)/A$.

On peut donc écrire *K* sous la forme :

$$\frac{\widehat{K}(r)}{\pi r^2} = \frac{\widehat{\lambda_r}}{\widehat{\lambda_A}} \quad \textbf{(40)}$$

La fonction *K*, après normalisation par la surface du cercle sur laquelle elle est appliquée, est le rapport entre la densité de voisins des points dans le cercle et la densité de voisins sur tout le domaine d'étude. Ce résultat permet de mieux comprendre le sens de la fonction *K* et de justifier la méthode de suppression du biais utilisée au paragraphe précédent.

La grandeur $\widehat{K}(r)/\pi r^2$ constitue une alternative avantageuse à la fonction $\hat{L}(r)$. Elle est sans dimension, sa valeur de référence est égale à 1. Sa valeur estimée est le rapport entre la densité de voisins observée à la distance r et la densité de voisins de référence. Les pics de $\widehat{K}(r)/\pi r^2$ correspondent aux rayons pour lesquels la densité de voisins est maximum.

Généralisation de la fonction de Ripley aux processus hétérogènes

La fonction K ne peut pas être appliquée aux processus non stationnaires. Pélissier et Goreaud (2001) proposent des méthodes de partition d'une zone d'étude hétérogène en surfaces plus petites homogènes. Cette méthode est applicable si le domaine d'étude est une juxtaposition de zones homogènes dont les limites peuvent être définies, par des critères liés à l'espace (le type de sol ou la pente par exemple pour un peuplement forestier) ou au semis de points (la hauteur des arbres,…). La technique consiste dans ce cas à calculer les propriétés locales de premier et de second ordre du processus de façon systématique, par exemple sur une grille régulière, et définir une zone comme homogène lorsque ces propriétés y sont à peu près constantes. La méthode est d'autant plus efficace que les transitions sont nettes et n'est pas utilisable face à des variations continues de la propriété testée.

Etat de l'art

La fonction D de Diggle et Chetwynd

Diggle et Chetwynd (1991) introduisent une généralisation de la fonction K pour l'étude de processus ponctuels non homogènes (d'intensité variable dans le domaine d'étude).

On considère un semis de points de deux types, le premier à étudier (appelés *cas*), le second utilisé comme référence (les *contrôles*). On peut calculer les deux fonctions K pour chacune des deux populations, K_c et K_0.

D est définie par :

$$D(r) = K_c(r) - K_0(r) \quad \textbf{(41)}$$

(41) : La fonction D de Diggle et Chetwynd

La distribution spatiale de tous les points étant donnée, l'hypothèse nulle est *l'étiquetage aléatoire* : chaque point peut être aléatoirement un cas ou un contrôle, la seule

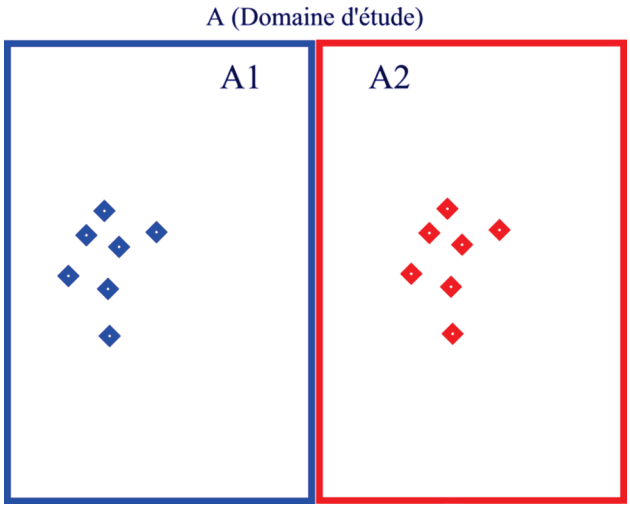


Figure 24 : Limite de l'application de la fonction de Diggle et Chetwynd

contrainte étant le respect du nombre total de points de chaque type. Sous cette hypothèse, $K_c(r) = K_0(r)$. Si $K_c(r) > K_0(r)$, les cas sont plus agrégés que les contrôles, et inversement. L'intervalle de confiance de l'hypothèse nulle est calculé en générant des jeux de points dont la position est celle du semis réel et l'étiquetage aléatoire.

Les processus ponctuels réels étant fréquemment hétérogènes, l'utilisation de la fonction *D* a connu un certain succès (Kempton et Taylor, 1976 ; Kingham *et al.*, 1995 ; Gatrell et Bailey, 1996 ; Gatrell *et al.*, 1996 ; Jones *et al.*, 1996 ; Sweeney et Feser, 1998 ; Feser et Sweeney, 2000; 2002 ; Jolles *et al.*, 2002 ; Kelly et Meentemeyer, 2002 ; Marcon et Puech, 2003).

La limite fondamentale de la fonction *D* peut être illustrée par un exemple (Figure 24). Considérons un domaine d'étude composé de deux sous-domaines identiques juxtaposés (*A1* et *A2*). Dans chacun de ces sous-domaines, on place une réalisation strictement identique du même processus ponctuel, mais on attribue une étiquette différente aux points selon le sous-domaine dans lequel ils se trouvent (les points bleus et les points rouges). On calcule la fonction *D* pour l'un des deux types de points.

Comme les deux semis sont identiques, les deux fonctions *K* pour les cas et les contrôles sont égales. La fonction *D* est donc toujours nulle, suggérant l'absence de structure spatiale, alors que les points rouges et bleus sont visiblement agrégés et qu'on est très loin de l'hypothèse nulle de l'étiquetage aléatoire.

La fonction *D* n'est pas puissante pour détecter les structures spatiales parce qu'elle calcule séparément les deux fonctions *K* puis les compare globalement. Une grande partie de l'information, la position relative des points, est donc perdue.

Enfin, la valeur de *D*, différence de deux fonctions *K*, n'est pas simple à interpréter. Toutes les applications citées dans la revue de la littérature se contentent de rechercher l'existence d'un écart à l'hypothèse nulle. Diggle et Chetwynd interprètent la valeur de *D(r)* multipliée par l'intensité des cas comme le nombre de cas excédentaires rencontrés dans le cercle de rayon *r* autour de chaque cas, le nombre de voisins attendu étant celui correspondant à une distribution des cas identique à celle des contrôles. On peut le vérifier facilement (les cas sont notés x_i^c et les contrôles x_i^0) :

$$\begin{aligned}\widehat{D}(r) &= \widehat{K}_c(r) - \widehat{K}_0(r) \\ &= \frac{1}{n_c(A)\widehat{\lambda_c}} \sum_{i=1}^{n_c(A)} \sum_{j=1, i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i^c - x_j^c\right\| \leq r\right) c(i,j,r) \\ &\quad - \frac{1}{n_0(A)\widehat{\lambda_0}} \sum_{i=1}^{n_0(A)} \sum_{j=1, i\neq j}^{n_0(A)} \mathbf{1}\left(\left\|x_i^0 - x_j^0\right\| \leq r\right) c(i,j,r)\end{aligned} \quad \textbf{(42)}$$

Pour alléger les écritures, notons $v_c(i,r) = \sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) c(i,j,r)$ le nombre de cas voisins du cas x_i^c, et $v_0(i,r)$ le nombre de contrôles voisins du contrôle x_i^0. Leur nombre moyen est noté respectivement $\overline{v_c}(r) = \frac{1}{n_c(A)} \sum_{i=1}^{n_c(A)} v_c(i,r)$ et $\overline{v_0}(r)$. On peut alors écrire :

$$\widehat{\lambda_c}\, \widehat{D}(r) = \overline{v_c}(r) - \overline{v_0}(r) \frac{\widehat{\lambda_c}}{\widehat{\lambda_0}} \quad \textbf{(43)}$$

En d'autres termes, $\widehat{\lambda_c}\, D(r)$ est égal au nombre moyen de cas voisins de chaque cas moins le nombre moyen de contrôles voisins de chaque contrôle multiplié par le rapport de l'intensité des cas sur celle des contrôles. Cette dernière valeur est précisément égale au nombre moyen de cas voisins dans le cadre de l'hypothèse nulle de l'étiquetage aléatoire. La différence est bien un nombre moyen de cas voisins excédentaire.

Cette interprétation illustre une autre limite de D : ses valeurs ne sont pas comparables quand r change. En effet le même nombre de points excédentaires n'a pas la même signification à petite distance (où le nombre de voisins attendu dans le cadre de l'hypothèse nulle est petit) et à grande distance (où le nombre de voisins attendu est grand).

La fonction D permet donc de rechercher l'existence de structures spatiales dans un espace hétérogène. Ses deux limites principales sont son manque de puissance (son incapacité à détecter certains types de structures) et l'impossibilité d'affecter un poids aux points.

La fonction K_{inhom} de Baddeley *et al.* (2000)

La fonction K_{inhom} est une généralisation de la fonction K de Ripley à des processus non stationnaires. La stationnarité est nécessaire pour la propriété de second ordre du processus ponctuel : on conçoit assez aisément que le calcul des interactions entre couples de points à une distance donnée n'a de sens que si elles sont identiques partout dans le domaine d'étude. On s'intéressera donc à des processus stationnaires au second ordre après repondération de leur intensité (*second-order intensity-reweighted stationary*).

Revenons à la définition de la fonction de corrélation des paires de points, g, équation (103). La fonction K est définie en intégrant la fonction g sur le cercle de rayon r, équation (3). Lorsque le processus n'est pas homogène, l'intégration donne :

$$K_{inhom} = \frac{1}{A}\mathbb{E}\left[\sum_i \sum_{j,i\neq j} \frac{\mathbf{1}(\|x_i - x_j\| \leq r)}{\lambda(x_i)\lambda(x_j)}\right] \quad \textbf{(44)}$$

$\lambda(x_i)$ est l'intensité du processus au point x_i. La fonction peut être estimée par :

$$\boxed{\widehat{K}_{inhom} = \frac{1}{A}\sum_i \sum_{j,i\neq j} \frac{\mathbf{1}(\|x_i - x_j\| \leq r)c(i,j,r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)}} \quad \textbf{(45)}$$

(45) : Fonction K_{inhom} de Baddeley *et al.*

L'indicatrice est corrigée des effets de bord par la méthode de Ripley, détaillée par Stoyan *et al.* (1987 p. 134-137). On voit immédiatement que si l'intensité est constante, on se ramène à la formule classique.

Sur le plan théorique, le traitement des processus non stationnaires est résolu. Sur le plan pratique, il en est tout autrement : la difficulté réside dans l'estimation des densités locales $\lambda(x_i)$ et $\lambda(x_j)$. La solution naturelle consiste à utiliser un estimateur par la méthode des noyaux (Diggle, 1985 ; Silverman, 1986) qui consiste à prendre en compte chaque point et ses voisins avec un poids décroissant avec la distance, le tout pondéré de façon que l'intégrale de l'intensité locale sur l'ensemble du domaine d'étude soit bien égale au nombre de points total. Les auteurs mentionnent un sévère biais dans l'estimation de la fonction K_{inhom} de cette façon pour un processus agrégatif. La raison est simple à comprendre : dans les agrégats, l'intensité observable (estimée en comptant le nombre de points) est très supérieure à l'intensité du processus. En d'autres termes, l'estimation d'un processus par la méthode des noyaux intègre dans sa propriété de premier ordre (l'intensité), ses propriétés de second ordre. Les auteurs proposent un meilleur estimateur pour les processus de Poisson inhomogènes, mais laissent les autres cas en suspens (notamment les processus dont la propriété de second ordre est négative). La conclusion est que la connaissance formelle au moins approximative du processus sous-jacent est indispensable pour distinguer l'effet de l'intensité de celui de l'agrégation ou de la dispersion.

La fonction de Baddeley *et al.* constitue un aboutissement théorique dans la caractérisation des processus ponctuels non stationnaires. Bizarrement, il n'a eu que peu d'impact, et est peu cité, surtout dans la littérature empirique (Bonneu, 2007 est une exception). Le problème fondamental est l'extrême difficulté tant

théorique que pratique de l'estimation des densités locales : un noyau trop petit accentue les variations de densité et gomme les interactions entre points, un noyau trop grand traite le processus comme s'il était homogène.

La fonction g_{inhom} de Baddeley *et al.* (2000)

Baddeley *et al.* introduisent sans aller plus loin un estimateur de la fonction g pour les processus hétérogènes, repris plus en détail par Law *et al.* (2009).

La densité de paires de points λ_2, définie dans l'équation (100), peut être estimée avec une fonction de noyau, comme le noyau uniforme :

$$k(\|x-y\|,r) = \begin{cases} 1/(2h) & \text{si } r-h \le \|x-y\| \le r+h \\ 0 & \text{sinon} \end{cases} \quad \textbf{(46)}$$

h est la bande passante choisie. Si deux points x et y sont distant de $r \pm h$, la fonction vaut $1/(2h)$.

Pour un processus homogène au second ordre, c'est-à-dire tel que $\lambda_2(x,y) = \lambda_2(\|x-y\|) = \lambda_2(r)$, l'estimateur est, avec une correction des effets de bord :

$$\widehat{\lambda_2}(r) = \frac{1}{2\pi r} \sum_{i=1}^{n(A)} \sum_{j=1, i\neq j}^{n(A)} k(\|x_i - x_J\|, r) c(i,j,r) \quad \textbf{(47)}$$

Finalement, g est estimé par :

$$\hat{g}_{inhom}(r) = \frac{1}{2\pi r} \sum_{i=1}^{n(A)} \sum_{j=1, i\neq j}^{n(A)} \frac{k(\|x_i - x_J\|, r) c(i,j,r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)} \quad \textbf{(48)}$$

La difficulté réside encore dans l'estimation des densités locales, compliquée par le choix de la bande passante pour l'estimation de la fonction de noyau.

La fonction K_d de Duranton et Overman

Duranton et Overman (2005) ont développé la fonction K_d, apparemment indépendamment des travaux de Ripley puisque ce dernier n'est pas cité, pour mesurer la concentration géographique des firmes manufacturières en Grande Bretagne.

Ils définissent $K_d(r)$ comme le nombre moyen de points localisés exactement à la distance r de chaque point (et non à une distance inférieure ou égale à r). Comme pour la fonction de Ripley, le calcul est réalisé par intervalles successifs. Elle est

ensuite lissée pour obtenir une courbe continue et normalisée pour que son intégrale entre 0 et l'infini soit égale à 1. Le lissage est réalisé par une fonction de noyau, comme dans (46). $K_d(r)$ est proportionnelle à $2\pi r\hat{\lambda}_2(r)$, sans correction des effets de bord ($c(i,j,r) = 1$).

La valeur de K_d est calculée pour toutes les distances choisies, sans correction des effets de bord, puis comparée à l'intervalle de confiance de l'hypothèse nulle d'une distribution aléatoire des points non sur l'ensemble de l'espace (le domaine d'étude n'est forcément pas défini avec certitude), mais sur l'ensemble des localisations observées sur le semis de point réel. L'agrégation est détectée à une distance donnée si la valeur de la fonction est significativement plus grande que celle de l'hypothèse nulle, la dispersion si elle est plus petite.

La Figure 25 présente la courbe de K_d pour un semis de points constitué de 9 agrégats répartis régulièrement. Les pics correspondent aux distances où le nombre de voisins est le plus important : dans les agrégats (autour de 0,5), à la distance entre agrégats (autour de 4, de 5,5 en diagonale et de 8). Les courbes sont calculées avec un pas de 0,5, un noyau d'Epanechnikov de bande passante égale au pas de calcul et les intervalles de confiances globaux sont calculés à partir de 10 000 simulations au seuil de 1%.

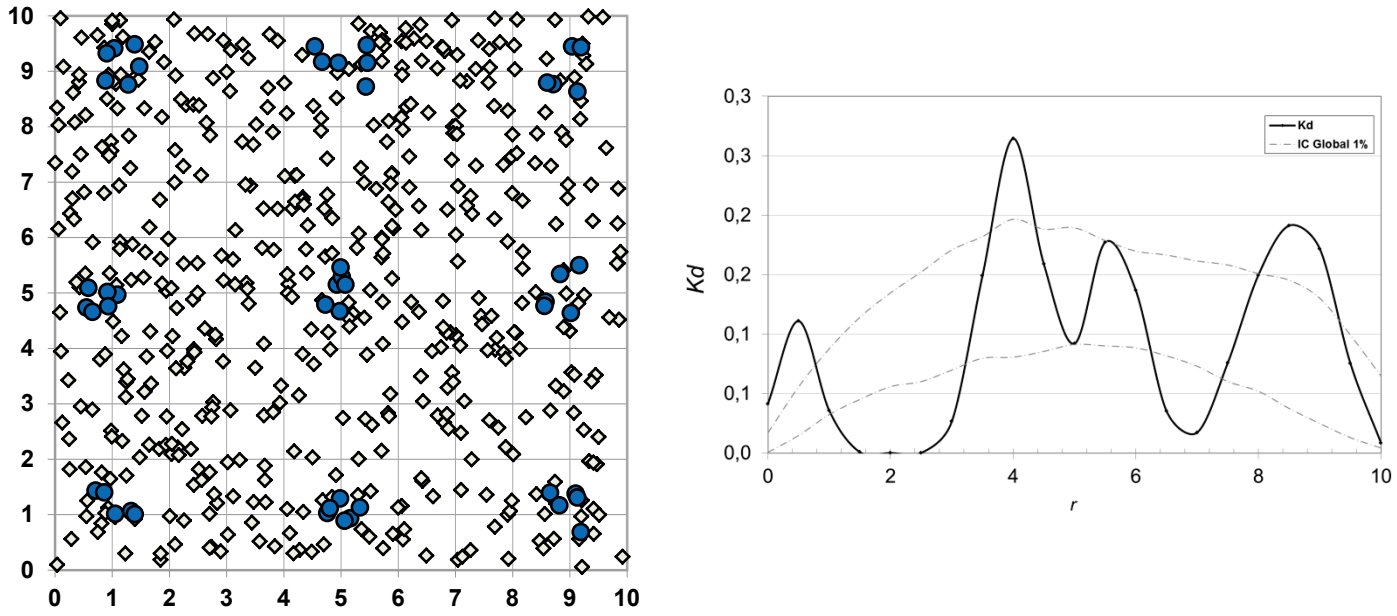


Figure 25 : Agrégats répartis régulièrement, exemple identique à la Figure 28. Carte des points et fonction K_d calculée pour les cercles bleus, distribués selon un processus de Matérn. Les losanges ne sont pas pris en compte.

Cette construction présente quelques inconvénients :

- Sa valeur ne peut pas être interprétée : l'agrégation ou la dispersion peuvent être détectées, mais pas quantifiées. L'intervalle de confiance de l'hypothèse nulle donne une idée du comportement de la fonction appliquée à un semis de points complètement aléatoire : elle croît d'abord parce que la densité de paires de points n'est pas normalisée par $1/(2\pi r)$. Elle décroît dès que les effets de bord deviennent prédominants, parce que les possibles voisins d'un point se raréfient quand la distance est grande (ils tomberaient au-delà du domaine d'étude).
- Comme l'intégrale de la fonction est égale à 1 pour le semis de points réel comme pour l'hypothèse nulle, si la courbe de K est au-dessus de celle de l'hypothèse nulle pour une plage de distances, elle sera nécessairement en dessous ailleurs. Autrement dit, l'observation de l'agrégation à une certaine distance entraîne mécaniquement la détection de la dispersion à une autre. Pour cette raison, les auteurs limitent leur analyse à une plage de distance relativement réduite (180 km pour l'ensemble des firmes de Grande Bretagne) qui est effectivement pertinente pour le problème traité. Mais rien n'assure que les agglomérations détectées ne soient pas qu'un artéfact dû à la dispersion à une échelle plus grande.

La fonction K_d de Duranton et Overman présente l'avantage par rapport à celle de Ripley de pouvoir prendre en compte un poids pour les points (ici, le nombre d'employés de chaque firme) en traitant alors chaque point comme un groupe de points superposés. L'hypothèse nulle est alors la distribution aléatoire de ces points pondérés sur l'ensemble des localisations possibles. Son deuxième avantage est d'être libérée des calculs de correction des effets de bord, ce qui rend son application beaucoup plus simple, mais au prix de résultats ininterprétables : K_d est la densité de probabilité de trouver un voisin à une distance donnée, non corrigée des effets de bord, non normalisée par l'intensité du processus.

La statistique *O-ring* de Wiegand *et al.*

Wiegand et Moloney (Wiegand *et al.*, 1999 ; Wiegand et Moloney, 2004) argumentent en faveur de l'utilisation d'une fonction de densité plutôt qu'une fonction cumulative, comme Duranton et Overman, pour obtenir des informations sur les phénomènes se déroulant à la distance r plutôt que jusqu'à la distance r.

Ils définissent la statistique *O-ring*, égale au produit de g par la densité des voisins :

$$O(r) = \lambda g(r) \quad \textbf{(49)}$$

Ils utilisent la méthode de correction des effets de bord par pixellisation présentée page 27 pour traiter des domaines de forme complexe.

Pour les processus hétérogènes, ils suggèrent d'estimer $\lambda(x)$ en tout point du domaine par un noyau constitué par un cercle de rayon choisi autour de x : la densité est estimée dans le cercle par le rapport du nombre de points sur le nombre de pixels multiplié par leur surface élémentaire. La statistique $O(r)$ n'est pas modifiée, mais la simulation de l'hypothèse nulle est faite dans un Poisson hétérogène d'intensité $\lambda(x)$.

Comme Duranton et Overman, Wiegand et Moloney renoncent à prendre en compte l'hétérogénéité dans leur statistique mais l'intègrent dans leur hypothèse nulle. La statistique n'est pas interprétable directement, mais sa significativité est détectée par sa sortie de l'intervalle de confiance. La Figure 26 présente un cas de processus de Poisson hétérogène traité de cette façon : l'intensité est estimée puis utilisée pour simuler l'hypothèse nulle. Les intervalles de confiance suivent aussi bien les courbes de L que de O, sauf à faible distance où une répulsion est mise en évidence.

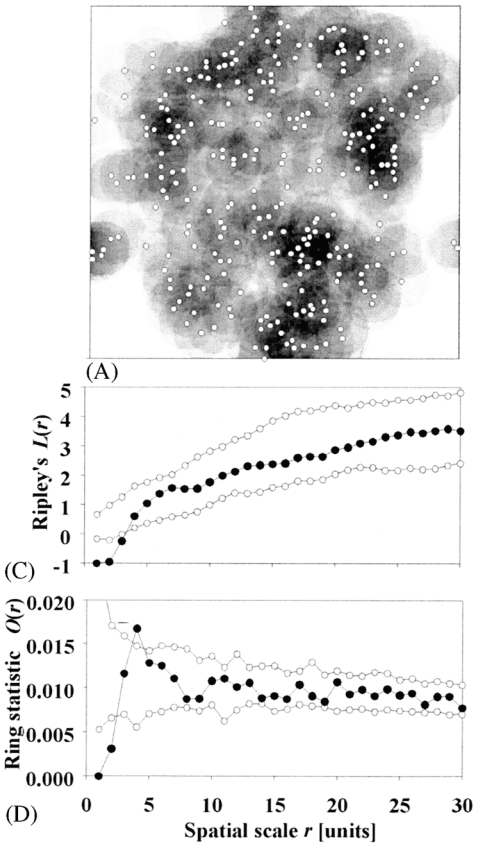


Figure 26 : Analyse d'un processus hétérogène par Wiegand et Moloney (2004, Fig. 7). La carte des points est en haut (A). Les points sont dessinés en blanc sur un fond d'autant plus sombre que la densité est grande. Les courbes de L et de O sont tracées (C et D). Les intervalles de confiance locaux de l'hypothèse nulle sont calculés en simulant un Poisson hétérogène.

Généralisation de K

L'objectif est maintenant de généraliser la fonction K pour la rendre opérationnelle sur des espaces hétérogènes et des points pondérés.

Estimateur probabiliste de K

On a vu à l'équation (40) que $\widehat{K}(r)/\pi r^2 = \widehat{\lambda_r}/\widehat{\lambda_A}$, c'est-à-dire le rapport entre la densité de voisins autour d'un point et la densité de points sur tout le domaine.

Définissons une épreuve de Bernoulli consistant à rechercher un voisin autour d'un point x_i dans une surface élémentaire dS du cercle de rayon r. Sa probabilité

de succès est $P_r = \lambda_r dS$, estimée par $\widehat{P_r} = \widehat{\lambda_r} dS$. Définissons une autre épreuve de Bernoulli consistant à rechercher un voisin autour du point x_i, cette fois sur l'ensemble du domaine d'étude. Sa probabilité de succès est de la même façon $P_A = \lambda_A dS$.

Finalement :

$$\frac{\widehat{K}(r)}{\pi r^2} = \frac{\widehat{\lambda_r}}{\widehat{\lambda_A}} = \frac{\widehat{P_r}}{\widehat{P_A}} \quad \textbf{(50)}$$

$K(r)/\pi r^2$ peut être interprété comme le rapport de deux probabilités de succès de lois de Bernoulli, que nous notons P_r et P_A.

La même interprétation peut être faite pour g. L'équation (103) peut être lue comme le rapport de la probabilité conditionnelle de trouver le point x_2 sachant l'existence du point x_1 (c'est-à-dire de trouver $x_2$voisin de x_1) sur la probabilité de trouver le point x_2 sans contrainte.

Espace hétérogène

La définition des épreuves de Bernoulli peut être changée facilement pour prendre en compte l'hétérogénéité de l'espace. Plutôt que de rechercher des voisins avec une probabilité égale dans des surfaces élémentaires, on va rechercher des voisins d'un certain type parmi l'ensemble des points existants, dont la position est considérée comme donnée.

Comme on l'a déjà fait précédemment, on appelle *cas* les points étudiés et *contrôles* les autres. L'épreuve de Bernoulli consiste à rechercher les cas parmi tous les points voisins de x_i. Sa probabilité de succès est estimée par la moyenne du rapport du nombre de cas sur le nombre de cas et de contrôles se trouvant dans la surface considérée (le cercle de rayon r ou le domaine d'étude entier). Définissons l'indicatrice $\mathbf{1}(\|x_i^c - x_j^c\| \leq r)$ qui vaut 1 si les deux points x_i^cet x_j^c sont des cas et si la distance entre eux est inférieure ou égale à r, 0 sinon. En détail :

- P_r est estimée par la moyenne sur l'ensemble des cas du rapport du nombre de cas voisins notés x_j^c sur le nombre de points (cas et contrôles, notés x_j) voisins : $\frac{1}{n_c(A)} \sum_{i=1}^{n_c(A)} \frac{\sum_{j=1, i\neq j}^{n_c(A)} \mathbf{1}(\|x_i^c - x_j^c\| \leq r) c(i,r)}{\sum_{j=1, i\neq j}^{n(A)} \mathbf{1}(\|x_i^c - x_j\| \leq r) c(i,r)}$. La correction des effets de bord de Besag ne dépend pas de j, donc elle se simplifie.
- P_A est estimée de la même façon, mais son expression est plus simple : pour tous les points x_i, le nombre de cas voisins sur l'ensemble du domaine est $n_c(A) - 1$ et le nombre de voisins de tous types est $n(A) - 1$.

Nous définissons donc la fonction $H_c = P_r/P_A$, généralisation de K aux espaces hétérogènes :

$$\widehat{H_c} = \frac{\sum_{i=1}^{n_c(A)} \frac{\sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i^c - x_j^c\right\| \leq r\right)}{\sum_{j=1,i\neq j}^{n(A)} \mathbf{1}\left(\left\|x_i^c - x_j\right\| \leq r\right)}}{n_c(A)} \Bigg/ \frac{n_c(A)-1}{n(A)-1} \qquad \textbf{(51)}$$

Poids des points

Nous définissons $M = P_r/P_A$ en autorisant une pondération des points, appliquée aux résultats des épreuves de Bernoulli. $w\left(x_j^c\right)$ est le poids du cas j, et W_c le poids total des cas, W celui de tous les points. Cette fois :

- P_r est estimée par la moyenne sur l'ensemble des cas du rapport du poids des cas voisins sur le poids de tous les points (cas et contrôles) voisins : $\frac{1}{n_c(A)} \sum_{i=1}^{n_c(A)} \frac{\sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i^c - x_j^c\right\| \leq r\right) w\left(x_j^c\right)}{\sum_{j=1,i\neq j}^{n(A)} \mathbf{1}\left(\left\|x_i^c - x_j\right\| \leq r\right) w(x_j)}$.
- $P_A$est estimée de la même façon. Son expression n'est plus aussi simple que pour les points non pondérés : pour chaque point x_i^c le rapport est $\left(W_c - w(x_i^c)\right)/\left(W - w(x_i^c)\right)$, sa valeur n'est plus la même pour chaque point. Sa valeur moyenne est $\frac{1}{n_c(A)} \sum_{i=1}^{n_c(A)} \frac{W_c - w(x_i^c)}{W - w(x_i^c)}$.

Nous retiendrons donc, après simplifications, l'estimateur suivant pour la fonction M :

$$\boxed{\widehat{M}(r) = \sum_{i=1}^{n_c(A)} \frac{\sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i^c - x_j^c\right\| \leq r\right) w\left(x_j^c\right)}{\sum_{j=1,i\neq j}^{n(A)} \mathbf{1}\left(\left\|x_i^c - x_j\right\| \leq r\right) w(x_j)} \Bigg/ \sum_{i=1}^{n_c(A)} \frac{W_c - w(x_i^c)}{W - w(x_i^c)}} \qquad \textbf{(52)}$$

(52) : Estimateur non biaisé de la fonction *M*

Version cas/contrôles

En général, le nombre total de voisins de tous types, y compris les cas, est la bonne référence : dans un peuplement forestier, les arbres étudiés occupent l'espace comme les autres ; le raisonnement est similaire en économie (Ellison et Glaeser, 1997 ; Combes et Overman, 2004).

Sans rien modifier au raisonnement, précédent, la fonction peut être adaptée aux échantillonnages dans lesquels la population totale est mieux représentée par les contrôles seuls que par l'ensemble des cas et des contrôles. Cette situation est fréquente en épidémiologie, où les contrôles sont choisis dans les protocoles pour

représenter la population totale, et les cas sont les malades. La version cas-contrôles de M est :

$$\widehat{M}_{cas}(r) = \sum_{i=1}^{n_c(A)} \frac{\sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i^c - x_j^c\right\| \leq r\right) w\left(x_j^c\right)}{\sum_{j=1,i\neq j}^{n_0(A)} \mathbf{1}\left(\left\|x_i^c - x_j^0\right\| \leq r\right) w\left(x_j\right)} \Bigg/ \frac{W_c(n_c - 1)}{W_0} \quad \textbf{(52)}$$

Le dénominateur se simplifie par rapport à la version standard, parce que le centre de chaque cercle n'est pas comptabilisé dans les contrôles.

Un exemple d'utilisation de cette version de M est donné Figure 51, page 173.

Indéterminations

Un problème apparaît pour les points autour desquels aucun voisin ne se trouve dans le cercle de rayon r. Le rapport $\frac{\sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}\left(\left\|x_i^c - x_j^c\right\| \leq r\right) w\left(x_j^c\right)}{\sum_{j=1,i\neq j}^{n(A)} \mathbf{1}\left(\left\|x_i^c - x_j\right\| \leq r\right) w\left(x_j\right)}$ est alors indéterminé (le numérateur et le dénominateur sont nuls).

Ces points ne seront pas pris en compte dans le calcul de M. Un point contribuera donc au calcul de M à la distance r si et seulement s'il a au moins un voisin d'un type quelconque.

Aux petites distances, il peut donc arriver que M ne soit pas définie (aucun point n'a de voisins) ou définie par peu de points. Dans ce dernier cas, la fonction est peu robuste, on peut s'en convaincre en observant l'étendue de l'intervalle de confiance de l'hypothèse nulle par exemple pour les premières valeurs de r sur la Figure 27, page 64.

Significativité

L'hypothèse nulle à laquelle on confronte la fonction M est comme précédemment la distribution aléatoire des points. Cependant, l'espace n'est plus homogène, et la distribution aléatoire des points doit être conforme à la propriété de premier ordre du processus analysé : l'objectif de M est de détecter la propriété de second ordre d'un processus ponctuel en s'affranchissant des variations de sa propriété de premier ordre.

Un semis de points généré dans le cadre de l'hypothèse nulle doit donc respecter la propriété de premier ordre du processus sous-jacent (les valeurs locales de l'intensité) dont le semis de point réel est une réalisation et ses points doivent être distribués indépendamment les uns des autres (propriété de second ordre égale à 1).

La difficulté pratique vient du fait qu'on ne connaît pas le processus ponctuel ayant donné naissance au semis de points réel, et qu'on ne dispose que d'une seule réalisation. La propriété de premier ordre est donc largement inconnue. À titre d'exemple, on peut se reporter au processus de Neyman-Scott (voir Figure 5, page 27) dont on a du mal à se convaincre de l'homogénéité à partir d'une seule réalisation.

Nous suivrons donc Duranton et Overman (2005) en redistribuant aléatoirement l'ensemble des points (qui conservent leur type et leur poids) sur l'ensemble des localisations réelles. Compte tenu des informations disponibles, il s'agit probablement de la meilleure façon de générer un semis de point dont l'intensité locale est une réalisation de celle du processus.

Exemples

Agrégats

On considère un ensemble de points de deux types différents : 500 contrôles distribués complètement aléatoirement et 50 cas générés par un processus de Matérn dans 9 agrégats de rayon 0,5. Le poids de tous les points est égal à 1.

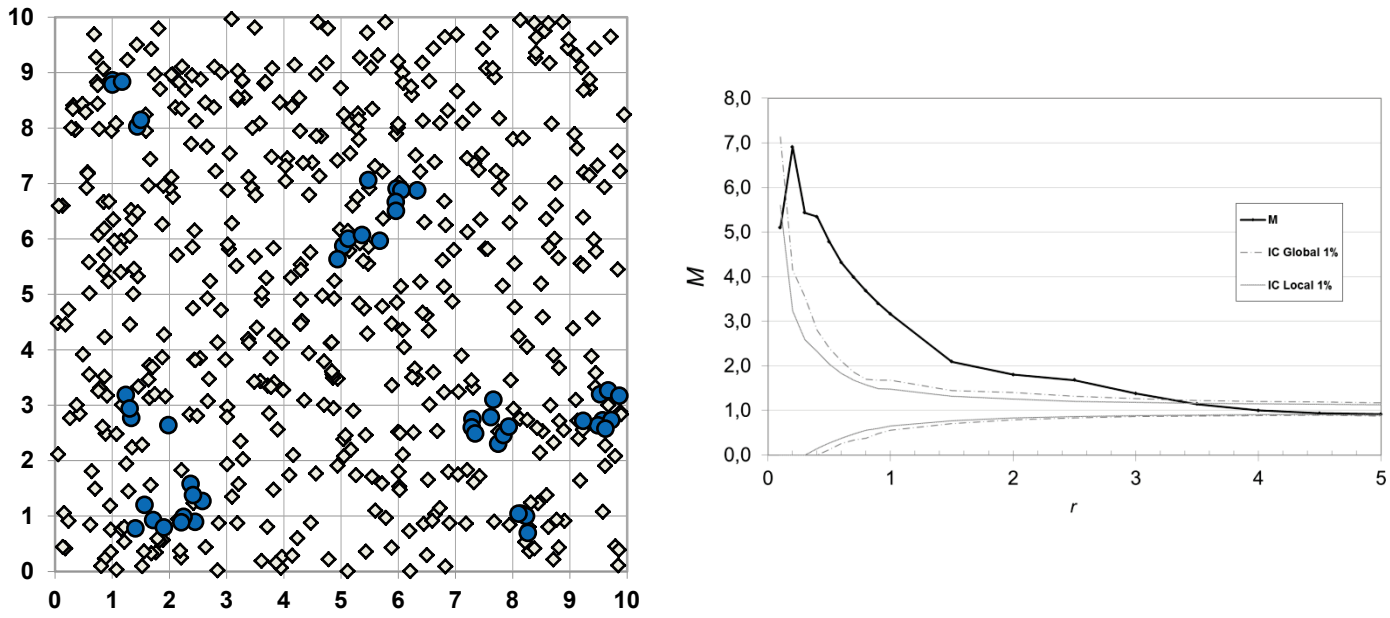


Figure 27 : Semis agrégé. Carte des points et fonction *M*.
Les contrôles (losanges) sont distribués complètement aléatoirement et les cas (cercles bleus) selon un processus de Matérn.

La carte est présentée en Figure 27. L'intervalle de confiance est calculé au seuil de 1% à partir de 10 000 simulations. La courbe est calculée avec un pas variable : 0,1 jusqu'à 1, 1 ensuite.

La courbe M a une allure similaire à celle de la courbe L : les pics positifs dénotent la concentration. Cependant, les pics ne correspondent pas à la taille des agrégats mais aux distances pour lesquelles l'intensité locale est la plus grande, c'est-à-dire à peu près à l'espacement entre les points dans les agrégats.

Régularité

L'exemple suivant (Figure 28) est identique au premier, à l'exception des agrégats qui ont été placés régulièrement sur la carte.

Les agrégats sont bien détectés à petite échelle, avec des valeurs proches de celles de l'exemple précédent. Les valeurs négatives, sous l'intervalle de confiance, mettent en évidence la régularité de l'espacement des agrégats.

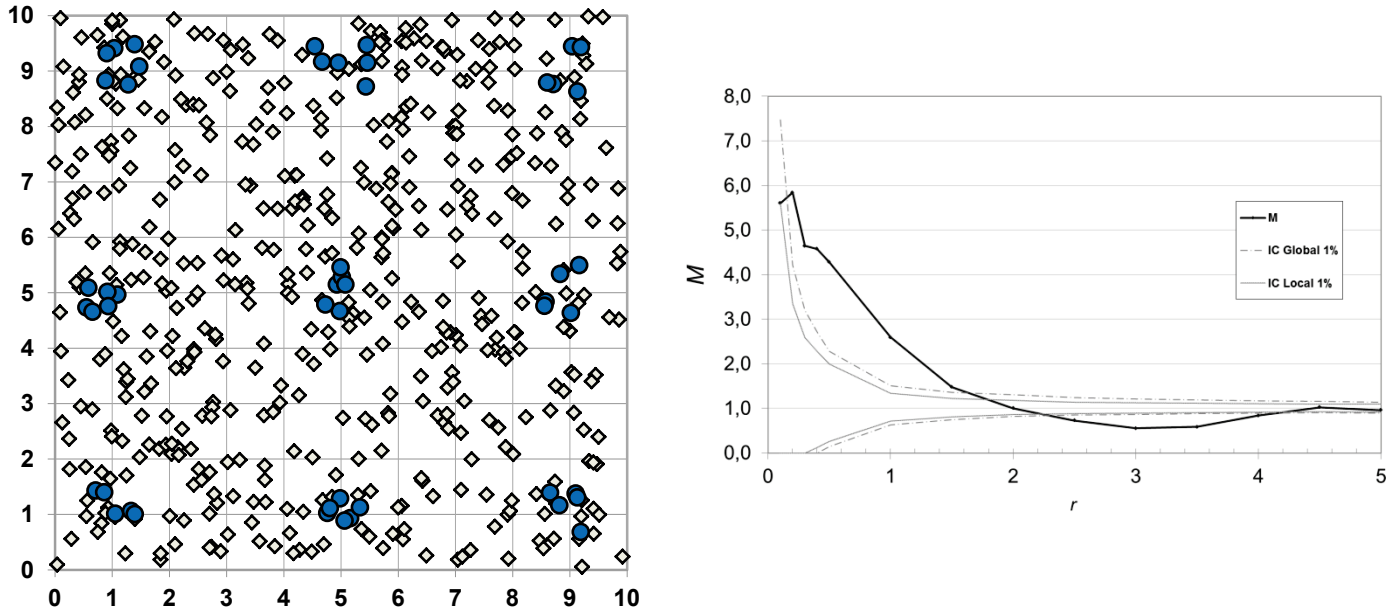


Figure 28 : Agrégats répartis régulièrement. Carte des points et fonction *M*.
Les contrôles (losanges) sont distribués complètement aléatoirement et les cas (cercles bleus) selon un processus de Matérn

Hétérogénéité

La Figure 29 présente une situation hétérogène, avec un gradient d'intensité de gauche à droite du domaine d'étude. Les contrôles comme les cas ont une intensité proportionnelle à $\ln(x+1)$, représentative d'un gradient de fertilité par exemple. Les cas sont distribués selon un processus de Poisson. Les cas sont tirés dans un processus de Thomas hétérogène. Le code pour R est le suivant :

```
>require ("spatstat")
>#  Poisson
>lambda <- function(x,y) {3*log(x+1)}
>pp <- rpoispp(lambda, win=owin(c(0,10), c(0,10)))
```

```
>#  Thomas
>kappa <- function(x,y) {(x+1)/10}
>mu <- function(x,y) {log(x+1)}
>ns=rThomas(kappa,.5,mu,owin(c(0,10), c(0,10)))
>#  Fusion
>total<-superimpose(controls=pp,cases=ns)
>plot(total)
```

Les agrégats sont très peu marqués. Ils sont détectés au seuil de 1% : le premier pic ($r = 0{,}6$) correspond à l'écart-type et le deuxième, plus discret, au double.

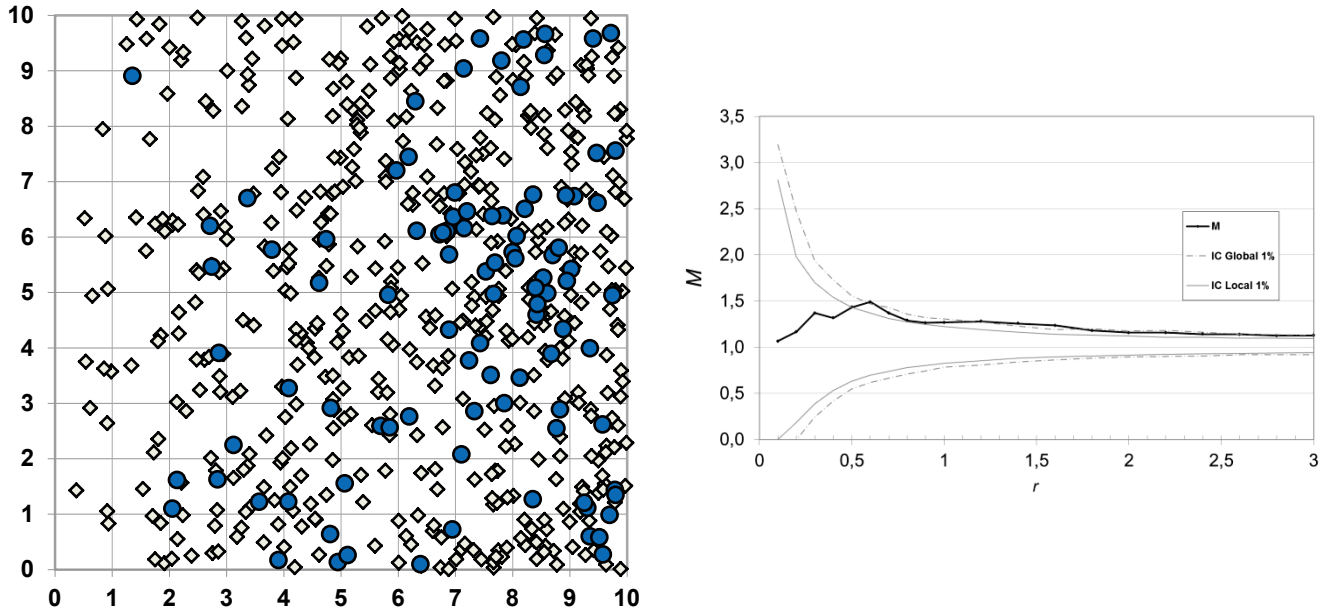


Figure 29 : Simulation d'un processus de Thomas hétérogène sur un Poisson hétérogène. Carte des points et fonction *M*.
Les contrôles (losanges) sont tirés dans un Poisson hétérogène d'intensité $3\ln(x+1)$, donnant un gradient de gauche à droite. Les cas (cercles) suivent un processus de Matérn (intensité des centres $(x+1)/10$, intensité des points dans les agrégats $0{,}5\ln(x+1)$). Il y a plus d'agrégats quand on progresse vers la droite, et de plus de points vers la droite des agrégats.

Cas extrêmes

Les indices de concentration prenant en compte le poids des individus doivent faire face à une difficulté particulière. L'analyse porte sur la structure des individus (les arbres dans notre cadre d'étude) alors que les calculs prennent en compte leur poids (la surface terrière par exemple). Si l'hypothèse nulle est l'indépendance entre les arbres, elle n'est pas identique à l'indépendance de la distribution des surfaces terrières. Ellison et Glaeser (1997) montrent que l'espérance, sous l'hypothèse de l'indépendance des établissements industriels, de leur indice de concentration géographique G (en termes d'employés) est proportionnelle à la concentration industrielle mesurée par l'indice d'Herfindahl (voir page 191), en d'autres termes que la distribution des employés sera d'autant moins homogène que les firmes seront de taille différente, même si aucune force

d'agglomération n'entre en jeu. Le contrôle de la distribution des poids des individus (la taille des établissements industriels d'Ellison et Glaeser ou la surface terrière des arbres), c'est-à-dire la prise en compte de ses effets pour ne pas attribuer à tort une concentration apparente des phénomènes d'agglomération des établissements ou des arbres alors qu'il ne s'agit que d'un artifice (toute la surface terrière d'un arbre est concentrée en un seul point) fait partie des critères de définition d'un bon indice de concentration spatiale, tels que définis par les économistes (Combes et Overman, 2004 ; Duranton et Overman, 2005).

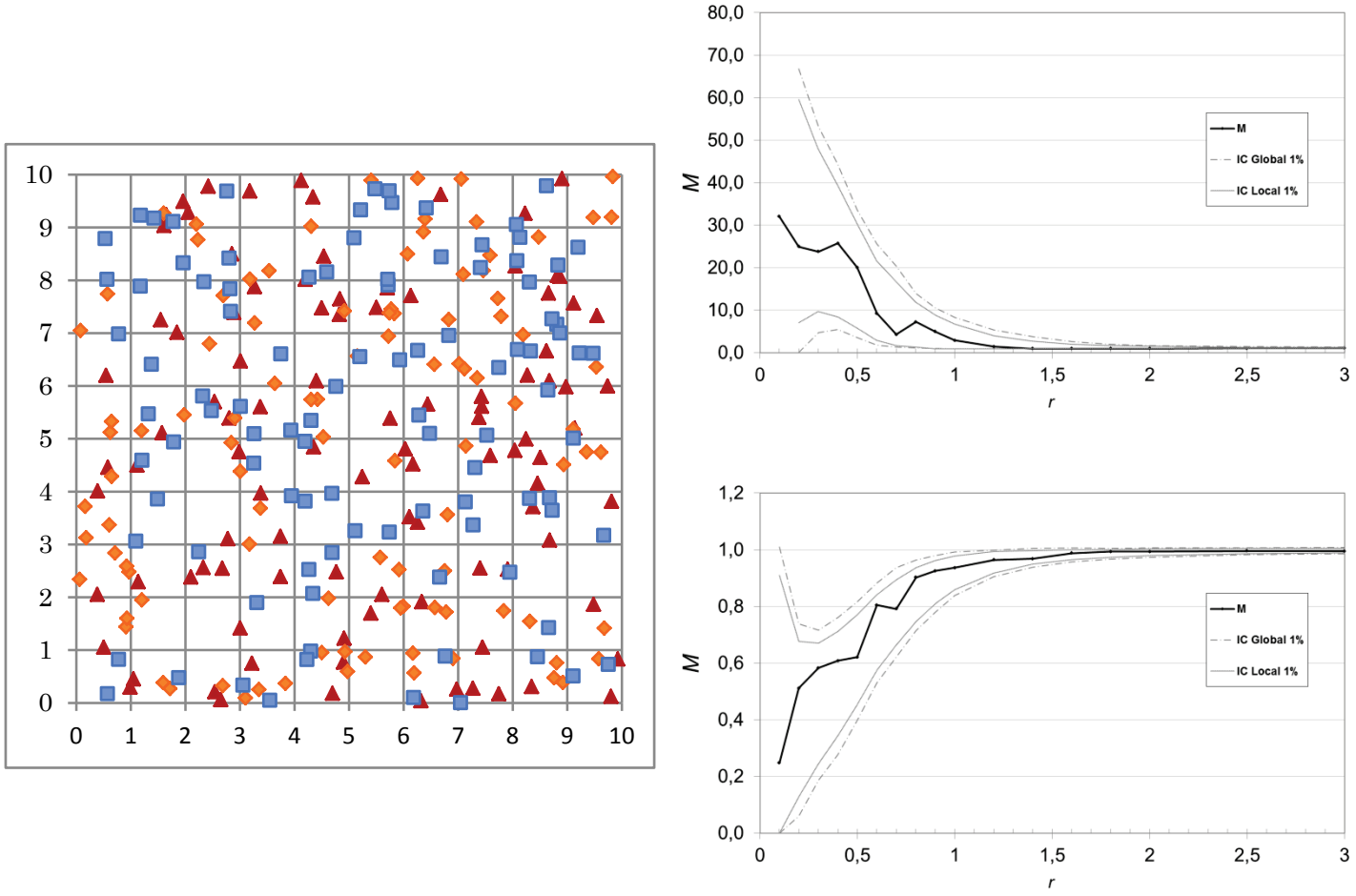


Figure 30 : Distribution complètement aléatoire de trois types de points de poids différent. Courbes *M* des points de poids 1 (en haut) et de poids 100 (en bas).

L'analyse des processus ponctuels s'applique par définition à des points, non pondérés, et la prise en compte de la taille des individus consiste implicitement à remplacer un point par un nombre de points correspondant à son poids. Si l'hypothèse nulle d'une distribution complètement aléatoire s'applique aux mètres carrés de surface terrière, elle ne correspond pas au problème généralement traité de l'indépendance de la localisation des arbres. Il ne semble pas possible de prendre en compte cette non-indépendance des points explicitement en compte dans le calcul de la fonction, mais la solution consiste, comme l'ont fait

avant nous Duranton et Overman (2005), à générer des jeux de points ayant la même pondération que les individus réels, et à calculer l'intervalle de confiance de l'hypothèse nulle par la méthode de Monte-Carlo à partir de ces points. L'intervalle de confiance correspond alors bien à l'hypothèse d'une distribution indépendante d'*arbres*, eux-mêmes présentant une certaine concentration de la surface terrière. Dans certains cas extrêmes, la valeur de référence de la fonction M, correspondant à l'indépendance des arbres, peut être différente de 1. On pourra l'estimer en retenant la valeur médiane des simulations. La Figure 30 présente un semis complètement aléatoire de 3 groupes de 100 points représentant des arbres de poids différents : 1 (triangles), 10 (losanges) et 100 (carrés) et les courbes de M pour les plus petits et les plus gros arbres. Les intervalles de confiance sont calculés au seuil de 1% à partir de 10 000 simulations.

Dans cet exemple, l'intervalle de confiance à très courte distance se situe entièrement au-dessus de 1 pour les petits arbres, au-dessous pour les plus gros. Les intervalles se recentrent rapidement sur la valeur 1 habituelle. L'explication est la suivante pour les gros arbres (elle est symétrique pour les petits) :

- Les points étant distribués de façon totalement aléatoire, les voisins rencontrés autour de chaque point dans un cercle de rayon r constituent une population équilibrée de petits, moyens et gros points. Notons a la proportion de la surface totale couverte par le cercle. La valeur individuelle de M attendue autour de chaque point est $\frac{a\times 99\times 100}{a\times(99\times 100+100\times(10+1))}\Big/\frac{99\times 100}{99\times 100+100\times(10+1)}=1$.
  On s'attend en effet à trouver dans toute la zone 99 points de poids 100 (100 points moins le centre du cercle) et 100 points de poids 1 et de poids 10. Parmi les voisins, on s'attend à trouver une proportion a de la population totale.
- A une distance, suffisamment petite, on ne trouvera que 2 points, ce qui exclut la possibilité d'en rencontrer un de chaque type. Les cas où un point de poids 100 manque seront à peu près aussi fréquents que ceux où un point de poids 1 ou de poids 10 manque. La valeur de M sera une moyenne des trois cas qui peut être calculée simplement : $\frac{1}{3}\left(\frac{0}{1+10}+\frac{100}{1+100}+\frac{100}{10+100}\right)\Big/\frac{99\times 100}{99\times 100+100\times(10+1)}\approx 0{,}7$. La valeur de M obtenue pour cette distance sera de l'ordre de 0,7 en absence de toute force de répulsion mais seulement parce que le poids moyen des voisins est surtout influencé par le cas où le point de poids 100 se situe hors du cercle.
- Intuitivement plutôt que mathématiquement, on peut comprendre que si les surfaces terrières étaient distribués indépendamment, les 100 unités d'un gros arbre seraient répartis de façon homogène sur le domaine. Vus d'un point de référence, quelques voisins se trouveraient à l'intérieur du cercle de voisinage. Comme ce n'est pas le cas, tous ces voisins sont superposés à l'emplacement de l'arbre, le plus souvent hors du cercle quand le

rayon est petit, comme si une force de répulsion agissait, d'où une valeur de *M* observée inférieure à 1.

Ces effets ne peuvent se produire que si tous les arbres d'un type étudié sont de la même taille, différente de la taille des autres. Dans le cas contraire, les effets de l'absence d'un point se compensent entre les cas et les contrôles. Enfin, ils ne sont sensibles qu'à très faible distance, tant que le nombre de voisins est inférieur au nombre de types. Si ces deux conditions ne sont pas remplies, la valeur médiane de l'intervalle de confiance de l'hypothèse nulle vaut approximativement 1. C'est le cas de la grande majorité des applications empiriques, à l'exception de celles visant à étudier justement la répartition spatiale des individus selon leur taille.

L'enseignement à tirer de cet exemple particulier est que la fonction *M* mesure numériquement la concentration ou la dispersion des surfaces terrières. Dans certains cas, la structure spatiale peut être due à la structure diamétrique. Si on s'intéresse, comme c'est généralement le cas, à la structure spatiale des arbres, l'intervalle de confiance de l'hypothèse nulle permet de détecter des forces d'agglomération ou de dispersion qui peuvent agir sur *M* même si leurs effets sont un effet net inférieur à celui de la structure diamétrique.

Les fonctions intertypes

Les fonctions intertypes permettent de détecter les interactions spatiales entre plusieurs (généralement deux) types de points. Les fonctions *K* et *M* intertypes sont des extensions des fonctions classiques. Les fonctions intertypes à marques continues permettent quant à elles de détecter l'autocorrélation spatiale de ces marques.

Les fonctions intertypes à marques discrètes

Les points peuvent être équipés de marques pour les reconnaître. Les marques discrètes permettent de définir des types de points (par exemple l'espèce des arbres) et les marques continues de noter une mesure (par exemple le diamètre des arbres). La fonction *L* a été rapidement utilisée (Diggle, 1983) pour évaluer la structure spatiale de semis de points de différentes marques. On s'appuie sur le même raisonnement, consistant à compter le nombre de voisins de chaque point, mais on s'intéresse maintenant aux voisins du type *2* autour des points du type *1*.

La fonction $K_{1,2}(r)$

Comme la fonction K originale, la fonction intertype est définie l'intégrale de la fonction g (107). La différence est que les points de référence (les centres des cercles) et les voisins ne sont pas du même type.

La fonction intertype de la structure spatiale des points de type *2* autour des points de type *1* est estimée, dans sa forme basique sans correction des effets de bord, par :

$$\widehat{K}_{1,2}(r) = \frac{1}{\widehat{\lambda_2}(A) n_1(A)} \sum_{i=1}^{n_1(A)} \sum_{j=1}^{n_2(A)} \mathbf{1}\left(\left\|x_i^1 - x_j^2\right\| \leq r\right) \quad \textbf{(53)}$$

(53) : Estimateur de la fonction intertype *K* de Ripley

$\widehat{K}_{1,2}(r)$ est égal au nombre moyen de voisins de type *2* autour de chaque point de type *1*, divisé par l'intensité des points de type *2*. La fonction intertype n'est pas concernée par le problème du biais traité page 50.

La fonction $L_{1,2}(r)$ est définie de la même façon que la fonction L originale :

$$L_{1,2}(r) = \sqrt{\frac{K_{1,2}(r)}{\pi}} - r \quad \textbf{(54)}$$

Correction des effets de bord

La correction des effets de bord est identique à celle de la fonction K non marquée. On retiendra la version suivante, avec la correction de Besag :

$$\widehat{K}_{1,2}(r) = \frac{1}{\widehat{\lambda_2}(A) n_1(A)} \sum_{i=1}^{n_1(A)} \frac{\pi r^2}{A_{ir}} \sum_{j=1}^{n_2(A)} \mathbf{1}\left(\left\|x_i^1 - x_j^2\right\| \leq r\right) \quad \textbf{(55)}$$

(55) : Fonction intertype *K* de Ripley après correction des effets de bord

Commutativité

On peut remarquer que $\widehat{\lambda_2}(A) = \frac{n_2(A)}{A}$. Il suffit de réarranger l'écriture de l'équation (51) pour montrer que $\widehat{K}_{1,2}(r) = \widehat{K}_{2,1}(r)$ en absence d'effets de bord.

La fonction intertype de Ripley mesure donc la co-agglomération des points de types *1* et *2*, sans donner de valeur préférentielle à l'une des deux marques.

Ce résultat n'est plus valide dès qu'on prend en compte la correction des effets de bord : la correction appliquée au point x_i^1 n'est pas égale à celle appliquée au point x_j^2, les deux estimateurs $\widehat{K}_{1,2}(r)$ et $\widehat{K}_{2,1}(r)$ sont donc légèrement différents. Diggle (1983) propose d'utiliser la moyenne des deux estimateurs. Lotwick et Silverman (1982) calculent la pondération optimale pour obtenir l'estimateur le plus efficace (ayant la plus faible variance) et concluent que, pour $n_1(A)$ et $n_2(A)$ suffisamment grands, le meilleur estimateur est : $\left(n_1(A)\widehat{K}_{2,1}(r) + n_2(A)\widehat{K}_{1,2}(r)\right)/\left(n_1(A) + n_2(A)\right)$.

Goreaud (2000) considère que la différence est suffisamment faible pour faire l'économie de calculs supplémentaires et se contente d'utiliser $\widehat{K}_{1,2}(r)$.

Hypothèse nulle

Les fonctions intertypes sont plus délicates à interpréter que la fonction de Ripley classique. $\hat{L}_{1,2}(r) > 0$ signifie que les points de type *1* et de type *2* sont plus agrégés les uns autour des autres que dans le cadre d'une distribution complètement aléatoire, mais cette agrégation peut être due aux relations entre les deux populations ou bien à la structure propre de chacune. L'hypothèse nulle de deux distributions complètement aléatoires ne permet pas de distinguer les deux cas.

Diggle (1983) propose deux hypothèses nulles, à choisir selon le problème à traiter :

- L'étiquetage aléatoire : on considère la répartition spatiale des points comme donnée, et on attribue aléatoirement les marques.
- L'indépendance des populations : on considère la répartition spatiale de chacune des populations comme donnée, et on change seulement leur position relative par une translation aléatoire. On considère la zone d'étude comme un tore : les points sortant de la zone d'étude après translation sont réintroduits à l'autre extrémité.

La question du choix de la bonne hypothèse nulle en fonction du problème est traitée en détail par Goreaud et Pélissier (2003). Ils montrent notamment qu'un mauvais choix peut aboutir à des contresens sur la structure détectée.

L'hypothèse nulle de l'indépendance des populations s'applique quand on cherche à détecter les relations entre deux populations *a priori* distinctes. Un exemple est la localisation relative de deux espèces dont on recherche les influences mutuelles. L'hypothèse de l'étiquetage aléatoire s'applique au contraire à des semis de points initialement unique dont on cherche à caractériser la répartition spatiale d'une transformation *a posteriori*, comme la mortalité des arbres.

L'hypothèse de l'indépendance des populations impose de générer des jeux de données aléatoires dans lesquels la structure propre chaque semis est conservée :

en pratique, on applique une translation aléatoire à l'une des deux populations. Si la fonction intertype sort de l'intervalle de confiance de cette hypothèse nulle, on peut affirmer que la cause est bien la relation entre les populations. Elle n'est applicable que si la forme de la zone d'étude permet de la traiter comme un tore. L'hypothèse de l'étiquetage aléatoire amène à conserver l'emplacement des points réels et à leur attribuer une étiquette aléatoire dans le respect des proportions d'origine des deux populations.

La fonction $M_{1,2}(r)$

La fonction $\widehat{M}_{1,2}(r)$ constitue une généralisation de la fonction intertype de Ripley similaire à celle présentée plus haut. L'équation (52) est modifiée pour prendre en compte le nombre de voisins de type *2* autour des points de type *1*. Le résultat est immédiat :

$$\widehat{M}_{1,2}(r) = \sum_{i=1}^{n_1(A)} \frac{\sum_{j=1}^{n_2(A)} \mathbf{1}\left(\left\|x_i^1 - x_j^2\right\| \le r\right) w\left(x_j^2\right)}{\sum_{j=1}^{n(A)} \mathbf{1}\left(\left\|x_i^1 - x_j\right\| \le r\right) w\left(x_j\right)} \Bigg/ \sum_{i=1}^{n_1(A)} \frac{W_2}{W - w\left(x_i^1\right)} \qquad \textbf{(56)}$$

(56) : Estimateur de la fonction intertype *M*

Les points x_i^1 appartiennent à l'espèce *1*, les points x_j^2 à l'espèce *2*. On calcule donc pour chaque point de l'espèce *1* le poids relatif des voisins de l'espèce *2* à la distance r (rapport des sommes au numérateur dans l'équation (52)). On rapporte cette moyenne au poids relatif de l'espèce *2* sur l'ensemble de la zone d'étude (dénominateur).

Cette fonction intertype présente les mêmes difficultés d'interprétation que la fonction intertype de Ripley : les deux hypothèses nulles doivent pouvoir être traitées, et, sous l'hypothèse d'indépendance des populations, la valeur de $M_{1,2}$ peut être due aux relations entre les deux semis de points (ce qui est son objectif) mais aussi à la structure propre de chacun. La technique consistant à translater un des semis de points par rapport à l'autre n'est pas utilisable ici, il s'agit donc de trouver une autre solution.

Test de l'hypothèse nulle d'étiquetage aléatoire

Le test ne présente aucune difficulté : les emplacements des points sont conservés, ainsi que leurs poids, et les types de points sont redistribués aléatoirement sur les points existants.

Test de l'hypothèse nulle d'indépendance des populations

Le test est nettement plus complexe : il s'agit de conserver la structure des deux types de points, mais seules les positions réelles de points peuvent être utilisées pour les simulations. La solution consiste à faire le test en deux temps.

Prise en compte de la structure du semis de points de type 1

La mesure de l'ensemble des localisations possibles des voisins des points de type *1* est la somme des poids des voisins de toutes les espèces. Si le semis *1* a une structure très marquée, disons très concentrée, le poids relatif des voisins de l'espèce *2* sera diminué, non par un déficit de points de type *2* mais par un excès de points de type *1* (voir la Figure 32, courbe $\widehat{M}_{2,1}(r)$ en haut à droite).

La façon la plus simple d'éliminer l'effet de la structure de *1* consiste à utiliser comme hypothèse nulle la disposition aléatoire des points des autres types étant donnée celle de *1*. En d'autres termes, on conservera les points de type *1* et on permutera de façon aléatoire tous les autres points sur l'ensemble des emplacements existants (hormis ceux occupés par les points de type *1*) pour obtenir un jeu de données aléatoire de référence.

Dans notre exemple, la valeur de la fonction intertype sera inférieure à 1 mais dans l'intervalle de confiance de l'hypothèse nulle dont la borne supérieure sera elle-même éventuellement inférieure à 1.

Prise en compte de la structure du semis de points de type 2

On ne peut pas, comme pour la fonction intertype *K*, conserver la structure de *2* en même temps que la structure de *1* dans l'hypothèse nulle : l'espace n'étant pas homogène, la seule façon de conserver la structure d'un semis de point est de le laisser en place. Conserver les deux structures de *1* et *2* reviendrait donc à éliminer tout aléa.

Seule la fonction $\widehat{M}_{2,1}(r)$ permet donc de prendre en compte la structure de *2*. On retiendra donc qu'il existe une relation significative à la distance r entre les points de types *1* et *2* si :

- $\widehat{M}_{1,2}(r)$ diffère significativement de sa valeur sous l'hypothèse nulle de la distribution aléatoire de tous les points n'appartenant pas au type *1*, les points de type *1* étant fixes.
- $\widehat{M}_{2,1}(r)$ diffère significativement de sa valeur sous l'hypothèse nulle de la distribution aléatoire de tous les points n'appartenant pas au type *2*, les points de type *2* étant fixes.

M intertype n'est applicable que si le nombre de types de points est supérieur à 2.

Les deux premiers exemples à suivre montrent que la structure propre de chacun des types de points est bien contrôlée par les intervalles de confiance. Le premier est un semis agrégé indépendant de semis complètement aléatoires, qui perturbe peu les fonctions intertypes. Le deuxième est un semis régulier qui déforme une des fonctions intertypes.

Les exemples suivants sont issus de données réelles, l'implantation des magasins de commerces de détail dans la ville de Lyon, qui mettent en évidence des interactions entre types de points.

Exemples

Semis complètement aléatoire et agrégé, indépendants

Cet exemple théorique (Figure 31) comprend deux semis de 100 points complètement aléatoires, de types *1* et *3*, et un semis de type *2* généré par un processus de Matérn de 25 points répartis en 5 agrégats.

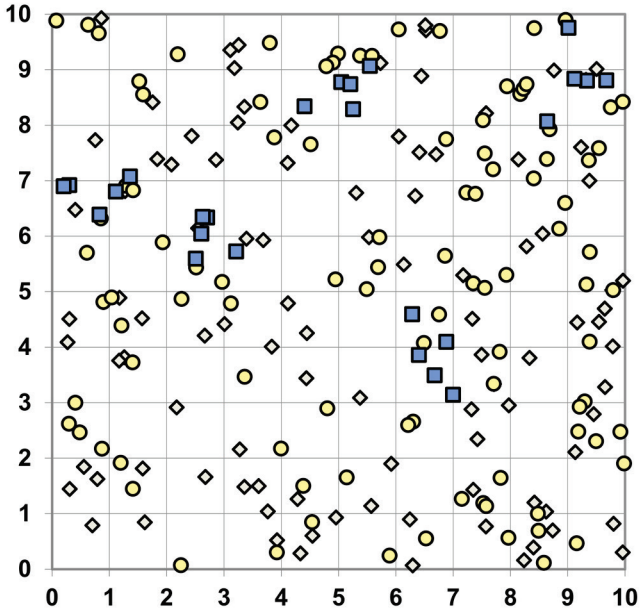


Figure 31 : Carte des points.
Deux semis de points poissonniens (losanges et cercles) et un semis de points agrégés (carrés), indépendants entre eux.

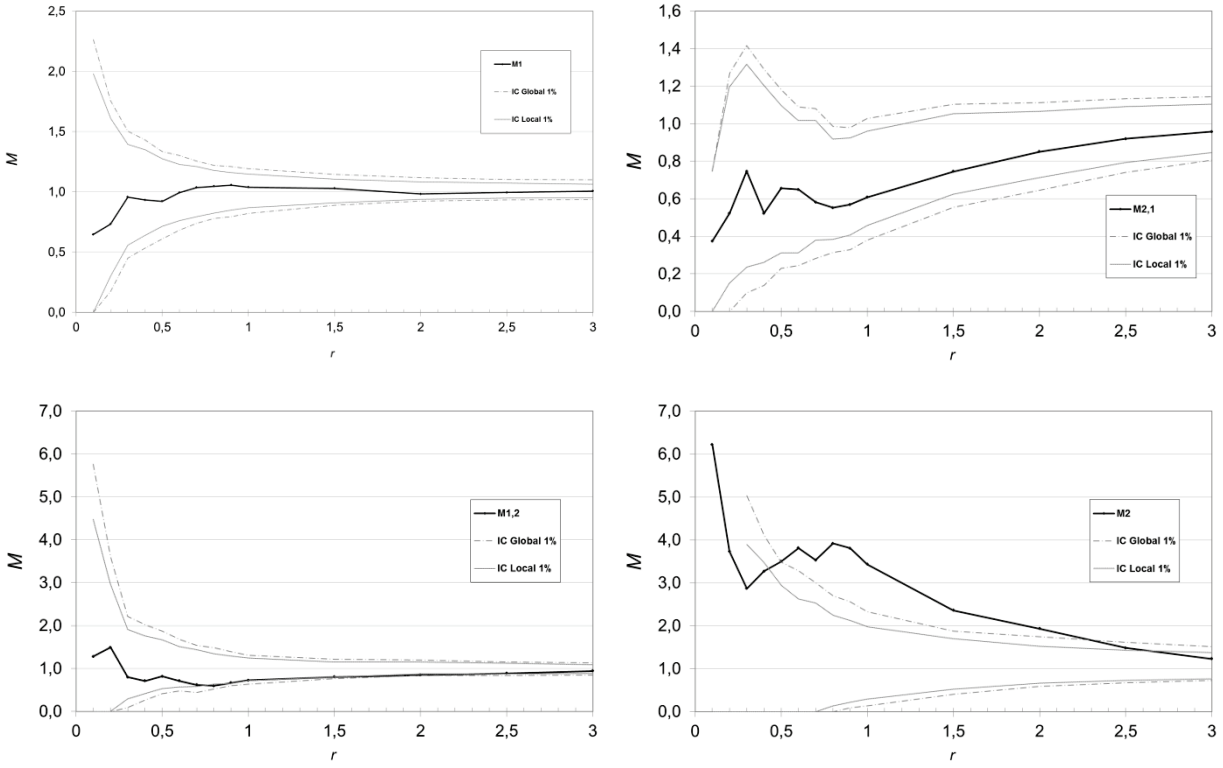


Figure 32 : Courbes M et M intertype pour un semis complètement aléatoire et un agrégé.
En haut à gauche, $\widehat{M}_1(r)$,pour les points de type *1*, distribués complètement aléatoirement. En bas à droite, $\widehat{M}_2(r)$ pour le semis de points agrégés. $\widehat{M}_{1,2}(r)$ et $\widehat{M}_{2,1}(r)$ sont les courbes intertypes, avec intervalles de confiance sous l'hypothèse nulle d'indépendance des populations.

Les courbes concernant le type *3* ne sont pas représentées puisque les types *1* et *3* ont la même structure.

On observe (Figure 32) que la seule courbe sortant de l'intervalle de confiance est $\widehat{M}_2(r)$, celle du semis agrégé. La courbe intertypes $\widehat{M}_{2,1}(r)$ n'est pas centrée sur la valeur 1 : il y a relativement moins de points du type *1*, complètement aléatoire, dans les agrégats du type *2*. L'intervalle de confiance de l'hypothèse nulle est déformé de façon identique.

Semis complètement aléatoire et régulier, indépendants

Cet exemple contient deux semis de 100 points complètement aléatoires, de types *1* et *3*, et un semis de type *2* régulier disposé sur une maille de 1x1.

La carte des points est en Figure 33.

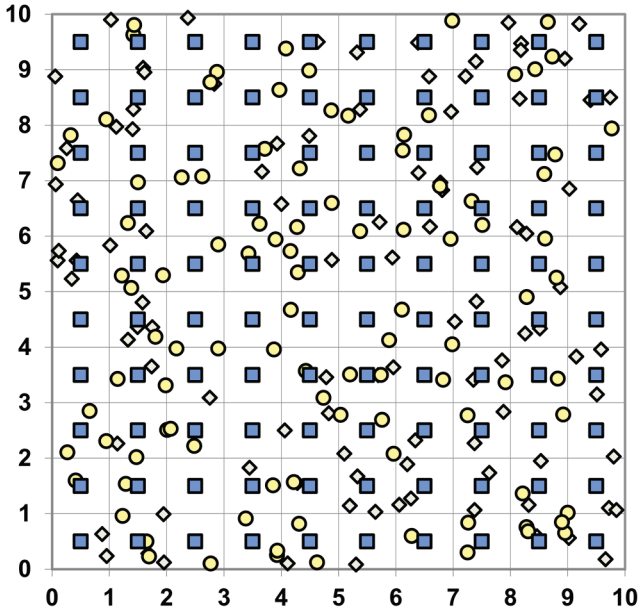


Figure 33 : Carte des points.
Deux semis de points poissonniens (losanges et cercles) et un semis de régulier (carrés), indépendants entre eux.

La seule courbe sortant de l'intervalle de confiance est $\widehat{M}_2(r)$, celle du semis régulier. Les courbes intertypes s'écartent sensiblement de 1, valeur correspondant à la répartition complètement aléatoire des deux semis, mais reste dans l'intervalle de confiance de l'hypothèse nulle, c'est-à-dire l'absence de relation entre les types *1* et *2*, $\widehat{M}_{1,2}(r)$ présente un déficit de voisins à petite distance puis un excédent à partir d'une distance correspondant à la demi-diagonale d'un carré de la grille du semis *2*, quand un point quelconque aura forcément plusieurs nœuds de la grille dans son voisinage. Symétriquement, la courbe $\widehat{M}_{2,1}(r)$ reste au-dessus de 1 à faible distance parce qu'aucun voisin des points de la grille n'appartient au type *2* (la proportion de voisins du type *1* est donc augmentée). A $r = 1$, la courbe présente une discontinuité (éliminée par le processus de lissage) correspondant à l'entrée des points de la grille les plus proches dans le cercle de voisinage. Les discontinuités suivantes, à $r = \sqrt{2}$ puis $r = 2$ sont masquées par le pas de calcul égal à 0,5 à partir de $r = 1$: comme on l'a vu plus haut, la régularité est difficile à détecter.

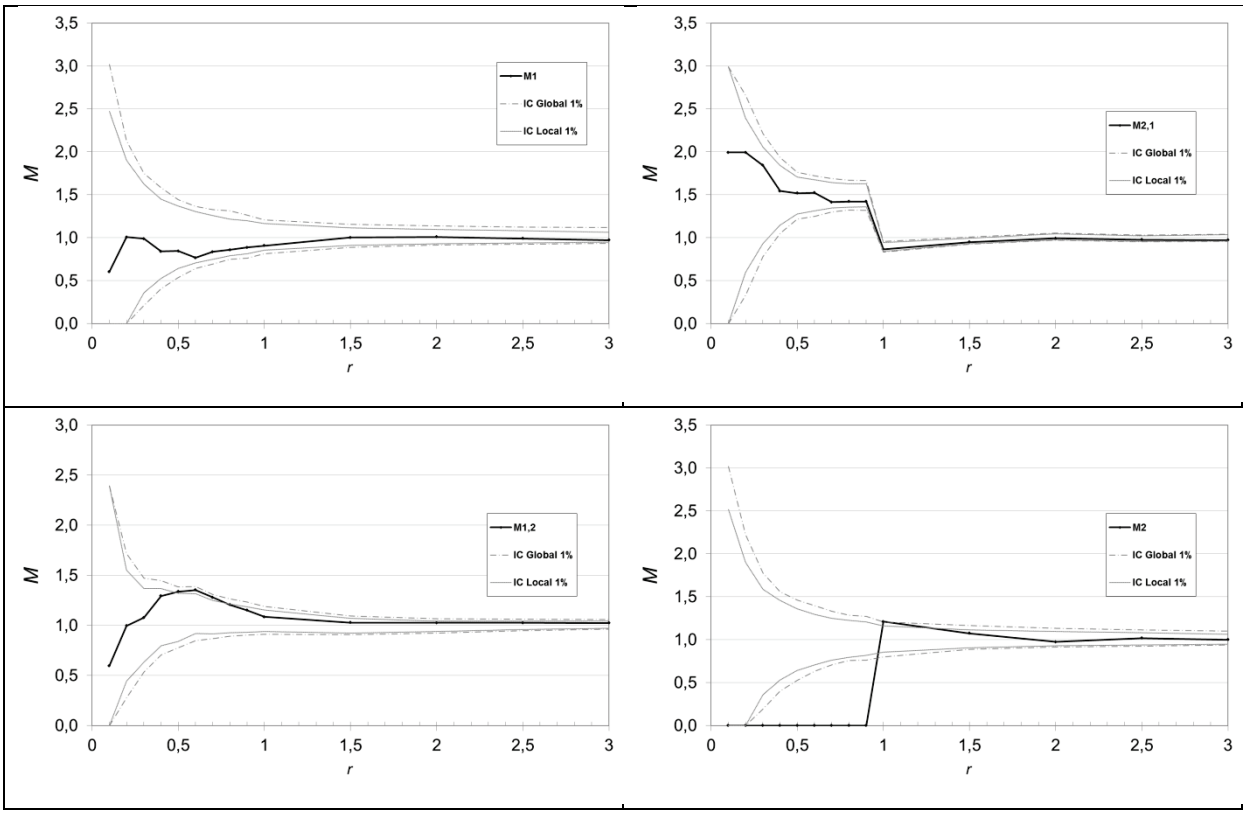


Figure 34 : Courbes M et M intertype pour un semis complètement aléatoire et un régulier. En haut à gauche, $\widehat{M}_1(r)$, pour les points de type *1*, distribués complètement aléatoirement. En bas à droite, $\widehat{M}_2(r)$ pour le semis de points réguliers. $\widehat{M}_{1,2}(r)$ et $\widehat{M}_{2,1}(r)$ sont les courbes intertypes, avec intervalles de confiance sous l'hypothèse nulle d'indépendance des populations.

Les commerces de détail de la ville de Lyon

Le jeu de données provient de la Chambre de Commerce de Lyon et contient les emplacements de tous les commerces de détail de l'aire urbaine, classés par secteur d'activité (l'année d'observation est 2005). Faute d'information suffisante sur la taille des établissements, un poids identique égal à 1 leur est attribué. La carte est en Figure 35. Les points cartographiés sont 220 pharmacies, 229 vendeurs de journaux et livres, 927 boutiques de vêtements et 2013 autres commerces. Le semis de points est très fortement hétérogène, avec une forte concentration des commerces en centre-ville et le long des artères commerçantes visibles sur la carte.

La théorie économique (Eaton et Lipsey, 1982) prévoit que les consommateurs cherchent à minimiser leurs déplacements pour les achats quotidiens et que les commerces correspondants ont intérêt à la co-agglomération pour fournir en un seul lieu l'ensemble des services. Elle est vérifiée sur la Figure 36, en haut à gauche. À cause de la concurrence, les commerces du même secteur ont tendance à se repousser (figure en bas à gauche). Les commerces de vêtement en revanche ont une stratégie différente et se localisent ailleurs (figure en haut à droite), ensemble (figure en bas à droite), pour former des agrégats de taille importante qui attirent les clients.

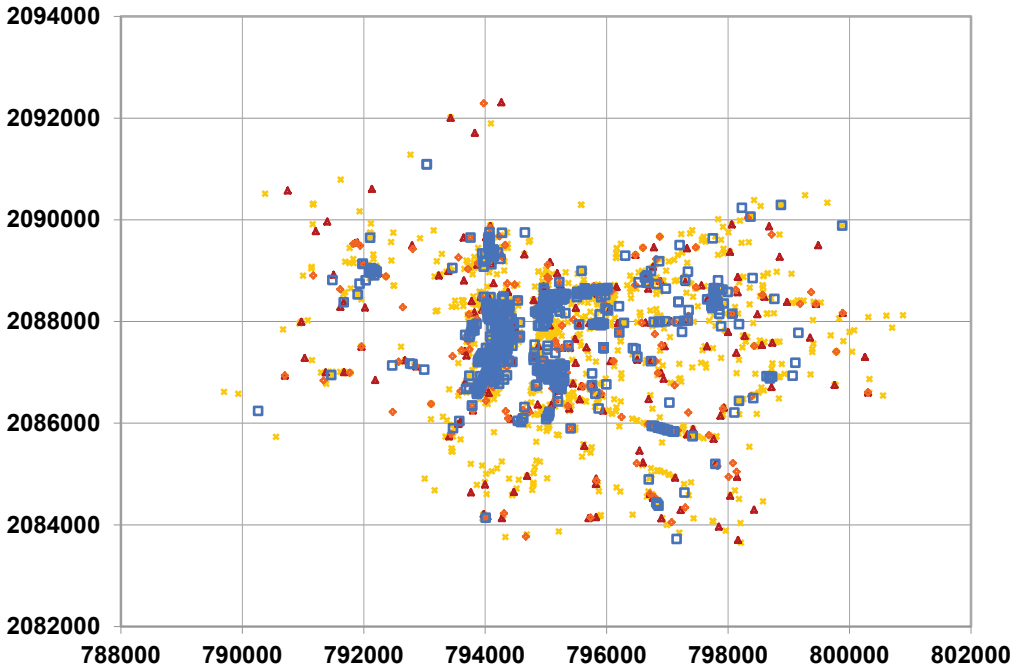


Figure 35 : Commerces de détail de la ville de Lyon. Carte des points.
Les pharmacies (triangles), les marchands de journaux (losanges) et de vêtements (carrés) sont détaillés, les autres commerces sont représentés par des croix. Les quadrillages ont deux kilomètres de côté.

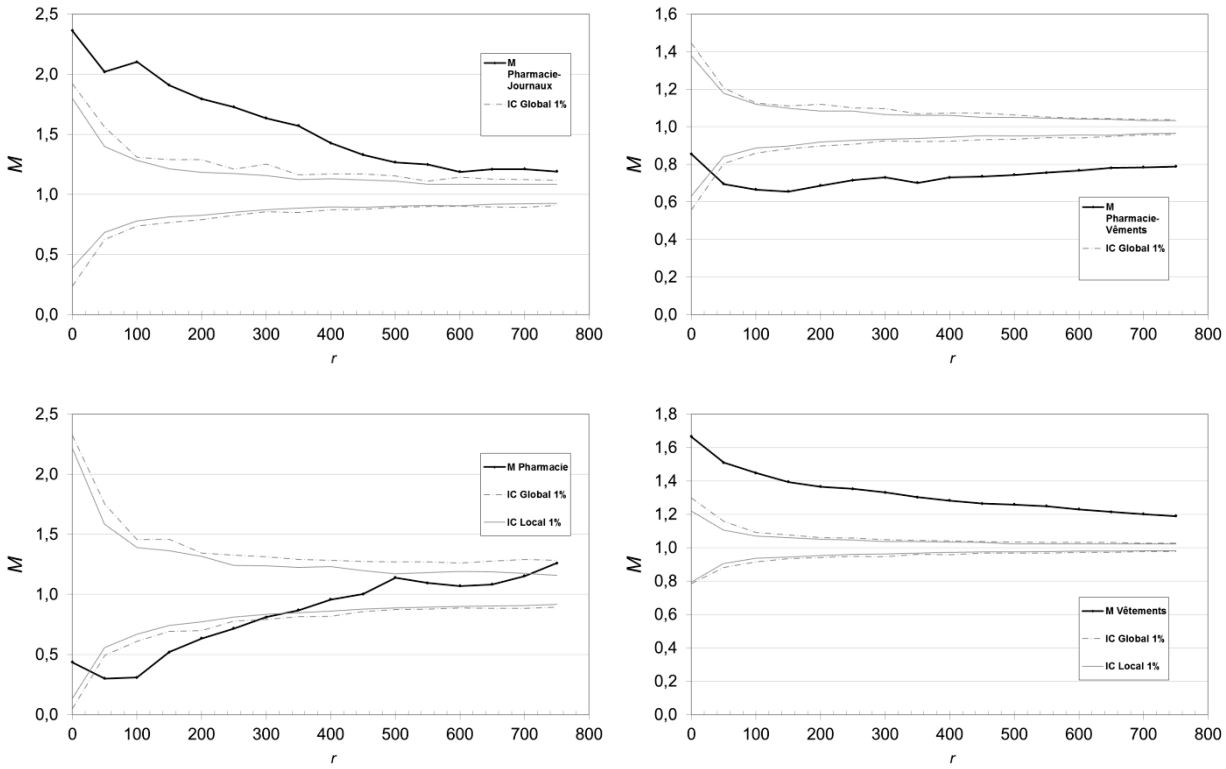


Figure 36 : Structure propre, co-agglomération et répulsion des commerces.
Les commerces des achats quotidiens s'attirent entre secteurs mais se repoussent dans le même secteur alors que les commerces de vêtements se localisent loin des premiers sous forme d'agrégats.

Les intervalles de confiance sont réalisés à partir de 1 000 simulations. L'hypothèse nulle est l'étiquetage aléatoire : les emplacements commerciaux pré-existent et les enseignes viennent les occuper. Les distances sont en mètres.

Version alternative de la fonction D

L'hypothèse nulle de l'étiquetage aléatoire utilisée pour la fonction D établit deux égalités (Diggle et Chetwynd, 1991 équation 3) :

$$K_c(r) = K_0(r) = K_{c,0}(r) \quad \textbf{(57)}$$

Diggle et Chetwynd ont choisi de définir la fonction D comme la différence entre les fonctions K des cas et des contrôles. Ils auraient pu s'intéresser à la différence $K_c(r) - K_{c,0}(r)$, également nulle dans le cadre de l'hypothèse nulle. Cette variante de la fonction D, que nous appellerons D_i puisqu'elle utilise comme référence la fonction K intertype, présente un avantage de taille : elle calcule les deux fonctions K autour des *mêmes* points, les cas, et échappe donc au problème illustré par la Figure 24. Son expression analytique est plus simple :

$$\begin{aligned} D_i(r) &= K_c(r) - K_{c,0}(r) \\ &= \frac{1}{n_c(A)} \sum_{i=1}^{n_c(A)} \left[\frac{1}{\widehat{\lambda_c}} \sum_{j=1, i\neq j}^{n_c(A)} \mathbf{1}\left(\|x_i^c - x_j^c\| \leq r\right) c(i,j,r) \right. \\ &\quad \left. - \frac{1}{\widehat{\lambda_0}} \sum_{j=1, i\neq j}^{n_0(A)} \mathbf{1}\left(\|x_i^0 - x_j^0\| \leq r\right) c(i,j,r) \right] = \frac{\overline{v_c}(i,r)}{\widehat{\lambda_c}} - \frac{\overline{v_0}(i,r)}{\widehat{\lambda_0}} \end{aligned} \quad \textbf{(58)}$$

D_i est égal à la différence moyenne, *autour de chaque cas*, des cas voisins rapportés à leur densité sur l'ensemble du domaine d'étude, et des contrôles voisins, normalisé de la même façon.

Unification des outils de caractérisation des processus ponctuels

L'objectif de cette conclusion du chapitre sur les statistiques spatiales continues est d'établir une classification des différentes fonctions dans un cadre commun applicable à g, K, g_{inhom}, K_{inhom}, K_d, O et M.

La première étape de la construction de la statistique consiste à compter les voisins d'un point, à la distance r ou jusqu'à la distance r, définissant selon le cas des fonctions de densité (g, g_{inhom}, K_d et O) ou des cumulatives (K, K_{inhom}, et M). Ces nombres de voisins sont ensuite comparés à une mesure de référence. Cette mesure est la surface du cercle (ou de la couronne) ou encore le nombre de voisins de tous types confondus. Selon la typologie fixée par Brülhart et Traeger (2005) :

- Les statistiques *topographiques* utilisent l'espace comme mesure de référence : le nombre de voisins est divisé par divisé par la surface du cercle ou cercle (ou de la couronne). g, K, g_{inhom}, K_{inhom}, , et O sont concernés.
- Les statistiques *relatives* comparent le nombre de voisins à un autre nombre de voisins : le nombre de cas est divisé par le nombre de voisins de tous types, cas et contrôles. C'est le cas de M.
- Les statistiques *absolues* n'ont pas de valeur de référence. K_d est une fonction absolue à ce stade.

Fonction	***Comptage autour de x_i***	***Observations***
$\widehat{\boldsymbol{K}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$ $\widehat{\boldsymbol{K}}_{\boldsymbol{inhom}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$	$v(x_i,r) = \sum_{j=1,i\neq j}^{n(A)} \frac{\mathbf{1}(\|x_i - x_j\| \leq r)c(i,j,r)}{\pi r^2 \hat{\lambda}(x_j)}$	Le nombre de voisins est corrigé des effets de bord et comparé au nombre de voisins attendus
$\widehat{\boldsymbol{g}}(\boldsymbol{r})$ $\widehat{\boldsymbol{g}}_{\boldsymbol{inhom}}(\boldsymbol{r})$ $\widehat{\boldsymbol{O}}(\boldsymbol{r})$	$v(x_i,r) = \sum_{j=1,i\neq j}^{n(A)} \frac{k(\|x_i - x_J\|,r)c(i,j,r)}{2\pi r \hat{\lambda}(x_j)}$	Idem, mais le nombre de voisins est estimé par un noyau.
$\boldsymbol{K_d(r)}$	$v(x_i,r) = \sum_{j=1,i\neq j}^{n(A)} k(\|x_i - x_J\|,r)$	Le nombre de voisins est estimé par un noyau, mais n'est comparé à rien.
$\widehat{\boldsymbol{M}}(\boldsymbol{r})$	$v(x_i,r) = \frac{\sum_{j=1,i\neq j}^{n_c(A)} \mathbf{1}(\|x_i^c - x_j^c\| \leq r)w(x_j^c)}{\sum_{j=1,i\neq j}^{n(A)} \mathbf{1}(\|x_i^c - x_j\| \leq r)w(x_j)}$	Le nombre de cas voisins est comparé au nombre de tous les voisins.

Tableau 2 : Estimation du nombre de voisins par les fonctions de mesure de la concentration spatiale.

La valeur obtenue autour de chaque point (Tableau 2) est ensuite moyennée sur l'ensemble des points (Tableau 3). Pour les fonctions géographiques, chaque point reçoit un poids inverse à l'intensité du processus autour de lui pour assurer un échantillonnage uniforme de l'espace. Chaque point reçoit le même poids dans K_d, et M.

Fonction	*Moyenne pour tous les x_i*	*Observations*
$\widehat{K}(r)/(\pi r^2)$ $\widehat{K}_{inhom}(r)/(\pi r^2)$ $\widehat{g}(r)$ $\widehat{g}_{inhom}(r)$ $\widehat{O}(r)$	$\bar{v}(r) = \sum_{i=1}^{n(A)} \frac{v(x_i, r)}{\hat{\lambda}(x_i)}$	La moyenne est inversement pondérée par l'intensité locale du processus pour assurer un échantillonnage homogène de l'espace.
$K_d(r)$	$\bar{v}(r) = \frac{1}{n(A)} \sum_{i=1}^{n(A)} v(x_i, r)$	La moyenne n'est pas pondérée.
$\widehat{M}(r)$	$\bar{v}(r) = \frac{1}{n_c(A)} \sum_{i=1}^{n_c(A)} v(x_i, r)$	La moyenne n'est pas pondérée.

Tableau 3 : Moyenne du nombre de voisins.

La dernière étape est la normalisation. La valeur attendue est 1 en cas d'indépendance des points pour $\widehat{K}(r)/(\pi r^2)$, $\widehat{K}_{inhom}(r)/(\pi r^2)$, $\hat{g}(r)$, $\hat{g}_{inhom}(r)$ et $\widehat{M}(r)$. Seul $\widehat{M}(r)$ nécessite une normalisation par le rapport du nombre de cas sur le nombre total de points. $\hat{O}(r)$ est multiplié par $n(A)/A$ pour atteindre sa valeur attendue égale à λ. Enfin, $K_d(r)$ est divisé par $n(A)(n(A)-1)$ pour que son intégrale soit égale à 1.

Les simulations de l'hypothèse nulle permettent de pallier les manques de la formulation analytique :

- K_d ne prend en compte aucune référence à la distribution totale de l'emploi (à laquelle la distribution d'un secteur économique est comparée), ni la concentration industrielle qu'elle cherche à contrôler. C'est la redistribution des points dans le cadre de l'hypothèse nulle qui permet de donner un sens aux résultats, pas la valeur de la statistique. La comparaison de K_d à son hypothèse nulle fournit un test de concentration relative pour des semis de points hétérogènes.
- M permet d'attribuer un poids aux points, mais ne fournit pas de correction pour ramener sa valeur de référence (en cas d'indépendance des points) à 1 en cas de forte structuration de la taille des points. C'est le rôle des simulations de l'hypothèse nulle, qui redistribue les points avec leurs poids d'origine.

- La statistique O ne prend pas en compte l'hétérogénéité du processus. Sa valeur peut être comparée à celles d'un processus hétérogène simulé, pour détecter la dépendance entre les points.
- Dans tous les cas, le modèle nul est un processus de Poisson. La valeur de g, celle de $K/(\pi r^2)$, celles de leurs versions inhomogènes, ou celle de M sont attendues à 1 dans ce cas. Pour rejeter l'hypothèse nulle d'un processus dépendant, comme un Strauss, les simulations sont le bon outil.
- Si l'on cherche à caractériser les propriétés du processus, et notamment la fonction g, la fonction M surpondère les points des zones denses. Si l'on cherche en revanche à caractériser le comportement des individus (choix de localisation,…), l'absence de pondération est plus appropriée.

La littérature statistique traite principalement des processus ponctuels homogènes, qui traitent le cas topographique. Les méthodes permettant de traiter des processus inhomogènes sont encore en cours de développement : l'estimation de l'intensité du processus en tout point ne pose pas de difficulté technique mais des problèmes méthodologiques.

La mesure de la concentration relative est nécessaire dans la plupart des applications économiques. C'est ce qui a motivé le développement de nouveaux outils, dont K_d et M. L'application de M en écologie est moins immédiate qu'en économie. En économie géographique, la question à traiter typiquement est celle de la structuration spatiale de la distribution des emplois, supposée agrégative en raison d'externalités positives (Marshall, 1890 ; Weber, 1909 ; Krugman, 1991). La distribution totale de l'emploi est très hétérogène, les indices de concentration géographique sont peu pertinents (Marcon et Puech, 2003), mais les indices relatifs à cette distribution totale permettent de détecter l'agrégation ou la répulsion de certains secteurs économiques (Combes et Overman, 2004 ; Duranton et Overman, 2005 ; Marcon et Puech, 2010). Si on s'intéresse à la structure spatiale d'un peuplement forestier, utiliser une fonction relative revient à caractériser la structure spatiale d'une espèce en contrôlant l'hétérogénéité par la distribution de l'ensemble des espèces. Cette méthode a un sens si l'espèce en question a une distribution possible équivalente à celle de l'ensemble des autres. Si ce n'est pas le cas, le semis de points de contrôle doit être adapté : par exemple, l'ensemble des espèces pionnières pour caractériser la structure d'une espèce pionnière. Cette démarche est en pratique assez difficile et les résultats peuvent être aussi discutables que ceux obtenus en assumant l'homogénéité du processus.

Les connaissances mathématiques sur les processus homogènes sont très avancées. De nombreux résultats théoriques existent, les propriétés de g et de K sont bien connues, y compris leur variance et même les valeurs critiques de K établies plus haut pour certains cas simples. La formulation analytique de leur espérance existe pour plusieurs processus classiques : Poisson homogène, mais aussi certains processus non indépendants (Illian et al., 2008). K_d a été en revanche cons-

truits sur une base empirique. Les valeurs de K_d calculées à partir des données sont d'ailleurs les indices eux-mêmes, et non les estimateurs de fonctions venant de la théorie sur les processus ponctuels. Le cas de M est intermédiaire.

Enfin, il existe un débat sur l'intérêt comparé des fonctions de densité et des cumulatives (Wiegand et Moloney, 2004 ; Law *et al.*, 2009 ; Marcon et Puech, 2010) qui selon les circonstances, peuvent donner des résultats plus clairs.

Au final, le praticien dispose d'une boîte à outils assez complète, mise à jour régulièrement dans la littérature écologique (Fortin et Dale, 2005 ; Illian *et al.*, 2008 ; Law *et al.*, 2009) qui ignore cependant encore largement les indices relatifs. Une approche pratique de la caractérisation des structures spatiales est présentée en conclusion générale de ce travail (page 109).

STATISTIQUE SPATIALE DISCRÈTE - ENTROPIE

On a commencé à étudier les structures spatiales longtemps avant Ripley. Beaucoup de mesures de la structuration spatiale s'appuient par exemple sur la variance du nombre de points comptés dans des quadrats (voir page 185), à la suite de Clapham (1936). L'analyse d'un semis de points en tant que tel, sans regroupement par quadrat, est plus récente (Clark et Evans, 1954). Les bases théoriques des mesures étaient souvent moins claires que celles de Ripley, mais l'homogénéité était presque toujours sous-entendue. La référence était aussi en général une distribution indépendante, mais ce n'est pas toujours le cas, y compris pour des indices d'importance majeure dans la littérature, comme celui de Gini (page 188).

L'approche utilisée ici est celle de la mesure de l'entropie, qui fournit un cadre clair et présente des propriétés intéressantes. L'entropie est la mesure du désordre d'un système, par exemple le nombre de ses états possibles.

Trois littératures parallèles ont développé les concepts et les outils de l'entropie : la physique statistique, la théorie de l'information avec les travaux de Shannon (1948) et la branche de l'économie qui s'intéresse à l'inégalité des revenus avec l'ouvrage de Theil (1967). La mesure de la diversité est une question centrale en écologie dont l'indice de Shannon est devenu un outil standard. L'indice de Simpson, développé dans un cadre différent, est aussi une mesure d'entropie. En économie, l'indice de Theil s'est répandu et a été utilisé pour mesurer la concentration absolue (Hart, 1971, l'applique à la concentration industrielle pour conclure que les mesures classiques sont plus pertinentes dans ce contexte) puis la concentration géographique (Brülhart et Traeger, 2005). L'objectif ici est d'unifier les approches des écologistes et des économistes pour produire un outil de mesure simultanée de la diversité et de la structure spatiale, dont une propriété essentielle est d'être décomposable.

Les résultats majeurs sont la production d'une statistique mesurant l'écart entre une distribution observée et une distribution attendue, qui peut caractériser aussi bien la concentration spatiale que la diversité. Ses propriétés de décomposabilité, appliquées au cas particulier de la biodiversité, permettent de formuler analytiquement la diversité β de Shannon. Enfin, les valeurs empiriques peuvent être testées contre l'hypothèse nulle d'égalité des distributions.

Les outils concurrents et complémentaires de mesure de la concentration spatiale sont présentés en Annexe 3 : Méthodes alternatives en statistiques spatiales discrètes. Le champ d'application et les limites de chacun sont discutés et une synthèse est proposée.

Les indices de diversité en écologie

L'indice de Shannon

La construction de l'indice de Shannon (Shannon, 1948 ; Shannon et Weaver, 1963) est fondée sur la quantité d'information. Considérons une placette forestière contenant S espèces végétales différentes. La probabilité qu'une plante choisie au hasard appartienne à l'espèce s est notée p_s. On prélève n plantes, et on enregistre la liste ordonnée des espèces des n plantes. Si n est suffisamment grand, le nombre de plantes de l'espèce i est np_s. On note L le nombre de listes respectant ces conditions :

$$L = \frac{n!}{\prod_{s=1}^{S}(np_s)!} \qquad \textbf{(59)}$$

Démonstration :

Le nombre de positions possibles dans la liste pour les individus de la première espèce est $C_n^{np_1}$. Le nombre de positions pour la deuxième espèce est $C_{n-np_1}^{np_2}$. Pour la i^{ème} espèce, le nombre est $C_{n-np_1-\cdots-np_{i-1}}^{np_i}$

Les produits de combinaisons se simplifient pour donner l'équation (59).

On peut maintenant écrire le logarithme de L : $\ln L = \ln n! - \sum_{s=1}^{S} \ln np_s$. On utilise l'approximation de Stirling, $\ln n! \approx n \ln n - n$, pour obtenir après simplifications :

$$\ln L = -n \sum_{s=1}^{S} p_s \ln p_s \qquad \textbf{(60)}$$

Il est possible d'obtenir des listes de plantes ne respectant pas les probabilités individuelles, mais comme on suppose n assez grand, leur probabilité d'occurrence est faible (loi des grands nombres) et elles peuvent être négligées.

$H = \ln L/n$ est l'indice de Shannon. À l'origine, Shannon a utilisé un logarithme de base 2 pour que H soit le nombre moyen de questions binaires (réponse oui ou non) nécessaire pour identifier l'espèce d'une plante.

L'indice est compris entre 0, cas où une seule espèce est de probabilité non nulle, et $\ln S$, le logarithme du nombre d'espèces si elles toutes équifréquentes.

L'indice de Simpson

On note p_s la fréquence de l'espèce s. L'indice de Simpson (1949), ou Gini-Simpson, est :

$$E_s = 1 - \sum_{s=1}^{S} p_s^2 \qquad \textbf{(61)}$$

Il peut être interprété comme la probabilité que deux individus tirés au hasard dans l'échantillon soit d'espèces différentes. Il est compris dans l'intervalle $[0; 1[$. Sa valeur diminue avec la régularité de la distribution : $E_s = 0$ si une seule espèce a une fréquence de 1, $E_s = 1 - 1/S$ si les S espèces présentes ont la même fréquence $p_s = 1/S$. La valeur 1 est atteinte pour un nombre S infini d'espèces, de fréquences nulles.

Entropie généralisée

Les premiers travaux consistant à généraliser l'indice de Shannon sont dus à Rényi (1961). L'entropie d'ordre α de Rényi est :

$$R_\alpha = \frac{1}{1-\alpha} \ln \sum_{s=1}^{S} p_s^\alpha \qquad \textbf{(62)}$$

Rényi pose également les axiomes pour une mesure d'entropie, dont :

- La symétrie : les espèces doivent être interchangeables, aucune n'a de rôle particulier et leur ordre est indifférent,
- La mesure doit être continue par rapport aux probabilités,
- La valeur maximale (fixée à 1) est atteinte si toutes les probabilités sont égales,

Il montre que R_α respecte les 3 axiomes.

Patil et Taillie (1982) ont montré de plus que :

- L'introduction d'une espèce dans une communauté augmente sa diversité (conséquence de la décroissance de la fonction d'information $g(p_s)$, voir Définition de l'entropie page 90),
- Le remplacement d'un individu d'une espèce fréquente par un individu d'une espèce plus rare augmente l'entropie à condition que $R(\wp)$ soit concave. Dans la littérature économique sur les inégalités, cette propriété est connue sous le nom de Pigou-Dalton (Dalton, 1920).

Hill (1973) transforme l'entropie de Rényi en « nombres de Hill », qui en sont simplement l'exponentielle :

$$N_\alpha = \left(\sum_{s=1}^{S} p_s^\alpha \right)^{\frac{1}{1-\alpha}} \quad \textbf{(63)}$$

Le souci de Hill est de rendre les indices de diversité intelligibles après l'article remarqué de Hurlbert (1971) intitulé « le non-concept de diversité spécifique ». Hurlbert reprochait à la littérature sur la diversité sa trop grande abstraction et son éloignement des réalités biologiques, notamment en fournissant des exemples dans lesquels l'ordre des communautés n'est pas le même selon l'indice de diversité choisi. Les nombres de Hill sont le nombre d'espèces équiprobables donnant la même valeur de diversité que la distribution observée. Ils sont des transformations simples des indices classiques :

- N_0 est le nombre d'espèces,
- $N_1 = e^H$, l'exponentielle de l'indice de Shannon,
- $N_2 = \frac{1}{1-E_s}$, l'inverse d'une des formes de l'indice de Simpson ($D_s = 1 - E_s$ est la probabilité que deux individus tirés au hasard soient de la même espèce).

Ces résultats avaient déjà été obtenus avec une autre approche par MacArthur (1965) et repris par Adelman (1969) dans la littérature économique.

Lande (1996) précise le concept de décomposabilité en postulant que les mesures de diversité doivent être concaves : la diversité d'un jeu de données regroupant plusieurs communautés doit être supérieure ou égale à la somme pondérée des diversités dans chaque communauté. De cette façon, il est possible de définir une diversité totale égale à la somme pondérée des diversités α (intra-communautés) et β (inter-communautés), toutes les diversités étant positives ou nulles. Il note que « la partition serait plus facilement interprétable si les différentes composantes de la diversité pouvaient être exprimés au moyen de la même formule ».

Lande rejette l'utilisation des nombres de Hill parce que N_2 n'est pas concave : dans certains cas, la diversité totale est inférieure à la somme des diversités intra.

Tsallis (1988) propose une classe de mesures appelée entropie généralisée et définie par :

$$S_\alpha = \frac{1}{\alpha - 1}\left(1 - \sum_{s=1}^{S} p_s^\alpha\right) \quad \textbf{(64)}$$

Tsallis a montré que les indices de Simpson et de Shannon étaient des cas particuliers d'entropie généralisée. Ces résultats ont été complétés par d'autres et repris en écologie par Keylock (2005) et Jost (2006; 2007). Nous en retiendrons que, à une normalisation éventuelle près :

- Le nombre d'espèces est S_0
- L'indice de Shannon est S_1
- L'indice de Gini-Simpson est S_2

Synthèse

Les différentes approches et généralisations ne donnent lieu finalement qu'à trois mesures, le nombre d'espèce, l'indice de Shannon et celui de Gini-Simpson. Les entropies généralisées d'ordres plus élevés ou négatifs n'ont pas trouvé d'application, de même que l'entropie de Rényi d'ordre différent de 1. L'intérêt de ces approches est d'avoir mis en évidence la différence entre les trois mesures :

- Le nombre d'espèces est la mesure qui donne le plus d'importance aux espèces rares : toutes les espèces ont la même importance, quel que soit leur effectif en termes d'individus. Il est bien adapté à une approche patrimoniale, celle du collectionneur qui considère que l'existence d'une espèce supplémentaire a un intérêt en soi, par exemple parce qu'elle peut contenir une molécule valorisable.
- L'indice de Shannon donne la même importance à tous les individus. Il est adapté à une approche d'écologue, intéressé par les interactions possibles : le nombre de combinaisons d'espèces en est une approche satisfaisante.
- L'indice de Gini-Simpson donne moins d'importance aux espèces rares. Il comptabilise les interactions possibles entre paires d'individus : les espèces rares interviennent dans peu de paires, et influent peu sur l'indice.

Les nombres de Hill, ou « nombres d'espèces équivalentes » ou « nombres d'espèces efficaces » permettent une appréhension plus intuitive de la notion de biodiversité (Jost, 2006).

Les indices d'inégalité en économie

L'indice de Theil

Theil (1967) a défini son indice d'inégalité des revenus comme :

$$T = \frac{1}{I}\sum_{i=1}^{I}\frac{y_i}{\overline{y}}\ln\frac{y_i}{\overline{y}} \quad \textbf{(65)}$$

Les données y_i peuvent être des effectifs, auquel cas la question est strictement identique à celle traitée en écologie, mais la littérature traite plus fréquemment des données de revenu (Shorrocks et Wan, 2005). y_i est le revenu par habitant de chacune des I régions, et $\overline{y} = \frac{1}{I}\sum_{i=1}^{I} y_i$ le revenu moyen sur toute la zone. Une présentation détaillée de l'indice est donnée par Conceição et Ferreira (2000).

L'indice varie entre 0 (une seule région capte tous les revenus) et $-\ln I/I$ si toutes les régions sont identiques.

Entropie généralisée

Une mesure généralisée de l'entropie a été développée par les économistes, parallèlement aux travaux des physiciens. Shorrocks et Wan (2005, p. 61) en résument l'historique. L'entropie généralisée d'ordre α est (nous suivons les notations de Brülhart et Traeger, 2005) :

$$\mathrm{GE}(\alpha) = \frac{1}{\alpha^2 - \alpha}\left(\frac{1}{I}\sum_{i=1}^{I}\left(\frac{y_i}{\overline{y}}\right)^{\alpha} - 1\right) \quad \textbf{(66)}$$

L'indice de Theil et les indices de variance relative sont des cas particuliers d'indices d'entropie généralisée.

> **Démonstration** :
>
> La forme générale des indices d'entropie généralisée est :
>
> $$\mathrm{GE}(\alpha) = \frac{1}{\alpha^2 - \alpha}\left(\frac{1}{I}\sum_{i=1}^{I}\left(\frac{y_i}{\overline{y}}\right)^{\alpha} - 1\right)$$
>
> Notons que $\sum_{i=1}^{x}\frac{y_i}{\overline{y}} = x$. On réécrit donc :

$$\mathrm{GE}(\alpha) = \frac{1}{\alpha^2 - \alpha}\left(\frac{1}{I}\sum_{i=1}^{I}\left[\left(\frac{y_i}{\overline{y}}\right)^{\alpha} - \frac{y_i}{\overline{y}}\right]\right)$$

Indice de Theil

On fixe $\alpha = 1$.

Lorsque $\alpha \to 1$, on peut écrire le développement limité à l'ordre 1 de $u^{\alpha} = u + (\alpha - 1)u\ln u + \cdots$

On substitue le développement limité dans $\mathrm{GE}(\alpha)$, quand $\alpha \to 1$:

$$\mathrm{GE}(\alpha) = \frac{1}{\alpha^2 - \alpha}\left(\frac{1}{I}\sum_{i=1}^{I}(\alpha - 1)\frac{y_i}{\overline{y}}\ln\frac{y_i}{\overline{y}}\right)$$

Après simplifications :

$$\mathrm{GE}(1) = \frac{1}{I}\sum_{i=1}^{I}\frac{y_i}{\overline{y}}\ln\frac{y_i}{\overline{y}}$$

Indice $\boldsymbol{GE(2)}$

Pour $\alpha = 2$:

$$\mathrm{GE}(2) = \frac{1}{2}\left(\frac{1}{I}\sum_{i=1}^{I}\left[\left(\frac{y_i}{\overline{y}}\right)^{2} - \frac{y_i}{\overline{y}}\right]\right)$$

En réduisant au même dénominateur, puis en factorisant et en séparant la somme en deux :

$$\mathrm{GE}(2) = \frac{1}{2\overline{y}^2}\left(\frac{1}{I}\sum_{i=1}^{I}y_i^2 - \overline{y}^2\right) = \frac{1}{2\overline{y}^2}\mathrm{Var}(y)$$

Finalement :

$$\mathrm{GE}(2) = \frac{1}{2}\mathrm{CV}(y)^2$$

À une constante près ($\frac{1}{2\overline{y}}$), l'indice est de la même forme que les indices de variance relative (page 185).

Théorie de l'information

Une présentation rapide de la théorie de l'information est utile pour une bonne compréhension des mesures d'entropie.

Les textes fondateurs sont Davis (1941) et surtout Theil (1967) en économétrie, et Shannon (1948 ; 1963) pour la mesure de la diversité. Une revue est fournie par Maasoumi (1993).

Définition de l'entropie

Considérons une expérience dont les résultats possibles sont $\{r_1, r_2, \dots, r_S\}$. La probabilité d'obtenir r_s est p_s, et $\mathcal{P} = \{p_1; p_2; \dots; p_S\}$. Les probabilités sont connues *a priori*. Tout ce qui suit est vrai aussi pour des valeurs de r continues, dont on connaîtrait la densité de probabilité.

On considère maintenant un échantillon de valeurs de r. La présence de r_s dans l'échantillon est peu étonnante si p_s est grande : elle apporte peu d'information supplémentaire par rapport à la simple connaissance des probabilités. En revanche, si p_s est petite, la présence de r_s apporte beaucoup d'information. On définit donc une fonction d'information, $g(p_s)$, décroissante quand la probabilité augmente, de $g(0) = +\infty$ (ou éventuellement une valeur strictement positive finie) à $g(1) = 0$. Chaque valeur observée dans l'échantillon apporte une certaine quantité d'information, dont la somme est l'information de l'échantillon.

La quantité d'information attendue de l'expérience est $\sum_{s=1}^{S} p_s g(p_s) = H(\mathcal{P})$. Si on choisit $g(p_s) = -ln(p_s)$, $H(\mathcal{P})$ est l'indice de Shannon, mais bien d'autres formes de $g(p_s)$ sont possibles. $H(\mathcal{P})$ est appelée *entropie*. C'est une mesure de l'incertitude (de la volatilité) du résultat de l'expérience. Si le résultat est certain, l'entropie est nulle. L'entropie est maximale quand les résultats sont équiprobables.

Si $\mathcal{P}$ est la distribution des fréquences des espèces dans une communauté, Patil et Taillie (1982) montrent que :

- Si $g(p_s) = \frac{1-p_s}{p_s}$, alors $H(\mathcal{P})$ est le nombre d'espèces S,
- Si $g(p_s) = -ln(p_s)$, alors $H(\mathcal{P})$ est l'indice de Shannon,
- Si $g(p_s) = 1 - p_s$, alors $H(\mathcal{P})$ est l'indice de Simpson.

Écarts entre distributions

Considérons maintenant les probabilités q_i formant l'ensemble $\mathcal{Q}$ obtenues par la réalisation de l'expérience. Elles sont différentes des probabilités p_i, par exemple parce que l'expérience ne s'est pas déroulée exactement comme prévu. On définit le gain d'information $I(\mathcal{Q},\mathcal{P})$ comme la quantité d'information supplémentaire fournie par l'expérience, connaissant les probabilités *a priori*. Ce gain d'information peut être vu comme une distance entre la distribution *a priori* et la distribution *a posteriori*. Il est possible que les distributions $\mathcal{P}$ et $\mathcal{Q}$ soit identiques, que le gain d'information soit donc nul, mais les estimateurs empiriques n'étant pas exactement égaux entre eux, des tests de significativité de la valeur de $\hat{I}(\mathcal{Q},\mathcal{P})$ seront nécessaires.

Quelques formes possibles de $I(\mathcal{Q},\mathcal{P})$ sont :

- La divergence de Kullback-Leibler (1951) connue par les économistes comme l'indice de dissimilarité de Theil (1967) :

$$T = \sum_{s=1}^{S} q_s \ln \frac{q_s}{p_s} \tag{67}$$

- Sa proche parente, appelée parfois deuxième mesure de Theil (Conceição et Ferreira, 2000, p. 34), qui inverse simplement les rôles de p et q :

$$L = \sum_{s=1}^{S} p_s \ln \frac{p_s}{q_s} \tag{68}$$

- L'entropie généralisée (Maasoumi, 1993), d'ordre γ :

$$I_\gamma(\mathcal{Q},\mathcal{P}) = \frac{1}{\gamma(\gamma+1)}\left[\sum_{s=1}^{S} q_s \left(\frac{q_s}{p_s}\right)^\gamma - 1\right] \tag{69}$$

Un calcul similaire à celui de la page 88 permet de montrer que $T = I_0$ et $L = I_{-1}$.

L'entropie généralisée de Brülhart et Traeger est un cas particulier de I_γ dans lequel les probabilités *a priori* sont égales (il suffit d'écrire $p_s = \frac{1}{S}$, $q_s = \frac{y_s}{x\bar{y}}$ et $\alpha = \gamma + 1$ pour retrouver l'équation (66)).

L'indice d'inégalité des revenus de Theil est un cas particulier de l'indice de dissimilarité. On considère l'ensemble des revenus y. La probabilité *a priori* qu'un élément du revenu (disons un euro) soit attribué à la région i est $p_i = 1/I$. La fré-

quence observée *a posteriori* est $q_i = y_i/(I\bar{y})$, la part de la région en termes de revenu.

Démonstration :

$$\hat{T} = \sum_{i=1}^{I} \frac{y_i}{I\bar{y}} \ln \frac{{}^{y_i}/_{I\bar{y}}}{{}^{1}/_{I}}$$
$$= \frac{1}{I} \sum_{i=1}^{I} \frac{y_i}{\bar{y}} \ln \frac{y_i}{\bar{y}}$$

Unification

On s'intéresse à des données discrètes, qui présentent des effectifs (par exemple des nombres d'arbres) par catégorie (espèces) localisés dans des zones (placettes forestières). Les catégories peuvent être regroupées (les espèces par genre puis par famille, les placettes par parcelle puis par forêt).

			Forêt k				Total
			Parcelle j		Parcelle $j+1$		
			Placette i	Placette $i+1$	Placette $i+2$	…	
Famille u	Genre t	Espèce s	y_{si}				y_{s+}
		Espèce $s+1$					
	Genre $t+1$	Espèce $s+2$					
		…					
Total			y_{+i}				y

Tableau 4 : probabilités attendues et distributions observées.

Le tableau sera appelé par la suite : « tableau espèces-placettes ». On note :

- y_{si} le nombre d'arbres de l'espèce s dans la placette i, y_{+i} est le nombre d'arbres de la placette i, toutes espèces confondues. y_{s+} est le nombre d'arbres total de l'espèce s.
- $\mathcal{I} = \{1,2,\dots,I\}$ l'ensemble des placettes, et $\mathcal{S} = \{1,2,\dots,S\}$ celui des espèces.

- $\mathcal{J}_1 = \{1,2,\dots,I_1\},\dots,\mathcal{J}_j = \{I_{j-1}+1,\dots,I_j\},\dots,\mathcal{J}_J = \{I_{J-1}+1,\dots,I_J\}$ l'ensemble des placettes appartenant à la parcelle j et $\mathcal{T}_t$ l'ensemble des espèces du genre t. J est le nombre de parcelles, I_j est l'indice de la dernière placette de la parcelle j et donc $I_J = I$.
- $\mathcal{K}_1 = \{1,2,\dots,J_1\},\dots,\mathcal{K}_k = \{J_{k-1}+1,\dots,J_k\},\dots,\mathcal{K}_K = \{J_{k-1}+1,\dots,J_K\}$ l'ensemble des parcelles appartenant à la forêt k et $\mathcal{U}_u$ l'ensemble des genres de la famille u.

Les effectifs y_{si} sont observés. Ils vont permettre d'estimer les probabilités $\mathcal{P}$ et $\mathcal{Q}$. Par exemple, si on s'attend à une distribution dans laquelle la probabilité qu'un individu se trouve dans une placette est proportionnelle à l'importance de la placette et de l'espèce, p_{si} sera estimé par $\hat{p}_{si} = {y_{+i}y_{s+}}/{y^2}$. Cette valeur n'est qu'un estimateur de la probabilité parce que y_{+i}, y_{s+} et y sont des estimateurs des tailles inconnues de la placette, de l'espèce et la communauté, obtenus en sommant les y_{si}, tirages de variables aléatoires Y_{si} dépendant de ces tailles.

Dans un premier temps, nous ne nous intéressons qu'à une espèce s sur l'ensemble des placettes, ou à l'ensemble des espèces sur une placette i. En d'autres termes, on ne dispose que des données de la première ligne et de la première colonne du tableau.

L'approche classique en écologie consiste à utiliser l'indice de Shannon pour mesurer la biodiversité sur la placette i. La probabilité qu'un individu soit de l'espèce s est estimée par $\frac{y_{si}}{y_{+i}}$, d'où : $\widehat{H} = -\sum_{s=1}^{S} \frac{y_{si}}{y_{+i}} \ln \frac{y_{si}}{y_{+i}}$.

Il s'agit d'une mesure absolue au sens de Brülhart et Traeger (2005) : elle ne dépend d'aucune référence extérieure comme les effectifs relatifs des espèces connus *a priori*. Ses valeurs extrêmes sont 0 si tous les arbres sont de la même espèce et lnS si les effectifs des espèces sont égaux.

La même méthode peut être utilisée pour mesurer la concentration spatiale absolue des arbres d'une espèce : $\widehat{H}' = -\sum_{i=1}^{I} \frac{y_{si}}{y_{s+}} \ln \frac{y_{si}}{y_{s+}}$. Cette mesure n'est jamais utilisée en écologie.

L'indice de Shannon et H' sont des cas particuliers de l'indice de Theil. En terme de diversité, si les probabilités *a priori* des espèces sont égales ($p_s = 1/S$) :

$$\widehat{T} = \sum_{s=1}^{S} \frac{y_{si}}{y_{+i}} \ln \frac{{y_{si}}/{y_{+i}}}{{1}/{S}} = \sum_{s=1}^{S} \frac{y_{si}}{y_{+i}} \ln \frac{y_{si}}{y_{+i}} + \ln S = \ln S - \widehat{H} \qquad \textbf{(70)}$$

De même en termes de concentration spatiale, si le choix des placettes est *a priori* équiprobable. Alors, $p_i = 1/I$, d' où :

$$\hat{T} = \sum_{i=1}^{I} \frac{y_{si}}{y_{s+}} \ln \frac{y_{si}/y_{s+}}{1/I} = \sum_{i=1}^{x} \frac{y_{si}}{y_{s+}} \ln \frac{y_{si}}{y_{s+}} + \ln I = \ln I - \hat{H}' \quad \textbf{(71)}$$

En conclusion, l'approche la plus générale est l'indice de dissimilarité de Theil (67)), qui compare une probabilité observée à une probabilité attendue. Sa forme a été établie pour la première fois par Kullback et Leibler (1951). Il ne s'agit pas vraiment d'une distance entre les distributions de probabilités parce qu'il n'est pas symétrique (les rôles de p et q ne sont pas interchangeables).

En termes de diversité par exemple, il est égal à la différence entre le logarithme du nombre d'espèces et l'indice de Shannon si les probabilités attendues sont les mêmes pour toutes les espèces. L'indice de Shannon est compris entre 0 (concentration maximale) et lnS (équirépartition) ; l'indice de Theil varie entre les mêmes bornes en sens inverse, donc en sens inverse de la diversité (il mesure l'écart à la diversité maximale) : il mesure la *spécialisation* des zones géographiques, notion familière aux économistes (Houdebine, 1999 par exemple).

En statistiques spatiales, les données seront des effectifs d'arbres d'une espèce dans différentes placettes. L'indice H' évalue la concentration absolue, sans référence à l'importance des parcelles. En introduisant une mesure d'importance des placettes n_{si}, l'indice de dissimilarité de Theil caractérise la concentration :

- Topographique si n_{si} représente la surface des placettes
- Relative si n_{si} représente le nombre total d'arbres, toutes espèces confondues, dans les placettes.

Cette approche géographique est directement transposable en termes de diversité, en utilisant les colonnes au lieu des lignes du tableau. L'indice de Shannon est la mesure classique de la diversité, absolue au sens où elle considère toutes les espèces comme équiprobables a priori. L'indice de dissimilarité de Theil mesure l'écart entre une distribution observée et une distribution attendue, par exemple la fréquence des espèces d'une communauté plus grande : on peut parler de diversité relative. Il n'y a pas d'équivalent pour la diversité de la concentration géographique.

Décomposition

Bourguignon (1979) définit une mesure d'inégalité décomposable comme respectant les propriétés suivantes :

- La population totale étant partitionnée, chaque partition recevant un poids, la composante intra-groupe de la mesure est égale à la somme pondérée des mesures dans chaque-groupe.
- La composante intergroupe est la mesure d'inégalité entre les groupes.
- La mesure totale est la somme des mesures intra et intergroupes.

Bourguignon montre que l'indice de Theil est la seule mesure décomposable, homogène de degré 0 et dont la somme des poids vaut 1.

Cutrini (2009) applique l'indice de Theil à un tableau de données dont les lignes sont des localisations emboîtées (régions dans pays) et les colonnes de secteurs d'activité économique. Il s'agit donc d'une version simplifiée (et transposée) du tableau espèces-placettes. Elle définit un indice de localisation globale égal à la fois à la somme des indices de concentration spatiale de tous les secteurs et à la somme des indices de spécialisation de toutes les régions. Cette approche peut être généralisée et les calculs simplifiés.

Règle générale

Le tableau espèces-placettes peut être présenté en termes de probabilités. Dans la distribution observée, $\hat{q}_{si} = {}^{y_{si}}/_{y}$. La probabilité attendue p_{si} sera définie différemment suivant la question étudiée et l'hypothèse choisie :

<table>
<tr><td colspan="3"></td><td>Forêt k ...</td><td></td><td rowspan="3">Total</td></tr>
<tr><td colspan="3"></td><td>Parcelle j ...</td><td></td></tr>
<tr><td colspan="3"></td><td>Placette i</td><td>...</td></tr>
<tr><td rowspan="2">Famille u

...</td><td rowspan="2">Genre t

...</td><td>Espèce s</td><td>p_{si}
$\hat{q}_{si} = {}^{y_{si}}/_{y}$</td><td></td><td>$\hat{q}_{s+} = {}^{y_{s+}}/_{y}$</td></tr>
<tr><td>...</td><td></td><td></td><td></td></tr>
<tr><td colspan="3">Total</td><td>$\hat{q}_{+i} = {}^{y_{+i}}/_{y}$</td><td></td><td>1</td></tr>
</table>

Tableau 5 : probabilités attendues et distributions observées.

$T_{si} = q_{si} \ln \frac{q_{si}}{p_{si}}$ est la contribution d'une cellule d'une tableau à l'entropie totale. Soit l'indice T_I défini sur un ensemble I de valeurs individuelles, par exemple l'ensemble des cellules du tableau ci-dessus : $T_I = \sum_{i \in I} q_i \ln \frac{q_i}{p_i} = \sum_{i \in I} T_{si}$. N'importe quel groupement des valeurs est possible : soit G un groupe, $G \subset I$, alors la contribution du groupe (T_G^{α}) à l'entropie totale est égale à la somme de son entropie après regroupement, qu'on appellera entropie gamma (T_G^{γ}), et de la valeur pondé-

rée (par le poids du groupe $\sum_{g\in G} q_g$)de son entropie entre individus (T_G^{β}). La somme des entropies individuelles des éléments du groupe est :

$$T_G^{\alpha} = \sum_{g\in G} q_g \ln \frac{q_g}{p_g} \quad \textbf{(72)}$$

L'entropie gamma de G est celle de la cellule unique obtenue après regroupement :

$$T_G^{\gamma} = q_G \ln \frac{q_G}{p_G} = \left(\sum_{g\in G} q_g\right) \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g} \quad \textbf{(73)}$$

La probabilité *a priori* (p_G) ou *a posteriori* (q_G) qu'un individu appartienne au groupe sont les sommes des probabilités de tous les éléments du groupe.

L'entropie inter-individus de G est :

$$T_G^{\beta} = \sum_{g\in G} \frac{q_g}{\sum_{g\in G} q_g} \ln \frac{\frac{q_g}{\sum_{g\in G} q_g}}{\frac{p_g}{\sum_{g\in G} p_g}} = \left(\sum_{g\in G} q_g\right)^{-1} \left[\sum_{g\in G} q_g \ln \frac{q_g}{p_g} - \left(\sum_{g\in G} q_g\right) \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g}\right] \quad \textbf{(74)}$$

À l'intérieur du groupe, la somme des probabilités vaut 1. Les probabilités intra-groupe sont égales aux probabilités de départ divisées par la probabilité totale du groupe.

Au total, la contribution de tous les éléments du groupe est bien égale à la somme de l'entropie gamma et de l'entropie inter-individus :

$$T_G^{\alpha} = T_G^{\gamma} + \left(\sum_{g\in G} q_g\right) T_G^{\beta} \quad \textbf{(75)}$$

Démonstration :

$$
\begin{aligned}
T_G^\gamma + \left(\sum_{g\in G} q_g\right) T_G^\beta & \\
&= \sum_{g\in G} q_g \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g} \\
&+ \left(\sum_{g\in G} q_g\right)\left(\sum_{g\in G} q_g\right)^{-1} \left[\sum_{g\in G} q_g \ln \frac{q_G}{p_G} - \left(\sum_{g\in G} q_g\right) \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g}\right] \\
&= \sum_{g\in G} q_g \ln \frac{q_g}{p_g} = T_G^\alpha
\end{aligned}
$$

L'égalité (75) peut être sommée pour tous les groupes G formant une partition de I pour obtenir :

$$\sum_{G\subset I} T_G^\alpha = \sum_{G\subset I} T_G^\gamma + \sum_{G\subset I}\left(\sum_{g\in G} q_g\right) T_G^\beta \qquad \textbf{(76)}$$

Dans la terminologie classique de l'analyse de variance, $\sum_{G\subset I} T_G^\alpha$ est l'entropie totale, qui se décompose en somme pondérée de l'entropie intra-groupe $\sum_{G\subset I}\left(\sum_{g\in G} q_g\right) T_G^\beta$ et de l'entropie inter-groupes $\sum_{G\subset I} T_G^\gamma$.

Localisation globale

On peut se placer dans un cadre complètement relatif : la probabilité *a priori* qu'un arbre se trouve dans la placette i est estimée par ${}^{y_{+i}}/_{y}$, la probabilité qu'il appartienne à l'espèce s par ${}^{y_{s+}}/_{y}$, et la probabilité qu'il appartienne à l'espèce s et se trouve dans la placette i par $\hat{p}_{si} = {}^{y_{+i}y_{s+}}/_{y^2}$. Les probabilités *a posteriori* sont $\hat{q}_{si} = {}^{y_{si}}/_{y}$. C'est l'approche habituelle des modèles cherchant à caractériser la concentration spatiale (Ellison et Glaeser, 1997) :

			Forêt k	…	Total
			Parcelle j	…	
			Placette i	…	
Famille u	Genre t	Espèce s	$\hat{p}_{si} = {}^{y_{+i}y_{s+}}/_{y^2}$ $\hat{q}_{si} = {}^{y_{si}}/_{y}$		$\hat{p}_{s+} = \hat{q}_{s+} = {}^{y_{s+}}/_{y}$
…	…	…			
Total			$\hat{p}_{+i} = \hat{q}_{+i} = {}^{y_{+i}}/_{y}$		1

Tableau 6 : probabilités attendues et distributions observées dans un cadre complètement relatif.

L'indice de localisation globale de Cutrini (2009) est :

$$L = \sum_{i=1}^{I}\sum_{s=1}^{S} q_{si}\ln\frac{q_{si}}{p_{si}}$$
$$\hat{L} = \sum_{i=1}^{I}\sum_{s=1}^{S} {}^{y_{si}}/_{y}\ln\frac{{}^{y_{si}}/_{y}}{{}^{y_{+i}y_{s+}}/_{y^2}} = \sum_{i=1}^{I}\sum_{s=1}^{S} {}^{y_{si}}/_{y}\ln\frac{{}^{y_{si}}/_{y_{+i}}}{{}^{y_{s+}}/_{y}} \quad \textbf{(77)}$$

La première forme est la plus compacte, la deuxième la plus développée et la dernière est celle utilisée par Cutrini. Elle montre que L peut être décomposé en la somme pondérée des concentrations spatiales de toutes les espèces :

$$\hat{L} = \sum_{s=1}^{S}\frac{y_{s+}}{y}\hat{T}_s \quad \textbf{(78)}$$

Nous allons montrer que tout regroupement par ligne ou par colonne fonctionne aussi simplement. Notamment, L est égal à la somme des spécialisations de toutes les placettes :

$$\hat{L} = \sum_{i=1}^{I}\frac{y_{+i}}{y}\hat{T}_i \quad \textbf{(79)}$$

> $T_i = \sum_{s\in\mathcal{S}} q_{si}\ln\frac{q_{si}}{p_{si}}$ est d'autant plus grand que la diversité est faible. C'est une mesure de *spécialisation* dans le vocabulaire habituel de l'économie géographique. Ses valeurs possibles varient de 0 quand les q_{si} sont toutes égales aux p_{si} à $-\ln(\min(p_{si}))$ quand tous les arbres sont de l'espèce attendue comme la plus rare (Mori *et al.*, 2005, note 13).

Commençons par regrouper tous les éléments d'une placette i. La somme des entropies, que l'on peut appeler spécialisation α de la placette, est :

$$\hat{T}_i^{\alpha} = \sum_{s \in \mathcal{S}} \frac{y_{si}}{y} \ln \frac{y y_{si}}{y_{s+} y_{+i}} \tag{80}$$

L'entropie γ est nulle parce que les probabilités marginales (somme des lignes ou sommes des colonnes) attendues ont été construites de façon à être égales aux distributions observées :

$$\hat{T}_j^{\gamma} = \left(\sum_{s \in \mathcal{S}} \frac{y_{si}}{y} \right) \ln \frac{\sum_{s \in \mathcal{S}} \frac{y_{si}}{y}}{\sum_{s \in \mathcal{S}} \frac{y_{+i} y_{s+}}{y^2}} = \frac{y_{+i}}{y} \ln \frac{\frac{y_{+i}}{y}}{\frac{y_{+i}}{y}} = 0 \tag{81}$$

L'entropie totale se résume donc à l'entropie β, c'est-à-dire la composante intra-placette. On vérifie que $\left(\sum_{s \in \mathcal{S}} \frac{y_{si}}{y} \right) \hat{T}_i^{\beta} = \hat{T}_i^{\alpha}$:

$$\frac{y_{+i}}{y} \hat{T}_i^{\beta} = \frac{y_{+i}}{y} \sum_{s \in \mathcal{S}} \frac{\frac{y_{si}}{y}}{\sum_{s \in \mathcal{S}} \frac{y_{si}}{y}} \ln \frac{\frac{\frac{y_{si}}{y}}{\sum_{s \in \mathcal{S}} \frac{y_{si}}{y}}}{\frac{\frac{y_{+i} y_{s+}}{y^2}}{\sum_{s \in \mathcal{S}} \frac{y_{+i} y_{s+}}{y^2}}} = \sum_{s \in \mathcal{S}} \frac{y_{si}}{y} \ln \frac{y y_{si}}{y_{s+} y_{+i}} = \hat{T}_i^{\alpha} \tag{82}$$

La même démonstration vaut pour le regroupement des éléments d'une espèce s en inversant le rôle des indices i et s.

Le regroupement de plusieurs placettes en une parcelle se fait également en ajoutant les entropies individuelles. Les entropies γ sont nulles ici encore :

$$\hat{T}_j^{\alpha} = \sum_{i \in \mathcal{J}_j} \hat{T}_i^{\alpha} = \hat{T}_j^{\beta} = \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \hat{T}_i^{\beta} \tag{83}$$

La nullité des entropies γ fait que l'entropie totale de la parcelle est égale à la somme des spécialisations des placettes. En regroupant toutes les placettes de la forêt, on obtient l'équation (83). Le raisonnement est identique pour montrer que la concentration spatiale d'un genre est égale à la somme des concentrations spatiales des espèces qui le composent.

Test d'une hypothèse nulle

La possibilité de comparer les données à une hypothèse nulle est apparue essentielle dès le début de la théorisation des processus ponctuels. De façon surprenante, la question est assez récente en statistiques spatiales discrètes, développées depuis longtemps en économie mais assez peu en écologie. La mesure traditionnelle de concentration spatiale en économie est l'indice de Gini (voir page 188). Il compare une statistique calculée à partir des données à sa valeur dans le cas d'une équirépartition parfaite, jamais atteinte, sans que l'on puisse savoir si la cause est un mécanisme précis ou une simple fluctuation stochastique. Ellison et Glaeser (1997) ont été les premiers à construire un indice de concentration spatiale (voir page 189) autour d'une distribution aléatoire des points, selon une probabilité proportionnelle à la taille des zones géographiques. Dans ce cadre, l'équirépartition parfaite est un extrême, l'autre étant celui de la concentration de tous les points dans une seule zone, et les deux reviennent à rejeter l'hypothèse nulle.

L'indice de Theil compare une distribution observée à une distribution attendue, il est donc naturel de choisir comme hypothèse nulle l'égalité entre les deux. Toutes les valeurs de T peuvent être testées par la méthode de Monte Carlo :

- Tirer toutes les valeurs de y_{si} selon la loi appropriée et calculer T
- Répéter l'opération un grand nombre de fois (par exemple 10 000) et retenir les percentiles correspondant au seuil de risque choisi (par exemple, la 251ème et la 9750ème valeur sont les bornes de l'intervalle de confiance de l'hypothèse nulle au seuil de 5% avec 10 000 simulations).

Dans le cadre relativiste de Cutrini, la probabilité *a priori* qu'un arbre soit de l'espèce s et dans la placette i est $\hat{p}_{si} = {y_{+i}y_{s+}}/{y^2}$ et les individus sont tirés indépendamment les uns des autres. Le nombre d'arbres attendus dans chaque case du tableau espèces-placettes suit donc une loi binomiale $\mathcal{B}(y, \hat{p}_{si})$. Pour tester la significativité de la valeur observée de L, il suffit de tirer tous les y_{si} selon cette loi et calculer les valeurs de L qui en résultent. La méthode est la même pour toute décomposition de L (concentration spatiale d'une espèce par exemple). De plus, chaque valeur y_{si} peut être testée analytiquement contre son hypothèse nulle d'appartenance à une loi binomiale : si le nombre total d'arbres y est assez grand, la loi binomiale tend vers une loi normale, et l'intervalle de confiance de la valeur attendue de y_{si} est $y\hat{p}_{si} \pm t_\alpha(y-1)\sqrt{y\hat{p}_{si}(1-p_{si})}$, où $t_\alpha(y-1)$ est la valeur critique de la loi de Student à $y-1$ degrés de liberté (1,96 au seuil $\alpha = 5\%$, puisque y est grand).

Lorsque le nombre d'individus y augmente, T tend vers une loi de χ^2, qui mesure la somme des écarts $(q_{si} - p_{si})^2$. La méthode de Monte-Carlo est préférable ici aux tests asymptotiques parce qu'elle permet de s'affranchir de l'hypothèse de convergence (il n'est jamais certain que y soit assez grand) et de tester n'importe quelle hypothèse sur une partie des données, par exemple la conformité au modèle nul de la distribution d'une seule espèce (sa distribution spatiale est-elle proportionnelle à la taille des placettes ou non ?) ou d'une seule placette (son niveau de biodiversité est-il exceptionnellement haut ou bas ?).

Application : Décomposition de l'indice de Shannon

Formulation de la diversité β

Pour décomposer la biodiversité de la parcelle j, dont les placettes sont $\mathcal{J}_j = \{I_{j-1} + 1, \ldots, I_j\}$, le regroupement se fait espèce par espèce :

	Parcelle j						Parcelle j
	Placette I_{j-1}	…	Placette i	…	PlacetteI_j		regroupée
…						→	
Espèce s			$\hat{p}_{si} = {y_{+i}}/{Sy}$ $\hat{q}_{si} = {y_{si}}/{y}$			→	$\hat{p}_{sj} = \sum_{i \in \mathcal{J}_j} {y_{+i}}/{Sy}$ $\hat{q}_{sj} = \sum_{i \in \mathcal{J}_j} {y_{si}}/{y}$
…						→	

Les équations (74), (75) et (76) sont appliquées directement à ce regroupement.

La contribution à l'entropie totale de l'espèce s dans la parcelle j est (zone gris clair à gauche du tableau) :

$$\hat{T}_{sj}^{\alpha} = \sum_{i \in \mathcal{J}_j} \hat{q}_{si} \ln \frac{\hat{q}_{si}}{\hat{p}_{si}} = \sum_{i \in \mathcal{J}_j} \frac{y_{si}}{y} \left(\ln \frac{y_{si}}{y_{+i}} + \ln S \right) \quad \textbf{(84)}$$

L'entropie gamma de l'espèce s dans la parcelle j est (zone grisée encadrée en pointillés) :

$$\hat{T}_{sj}^{\gamma} = \hat{q}_{sj} \ln \frac{\hat{q}_{sj}}{\hat{p}_{sj}} = \frac{\sum_{i \in \mathcal{J}_j} y_{si}}{y} \left(\ln \frac{\frac{\sum_{i \in \mathcal{J}_j} y_{si}}{y}}{\frac{\sum_{i \in \mathcal{J}_j} y_{+i}}{y}} + \ln S \right) \quad \textbf{(85)}$$

L'entropie inter-individus de l'espèce s dans la parcelle j est :

$$\left(\sum_{i \in \mathcal{J}_j} \hat{q}_{si} \right) \hat{T}_{sj}^{\beta} = \left(\sum_{i \in \mathcal{J}_j} \hat{q}_{si} \right) \sum_{i \in \mathcal{J}_j} \frac{\hat{q}_{si}}{\sum_{i \in \mathcal{J}_j} \hat{q}_{si}} \ln \frac{\frac{\hat{q}_{si}}{\sum_{i \in \mathcal{J}_j} \hat{q}_{si}}}{\frac{\hat{p}_{si}}{\sum_{i \in \mathcal{J}_j} \hat{p}_{si}}} = \sum_{i \in \mathcal{J}_j} \frac{y_{si}}{y} \ln \frac{\frac{y_{si}}{\sum_{i \in \mathcal{J}_j} y_{si}}}{\frac{y_{+i}}{\sum_{i \in \mathcal{J}_j} y_{+i}}} \quad \textbf{(86)}$$

On sait (75) que $T_{sj}^{\alpha} = T_{sj}^{\gamma} + \left(\sum_{i \in \mathcal{J}_j} q_{si} \right) T_{sj}^{\beta}$

Chacun de ces trois termes peut maintenant être sommé sur toutes les espèces pour faire apparaître les mesures de biodiversité :

$$\begin{aligned} \sum_s \hat{T}_{sj}^{\alpha} &= \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \ln S + \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{y_{si}}{y_{+i}} = \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \ln S - \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \hat{H}_i^{\alpha} \\ &= \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \ln S - \hat{H}_{\alpha} \end{aligned} \quad \textbf{(87)}$$

> La diversité α est la somme pondérée des diversités α de chaque placette. Dans chacune de ces placettes, la diversité α estimée par $\hat{H}_i^{\alpha}$ ne prend en compte que les fréquences relatives des espèces y_{si}/y_{+i} à l'intérieur de la placette, sans référence aux données hors de la placette.

$$\sum_s \hat{T}_{sj}^{\gamma} = \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \ln S + \sum_s \frac{\sum_{i \in \mathcal{J}_j} y_{si}}{y} \ln \frac{\frac{\sum_{i \in \mathcal{J}_j} y_{si}}{y}}{\frac{\sum_{i \in \mathcal{J}_j} y_{+i}}{y}} = \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \ln S - \hat{H}_{\gamma} \quad \textbf{(88)}$$

$$\begin{aligned} \sum_s \left(\sum_{i \in \mathcal{J}_j} \hat{q}_{si} \right) \hat{T}_{sj}^{\beta} &= \sum_{i \in \mathcal{J}_j} \sum_s \frac{y_{si}}{y} \ln \frac{\frac{y_{si}}{\sum_{i \in \mathcal{J}_j} y_{si}}}{\frac{y_{+i}}{\sum_{i \in \mathcal{J}_j} y_{+i}}} = \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{\frac{y_{si}}{y_{+i}}}{\frac{\sum_{i \in \mathcal{J}_j} y_{si}}{\sum_{i \in \mathcal{J}_j} y_{+i}}} \\ &= \sum_{i \in \mathcal{J}_j} \frac{y_{+i}}{y} \hat{H}_i^{\beta} = \hat{H}_{\beta} \end{aligned} \quad \textbf{(89)}$$

La diversité β est la somme pondérée des diversités β de chaque placette. Dans chacune de ces placettes, la diversité β estimée par $\widehat{H}_i^{\beta}$ est une divergence de Kullback-Leibler. La fréquence attendue pour chaque espèce est celle observée dans le groupe (ici la parcelle), $\sum_{i\in\mathcal{J}_j} y_{si}/\sum_{i\in\mathcal{J}_j} y_{+i}$, alors que la fréquence observée est celle de la placette : y_{si}/y_{+i}.

En combinant les équations (87), (88) et (89), la décomposition de la biodiversité est établie pour les estimateurs. Comme ces estimateurs sont consistants, l'égalité vaut pour les variables aléatoires :

$$H_\gamma = H_\alpha + H_\beta \tag{90}$$

En passant par les nombres de Hill, Jost (2007) montre que l'indice de Shannon est le seul pouvant être décomposé de cette façon. Mais il n'explicite pas $\widehat{H}_\beta$, seulement obtenu par la différence $\widehat{H}_\gamma - \widehat{H}_\alpha$. La forme de H_β avait été établie par Ricotta et Avena (2003), sans la relier celle de H_α et H_γ. Enfin, l'idée de la décomposition de la divergence de Kullback-Leibler, mais avec une approche différente, a été publiée par Ludovisi et Taticchi (2006).

La décomposition ci-dessus explicite les valeurs des différents niveaux de diversité et montre que la forme est bien la même dans tous les cas : une divergence de Kullback-Leibler entre une distribution observée et une distribution attendue.

Indice	*Distribution observée*	*Distribution attendue*	*Formule*
$\boldsymbol{H_\alpha}$	Fréquence des espèces dans la placette	Fréquences égales, hors formule	$\widehat{H}_i^{\alpha} = \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{y_{si}}{y_{+i}}$ $\widehat{H}_\alpha = \sum_{i\in\mathcal{J}_j} \frac{y_{+i}}{y} \widehat{H}_i^{\alpha}$
$\boldsymbol{H_\beta}$	Fréquence des espèces dans la placette	Fréquence des espèces dans la parcelle	$\widehat{H}_i^{\beta} = \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{\frac{y_{si}}{y_{+i}}}{\frac{y_{s+}}{y}}$ $\widehat{H}_\beta = \sum_{i\in\mathcal{J}_j} \frac{y_{+i}}{y} \widehat{H}_i^{\beta}$
$\boldsymbol{H_\gamma}$	Fréquence des espèces dans la parcelle	Fréquences égales, hors formule	$\widehat{H}_\gamma = \sum_s \frac{y_{s+}}{y} \ln \frac{y_{s+}}{y}$

Tableau 7 : Probabilités attendues et observées pour la définition de l'indice de diversité de Shannon.

Le cas particulier dans lequel seulement deux niveaux existent, par exemple parcelles et forêt, fournit une expression plus simple des formules : $\sum_j y_{sj} = y$, la somme des nombre d'individus dans les parcelles est égal au nombre d'individus dans la forêt, d'où les équations du Tableau 7.

Les parcelles peuvent à leur tour être regroupées en forêts, la diversité γ de la parcelle devenant diversité α pour la forêt. La décomposition ou le regroupement peuvent être effectués sur un nombre quelconque de niveaux.

Test de significativité

L'objectif est de tester si deux placettes ne sont pas simplement deux échantillons d'une même communauté, dont les différences ne sont que des fluctuations dues au hasard. Sous l'hypothèse nulle, les observations $\hat{q}_{si}$ sont des réalisations des mêmes probabilités p_{s+}.

Le test est réalisé de la façon suivante :

- Chaque valeur y_{si} est tirée dans une loi binomiale $\mathcal{B}\left(y, {}^{y_{s+}}/_{y}\right)$ et H_β est calculée,
- La simulation est répétée un grand nombre de fois, par exemple 10 000, et les valeurs extrêmes sont éliminées. Au seuil de risque $\alpha = 5\%$, les 251ème et 9750ème valeurs simulées définissent les bornes de l'intervalle de confiance de l'hypothèse nulle.

L'hypothèse nulle est rejetée si la valeur observée de H_β n'est pas dans cet intervalle, en général au-delà de la borne supérieure. Il peut arriver que les deux placettes soient plus semblables que sous l'hypothèse nulle, c'est-à-dire que les fréquences varient moins que dans le tirage d'une loi binomiale, si deux placettes ont été plantées avec le même nombre d'arbres de chaque espèce par exemple.

Le code proposé pour R (R Development Core Team, 2010) se trouve en annexe 4.

Exemples

Données simulées

Des exemples théoriques sont utiles pour comprendre les déterminants de $\widehat{H}_\beta$, notamment la richesse de la communauté et l'effort d'échantillonnage. Deux distributions de fréquences sont tirées au hasard, de respectivement 20 et 40 espèces. Les fréquences de chaque espèce sont tirées dans la même loi uniforme et normalisées pour que leur somme soit égale à 1. Une paire de placettes est en-

suite tirée 1 000 fois selon ces fréquences, avec une espérance de 500 ou 5 000 points.

$\widehat{H_\beta}$ est calculée pour chaque paire de placettes et les résultats sont affichés sous la forme d'un histogramme des fréquences, lissé pour obtenir une densité de probabilité par la fonction `density` de R.

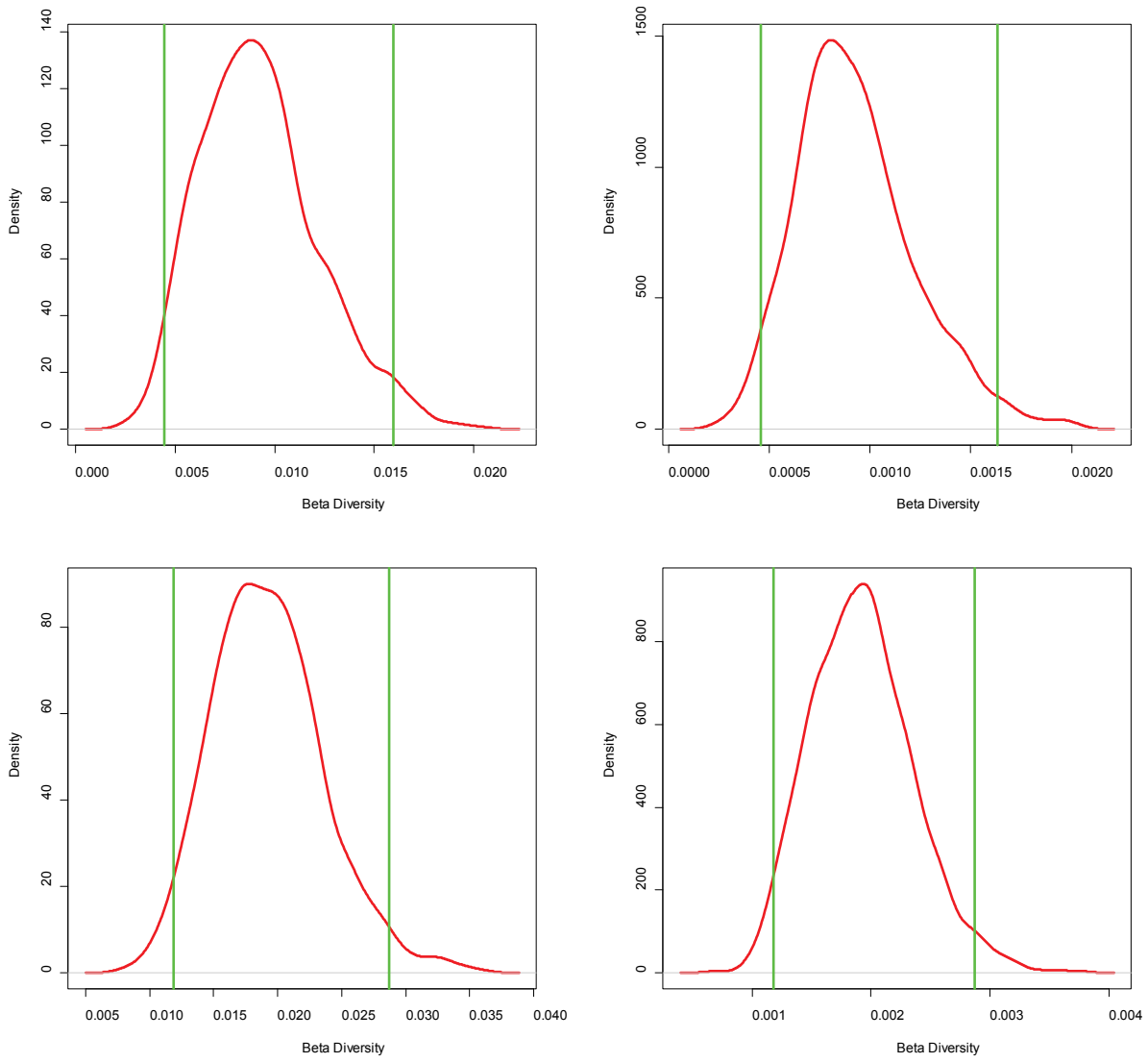


Figure 37: Densités de probabilité de H_β obtenues à partir de 1 000 simulations du modèle présenté dans le texte.
Deux placettes forestières sont tirées dans la même communauté. H_β n'est pas nulle à cause des différences stochastiques entre les tirages. Les barres verticales sont les 5ème et 95ème centiles. La première colonne correspond à des placettes de 500 arbres environ, la seconde de 5 000 arbres, la première ligne à 20 espèces, la seconde à 40. Toutes choses égales par ailleurs, H_β décroît avec le nombre d'arbres et croît avec le nombre d'espèces.

Les résultats se trouvent en Figure 37.

La valeur de $\widehat{H}_\beta$ calculée entre deux placettes ne change pas si les effectifs sont multipliés par 10 sans changer les fréquences. Mais l'hypothèse nulle du test est que les deux placettes sont issues de la même communauté : quand plus d'individus sont échantillonnés, les fréquences observées convergent vers leur probabilité à cause de la loi des grands nombres. Une valeur observée de $\widehat{H}_\beta = 0.002$ montre une différence significative entre deux placettes de 5 000 arbres (Figure 37, en haut à droite), mais si les placettes ne contiennent que 500 individus (en haut à gauche), cette valeur est en dessous de la borne inférieure de l'intervalle de confiance et indique que les placettes sont probablement trop similaires pour que ce soit simplement le résultat du hasard.

$\widehat{H}_\beta$ tend à augmenter avec le nombre d'espèces. La borne supérieure de l'intervalle de confiance avec 20 espèces (en haut) correspond approximativement à la borne inférieure avec 40 espèces (en bas).

Données réelles

Le test est appliqué à des données réelles, quatre placettes de 1 ha de forêt tropicale sur les dispositifs forestiers de Paracou (Gourlet-Fleury *et al.*, 2004) et des Nouragues (Bongers *et al.*, 2001), en Guyane française. Pour la clarté du raisonnement, nous admettrons que ces placettes représentent des environnements contrastés et constituent ensemble un échantillon représentatif de chaque forêt.

Les caractéristiques de chaque placette sont résumées dans le Tableau 8.

Placette	*NH20*	*NL11*	*P006*	*P018*	*Total*
Nombre d'arbres	558	515	643	481	2197
Nombre d'espèces	203	182	147	149	425
$\widehat{H}$	4,74	4,63	4,19	4,42	5,29
Nombre de Hill	114	103	66	83	199

Tableau 8 : Résumé des quatre placettes de 1 hectare. Les deux premières sont situées aux Nouragues, les deux dernières à Paracou. $\widehat{H}$ est la diversité α de Shannon.

Le premier résultat est que les placettes des Nouragues sont plus diverses que celles de Paracou. Les nombres de Hill donnent une représentation intuitive du niveau de diversité : par exemple, la placette NH20 est aussi diverse que le serait une placette de taille identique avec 114 espèces de fréquence égale, alors que la placette P006 l'est à peu près deux fois moins.

Le Tableau 9 montre comment la diversité peut être décomposée entre les forêts puis entre les placettes ou regroupée dans l'autre sens. Les valeurs de $\widehat{H}_\beta$ peuvent

être testées contre l'hypothèse nulle d'absence de différence entre les placettes ou les forêts. Par exemple, la diversité β entre les deux placettes des Nouragues attendue sous l'hypothèse nulle est 0,144, ce qui correspond à un nombre de Hill de 1,16 (Figure 38).

Placette	***NH20***	***NL11***	***P006***	***P018***
$H_{placettes}$ ***pondéré***	2,46	2,22	2,40	1,89
$H_{\beta}^{placettes}$	0,42 (N=1,52)		0,45 (N=1.56)	
$H_{forêt}$	5,11 (N=165)		4.74 (N=114)	
$H_{forêt}$ ***pondéré***	2,49		2.42	
$H_{\beta}^{forêt}$	0,38 (N=1,46)			
H_{total}	5,29 (N=199)			

Tableau 9 : Regroupement successif des placettes des Nouragues et de Paracou. La première ligne contient la diversité α des placettes, pondérée par le nombre d'arbres. La deuxième ligne contient les valeurs de diversité β entre les placettes. La somme des deux donne la diversité γ de la forêt (ligne 3). À son tour, celle-ci peut être considérée comme la diversité α au niveau de regroupement supérieur. Sa valeur pondérée (ligne 4) est ajoutée à celle de la diversité β entre forêts (ligne 5) pour donner la diversité totale (ligne 6).

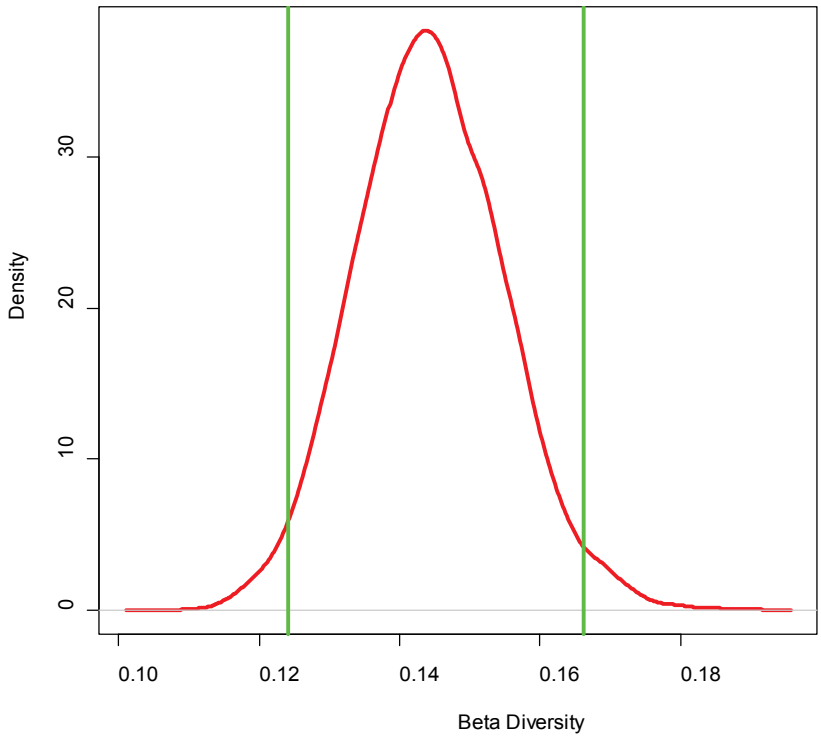


Figure 38: Densités de probabilité de H_{β} sous l'hypothèse nulle pour les placettes des Nouragues. Les traits verticaux sont les bornes de l'intervalle de confiance.

La plage possible des nombres de Hill va de 1 (distribution des fréquences exactement identique) à 2 (nombre d'arbres identiques sans aucune espèce en com-

mun). Les valeurs inférieures à 0,1 (N=1,10) ou supérieures à 0,2 (N=1,22) ont une probabilité si faible qu'elles peuvent être considérées comme jamais atteintes si les deux placettes proviennent de la même communauté. La valeur réelle observée aux Nouragues est 0,42, très au-dessus de l'intervalle de confiance. Toutes les valeurs de diversité β du tableau sont hautement significatives (au-delà de 99,99%). On peut observer que la diversité à l'intérieur des forêts est similaire à celle entre les forêts (tous les nombres de Hill autour de 1,5). Les placettes auraient pu être groupées directement. Dans ce cas, la diversité entre placettes aurait été de 0,81 (significative au-delà de 99,99%). Le nombre de Hill correspondant est 2,25, ce qui signifie que les 4 placettes sont aussi différentes que 2,25 placettes de même effectif sans espèces communes.

CONCLUSION

Les informations qu'un écologue cherche à retirer immédiatement de l'observation d'un semis de points représentant une communauté végétale sont les densités de chaque espèce et l'existence ou non d'interactions entre les plantes. Ceci est en substance l'introduction de l'article de Pielou (1959) et la motivation des développements méthodologiques présentés ici.

Caractériser

Les méthodes ont largement progressé depuis l'article de Pielou, avec un saut qualitatif net correspondant à l'introduction de la fonction K par Ripley (1976; 1977). Des revues ou des manuels sont régulièrement produits pour permettre aux praticiens de comprendre et utiliser les nouveaux outils (Perry *et al.*, 2002 ; Fortin et Dale, 2005 ; Law *et al.*, 2009). En amont, littérature sur les processus ponctuels s'est enrichie : une synthèse de l'état de l'art peut être trouvée dans Møller et Waagepetersen (2004). Entre cette approche théorique et les manuels pour praticiens se trouve une place pour une approche assez rigoureuse sur le plan mathématique mais orientée vers les applications empiriques (Diggle, 1983 ; Stoyan *et al.*, 1987 ; Cressie, 1993 ; Illian *et al.*, 2008 par exemple). L'objectif de ce travail était de proposer des avancées méthodologiques dans cet esprit.

Une façon de synthétiser les nombreuses approches passées en revue est une typologie des outils, dans le but de montrer leur cohérence plus que de fournir un guide pratique (on se référera pour cela à la liste plus haut). L'accent est mis ensuite sur le problème nommé « MAUP » par les économistes, qui affecte les statistiques discrètes : la délimitation des zones peut créer ou masquer des structures sans rapport avec le phénomène étudié, ce qui constitue une motivation supplémentaire pour traiter des données spatialisées.

Synthèse

Les indices peuvent être regroupés selon trois critères :

- Tout d'abord, leur champ d'application : en espace continu ou sur des quadrats

- Ensuite, selon leur appréhension de l'espace (le vocabulaire est celui de Brülhart et Traeger, 2005) : les indices absolus qui ne concernent que les quadrats considèrent que chacun doit contenir le même nombre d'objets et mesurent les écarts à cette égalité, les indices topographiques considèrent que le nombre d'objets ou leur poids doit être proportionnel à la surface et enfin les indices relatifs utilisent une autre référence que la surface, par exemple la surface terrière totale des arbres dans une parcelle.
- Enfin, selon leur valeur de référence : les moins élaborés fournissent un résultat à comparer à deux valeurs extrêmes correspondant à une équirépartition parfaite, c'est-à-dire un maximum de dispersion, et une concentration maximum de tous les objets en un lieu unique. Les autres permettent de comparer le résultat à une hypothèse nulle, généralement l'indépendance entre les points, accompagnée d'un intervalle de confiance permettant de conclure à la significativité de la structure observée.

Les tableaux ci-dessous résument les types d'indices :

- Fondés sur les quadrats :

Référence / ***Type d'indice***	***Extrêmes***	***Hypothèse nulle***
Absolu	Indice d'Herfindahl (page 191)	-
Topographique	Indices d'entropie généralisée GE(α)	Variance relative (page 185) Méthode de Greig-Smith (page 186)
Relatif	Indice de Gini (page 188) Indice *G* d'Ellison et Glaeser (page 189) Indices d'entropie généralisée GE(α)	Indice *γ* d'Ellison et Glaeser (page 189) MTAD de Rysman et Greenstein (page 194)

- En espace continu :

Référence / ***Type d'indice***	***Extrêmes***	***Hypothèse nulle***
Absolu	-	-

Référence *Type d'indice*	*Extrêmes*	*Hypothèse nulle*
Topographique	-	Indice *R* d'Evans et Clark (page 167) Fonctions *F* et *G* de Diggle (page 168) Fonctions *J* de Van Lieshout et Baddeley (page171) *g et K* de Ripley, *L* de Besag g_{inhom} et K_{inhom} de Baddeley *et al.* *O-ring* de Wiegand *et al.*
Relatif	-	T_k de Cuzick et Edwards K_d de Duranton et Overman *D* de Diggle et Chetwynd *M* de Marcon et Puech

La MAUP

Toutes les méthodes fondées sur les quadrats nécessitent un découpage de l'espace *a priori*. En foresterie, on dispose couramment de données par parcelle, en économie de données par unités administratives plus ou moins détaillées. Or, toutes les mesures sont sensibles à l'échelle géographique retenue.

Ce résultat est connu sous le nom du Problème des Unités Spatiales Modifiables (*Modifiable Areal Unit Problem* – MAUP) dont le terme a été introduit par Openshaw et Taylor (1979). Une littérature abondante s'est développée à ce sujet en économie (par exemple Arbia, 1989 ; Fotheringham et Wong, 1991 ; Amrhein, 1995 ; Morphet, 1997). Le problème est résumé par Morphet (1997, page 1039) : « *le résultat sera sensible à la forme, à la taille et à la position des unités spatiales choisies* »[1]. Les manifestations de la MAUP sont essentiellement le problème d'échelle et le problème d'agrégation.

Le problème d'échelle est dû à la résolution du zonage choisi. Pour l'expliquer, considérons une distribution de points sur un territoire, Figure 39 (-a-), et deux découpages possibles : en 4 zones (-b-), puis 8 zones par redécoupage (-c-).

[1] « *The result will be sensitive to the shape, size, and position of the areal units chosen* ».

Une certaine hétérogénéité est détectée dans la figure c alors que la figure b montre une parfaite régularité de la distribution (1 point par zone). Le découpage en b est trop grossier et masque l'hétérogénéité à petite échelle. Le problème inverse est également possible : un découpage trop détaillé peut masquer la concentration en découpant les agrégats.

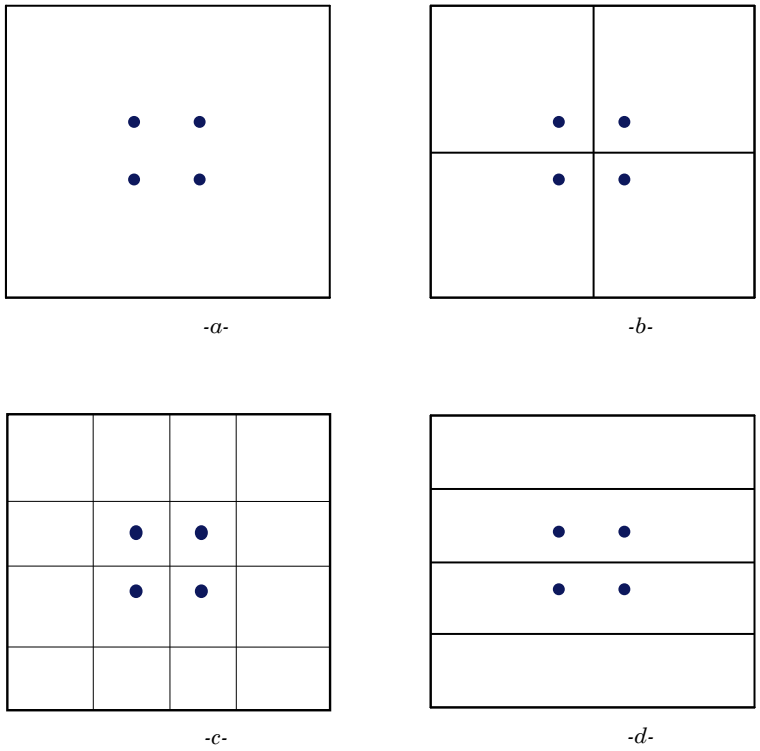


Figure 39 : Illustration du problème d'échelle et d'agrégation

Le problème d'agrégation se manifeste par des résultats différents selon le découpage géographique. Ce problème est illustré sur la Figure 39 par les cas b et d. Chaque territoire est divisé en 4 zones de mêmes tailles, mais sur la figure b la distribution est régulière alors que sur la figure d, deux quadrats n'ont aucun point. Quel que soit l'indice utilisé, l'agrégation sera plus grande en d.

Comme le souligne Arbia (2001, page 413) : « *nous pouvons facilement imaginer que la situation soit encore pire dans les cas réels, où les unités spatiales sont irrégulières en taille et en forme* »[2]. Ainsi, tous les indices reposant sur un découpage de l'espace doivent se soumettre aux critiques de la MAUP si le découpage n'est pas lié au phénomène observé ; en d'autres termes si le découpage est arbitraire (parcelles forestières) ou sans rapport (découpage administratif). Une précaution possible, mise en place par Barrios *et al.* (2003) consiste à vérifier l'absence

[2] « *We can easily imagine that the situation is even worse in real cases, where the spatial units are irregular in size and shape* »

d'autocorrélation spatiale (voir page 195) entre les zones et changer d'échelle si nécessaire : évaluer la concentration spatiale au niveau des régions si les départements sont autocorrélés par exemple.

Obtenir des informations sur les processus écologiques

Watt (1947) a été apparemment le premier à mettre en rapport clairement les processus écologiques et les structures spatiales observées en décrivant des successions de stades caractéristiques de communautés, responsables de structures caractéristiques observables. Parcourir le chemin inverse, en partant des structures pour en déduire les processus, est tentant mais risqué.

Pielou (1962) utilise la méthode de détection des structures spatiales développée dans un article précédent (Pielou, 1959) pour détecter la **compétition** à courte distance entre les arbres, entraînant une certaine régularité de la distribution spatiale. Cette approche a été largement reprise depuis : un semis de points est analysé par des méthodes non paramétriques en général pour détecter un processus écologique attendu : la compétition éloigne les plantes et donne des distributions plus régulières, la limitation de la dispersion crée des distribution agrégées, la concurrence ou la facilitation interspécifique est détectée par des attractions ou des répulsions intertypes (Szwagrzyk, 1990 ; Duncan, 1991 ; Szwagrzyk et Czerwczak, 1993 ; Wei et Skarpe, 1995 ; Haase *et al.*, 1996 ; Kuuluvainen *et al.*, 1996 ; Haase *et al.*, 1997 ; Martens *et al.*, 1997 ; Wiegand *et al.*, 1998 ; Nanami *et al.*, 1999 ; McDonald *et al.*, 2003 ; Wang *et al.*, 2003). Aldrich *et al.* (2003) disposent de 60 années de recul qui leur permettent de mettre en évidence l'évolution au cours du temps de la structure spatiale d'une forêt vieillissante.

La structure spatiale de la **mortalité** est souvent étudiée pour en comprendre les causes (Sterner *et al.*, 1986 ; Kenkel, 1988 ; Rebertus *et al.*, 1989 ; Fulé et Covington, 1998 ; Cole et Syms, 1999 qui recherchent la cause de la mortalité d'algues ; He et Duncan, 2000) : le rejet de l'hypothèse nulle d'étiquetage aléatoire montre qu'un processus non totalement aléatoire entre en jeu (Goreaud et Pélissier, 2003).

Les études sont nombreuses (la liste des références citées ne concerne que des applications de K et n'est pas exhaustive) mais l'approche est presque toujours descriptive et qualitative, sans moyen de tester si l'hypothèse écologique appuyée par la structure spatiale (caractérisée quant à elle avec certitude) est bien la

bonne. L'étape suivante est clairement la modélisation explicite de la structure d'un peuplement et la vérification de la validité du modèle sur les données.

Inférer

L'essentiel des méthodes permettant de caractériser la structure spatiale est non paramétrique. Une statistique est comparée à une hypothèse nulle et fournit donc un test. L'ajustement des données à un modèle connu, c'est-à-dire l'inférence, est très peu développée en écologie. Les outils mathématiques nécessaires à l'inférence des processus courants existent même s'ils sont compliqués (Møller et Waagepetersen, 2004), et certains sont disponibles dans R, dans le module Spatstat (Baddeley et Turner, 2005), ce qui en rend l'usage accessible aux non spécialistes.

Quelques exemples sont passés en revue ici, pour montrer l'intérêt de cette approche. Le premier est l'inférence d'un processus de Thomas qui décrit bien un peuplement dont les centres des agrégats sont les pieds-mères, qui dispersent leurs graine selon une distribution gaussienne. Une description détaillée permet de comprendre la démarche.

Estimation d'un processus de Thomas

Le processus de Thomas (*cf.* page 144) est utilisé par Plotkin *et al.* (2000) avec de bons résultats, jugés surprenant par les auteurs étant donnée sa simplicité. Diggle (1983 p. 75) calcule la valeur théorique de K pour un tel processus de densité d'agrégats ρ et dispersion σ :

$$K_{NS}(r) = \pi r^2 + \frac{1 - e^{\frac{-r^2}{4\sigma^2}}}{\rho} \quad \textbf{(91)}$$

Connaissant les valeurs estimées de K pour un semis de points réel, on peut estimer les paramètres ρ et σ du processus de Neyman-Scott sous-jacent supposé en minimisant (Diggle, 1983 p. 74) :

$$\Delta = \int_0^R (K(r)^c - K_{NS}(r)^c)^2 dr \quad \textbf{(92)}$$

R est le rayon maximum pris en compte. Diggle estime que l'ajustement est peu sensible à sa valeur. c est une constante permettant d'atténuer le fait que les va-

leurs de K augmentent rapidement avec r : en leur affectant une puissance inférieure à 1, les petits écarts sont moins écrasés par les grands. Diggle conseille d'utiliser une valeur de l'ordre de 0,5 pour des semis de points réguliers et 0,25 pour des processus agrégés, valeur reprise par Plotkin *et al.*

Comme les valeurs de K estimées sont discrètes, l'ajustement est réalisé en pratique en calculant pour chaque valeur de r la valeur de K_{ns} correspondante à partir de valeurs arbitraires des paramètres. La somme des carrés des écarts entre $K(r)^c$ et $K_{NS}(r)^c$, pondérés par le pas de calcul entre r et sa valeur précédente, est la valeur à minimiser en modifiant ρ et σ. Plotkin *et al.* (2000) utilisent comme valeurs de départ l'inverse de la plus grande valeur de K et le quart de R.

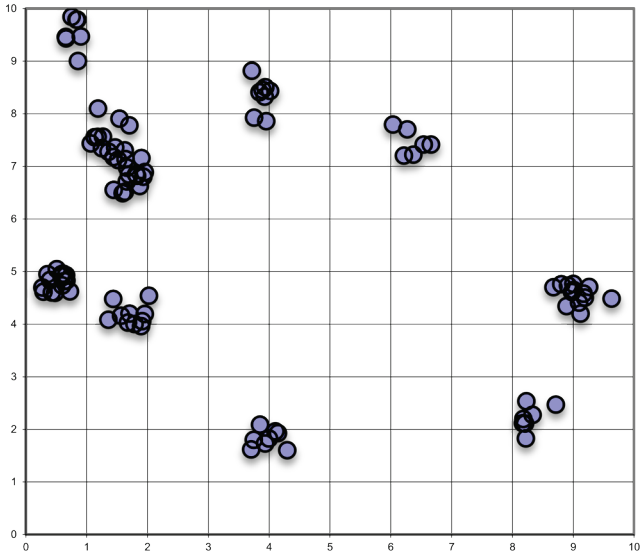


Figure 40 : Simulation d'un processus de Neyman Scott, Carte des points

L'ajustement des paramètres peut donc être réalisé simplement par la méthode de Newton, par exemple avec le solveur d'Excel. Un exemple est donné par la Figure 40 : un Processus de Neyman-Scott gaussien de paramètres $N_2 = 100$ (nombre de points), $\rho = 0{,}1$ (densité des agrégats) et $\sigma = 0{,}2$ (écart-type de l'intensité de probabilité) est simulé. On observe 10 agrégats (dont deux coalescents en haut à gauche de la figure) de 10 points en moyenne, 99% environ des points se trouvant dans un rayon de $3\sigma = 0{,}6$ autour du centre des agrégats. La fonction K a été calculée par pas de 0,2 jusqu'à $R = 3$, et l'estimation des paramètres effectuée. Les paramètres trouvés sont $\rho = 0{,}09$ et $\sigma = 0{,}22$. La Figure 41 montre

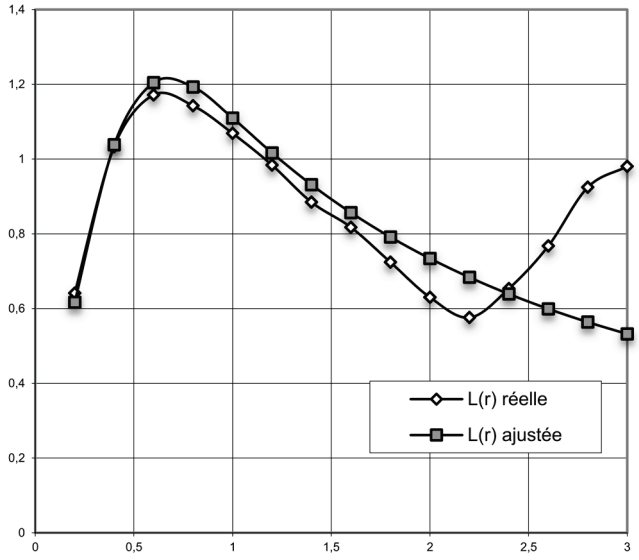


Figure 41 : Ajustement des courbes *L* du semis de point réel et calculée à partir des paramètres ajustés.

l'ajustement entre la courbe L (plus lisible que K) du semis de points réel et les valeurs calculées par l'équation (91) à partir des paramètres ajustés. Les ajustements sont efficaces malgré le pic secondaire de la courbe L due à la coalescence de deux agrégats. En limitant R à 2, les paramètres sont retrouvés avec une erreur inférieure à 1%.

Autres ajustements de modèles

Diggle et Rowlingson (1994) proposent un processus de Poisson hétérogène pour décrire les cas de maladie près d'une source de risque potentielle (l'exemple présenté concerne les cas d'asthme autour de trois usines en Grande-Bretagne). L'intensité du processus est fonction de la distance de la source et d'un paramètre à estimer par le maximum de vraisemblance. La valeur du paramètre, et surtout sa significativité, montrent l'importance de la source.

Kenkel (1993) analyse l'inhibition entre les individus d'une espèce herbacée dans le sous-bois d'une forêt canadienne. Il ajuste un processus ponctuel ad hoc selon deux paramètres : l'intensité de la répulsion et la distance à laquelle elle est sensible.

Dans un contexte différent, à partir de données discrètes, Kim *et al.* (2000) supposent que le nombre d'établissements industriels rencontrés dans chaque région est une réalisation d'une loi binomiale négative dont un paramètre décrit la variabilité de la distribution, donc le niveau d'agglomération, qui peut être traduit directement comme une mesure des externalités positives.

Plotkin *et al.* (2000) remarquent que les courbes aire-espèces prédites par un modèle de placement complètement aléatoire (Coleman, 1981) des arbres sont surestimées par rapport à la réalité, à cause de l'agrégation des espèces. Ils choisissent donc de simuler des peuplements virtuels selon un processus de Poisson agrégé dont les paramètres sont estimés à partir du peuplement réel (et obtiennent alors des courbes aire-espèces correctes).

Dans le même article, ils utilisent la simulation pour tester la significativité de la relation entre la présence d'une espèce et les facteurs du milieu. Les tests classiques comme le χ^2 supposent l'indépendance de la distribution des individus, qui est violée quand une espèce a une structure agrégative.

Dans les deux cas, les outils classiques supposent l'indépendance des individus. La violation de cette hypothèse fausse gravement les résultats. La prise en compte analytique de la dépendance est trop complexe et la simulation (méthode de Monte Carlo) est efficace. L'enjeu est donc de simuler des peuplements réalistes à partir d'un processus vraisemblable dont on aura pu estimer les paramètres à partir des données existantes.

Les processus mimétiques (Goreaud *et al.*, 2004 ; Ngo Bieng, 2007 ; Ngo Bieng *et al.*, In press) sont des processus de Strauss multi-étages et intertypes dans lesquels les paramètres d'interaction de paires de points sont l'écart entre les fonctions *K* du semis de points en cours de simulation et celles du semis de points observé. Les paramètres du processus de Strauss ne sont pas directement inférés, mais la structure du peuplement simulé est similaire à celle du peuplement réel.

Ces méthodes nécessitent avant tout le choix du processus sous-jacent, qui nécessite un modèle théorique ou des connaissances particulières, et une méthode mathématique correcte pour l'estimation du ou des paramètres, développée pour le modèle particulier. Leur généralisation et leur utilisation en routine ne sont donc pas envisageables. En revanche, les résultats sont souvent plus riches d'enseignement si les paramètres sont interprétables, et la possibilité de simuler d'autres distributions à partir des paramètres inférés ouvre la possibilité de tests par la méthode de Monte-Carlo.

Modéliser

Pielou toujours (1960) développe un modèle simple pour mettre en évidence la structure spatiale des peuplements végétaux : des plantes sont placées de façon complètement aléatoire dans un domaine, avec une intensité choisie. Chaque plante reçoit un espace vital circulaire dont le rayon varie aléatoirement dans un intervalle choisi aussi. Les plantes sont tirées séquentiellement et leur espace tracé. Si une plante tombe dans l'espace d'une précédente ou trop près pour installer son propre espace, elle est tirée à nouveau. Ce simple modèle géométrique dans lequel aucune plante ne peut chevaucher l'espace de sa voisine montre que les semis de points obtenus sont à coup presque sûr agrégés : les grosses plantes laissent des espaces remplis par des agrégats de petites. Les distributions régulières n'apparaissent que quand la taille des cercles varie peu (les cercles se rangent alors pour occuper l'espace efficacement) ou, par hasard, quand la densité est faible. Ce dernier cas est rejeté par Pielou comme peu réaliste : sur le terrain, les faibles densités limitent la compétition et génèrent plutôt des distributions complètement aléatoires.

Ces résultats montrent bien la limite de l'approche précédente : la recherche de distributions régulières pour mettre en évidence la compétition est probablement vouée à l'échec si les arbres sont de taille variable, et l'argument de l'agrégation en appui à une hypothèse écologique (dispersion limitée, clonage) est faible parce qu'aucun mécanisme biologique n'est nécessaire.

Ce modèle est repris par Hanus *et al.* (1998) et appliqué à des peuplements de Douglas : à partir du nombre d'arbres d'un peuplement réel et de leurs diamètres, les positions sont simulées. Les tailles des cercles sont choisies proportionnellement à la taille des couronnes des arbres de même taille en croissance libre (le coefficient de proportionnalité est choisi pour que la totalité des arbres puisse entrer dans la surface disponible). Le réalisme des peuplements simulés est validé par la comparaison des fonctions K de Ripley. Les structures spatiales des peuplements réels se trouve dans l'intervalle de confiance de celles des peuplements simulés, ce qui valide la méthode. La conclusion à tirer de cette application empirique est que la géométrie est bien à elle seule un déterminant majeur de la structuration spatiale.

L'utilisation des processus ponctuels en tant que tels comme modèles écologiques n'est probablement pas appropriée : les processus ponctuels fournissent un semis de points instantanément, avec ou sans interactions, alors que les processus écologiques se déroulent dans le temps. Un peuplement forestier dont la distribution spatiale est indiscernable d'un Poisson homogène, permettant de conclure qu'une espèce a une distribution complètement aléatoire, ne donne une information que sur le résultat de la compétition et des autres interactions ayant eu lieu entre les arbres et aboutissant à une distribution équivalente à celle obtenue par un processus ponctuel dans lequel les points n'interagissent pas. Les processus ponctuels ne fournissent pas de modèles mécanistes (Law *et al.*, 2009) mais peuvent permettre de comparer le résultat d'un modèle à un peuplement réel.

Les processus de Markov simulés par naissances et morts de points sont peut-être une voie de recherche intéressante si un processus biologique peut être rapproché de l'algorithme MCMC qui n'est pour l'instant qu'un outil de simulation.

Bilan et Perspectives

Ce travail a apporté quelques résultats : le calcul de l'intervalle de confiance de la fonction K est un outil manquant aux praticiens (Law *et al.*, 2009), la fonction M est un outil de caractérisation des processus inhomogènes dans un cadre relatif utile (cité, y compris dans ses versions préliminaires, par Rysman et Greenstein, 2003 ; Brülhart et Traeger, 2005 ; Fernandez-Gonzalez *et al.*, 2005 ; Maré, 2005 ; Zhu et Chen, 2007 ; Qi *et al.*, 2008 ; Bonneu, 2009 ; Mateu Mahiques *et al.*, 2009 ; Jensen et Michel, in press), même si son intérêt n'est pas évident en écologie forestière, où l'approche topographique est plus naturelle. Pour les données discrètes, l'unification des mesures de diversité et de concentration spatiale, vues sous l'angle de l'entropie, fournit un cadre de travail dont le premier résultat est

la clarification de l'indice β de Shannon. Ce travail n'en est qu'à son début, plus abouti sur les mesures de la diversité que sur celle des structures spatiales. Le champ des recherches méthodologiques est encore vaste.

BIBLIOGRAPHIE

Adelman, M. A. (1969). Comment on the "H" Concentration Measure as a Numbers-Equivalent. *The Review of Economics and Statistics* **51**(1): 99-101.

Aldrich, P. R., Parker, G. R., Ward, J. S. et Michler, C. H. (2003). Spatial dispersion of trees in an old-growth temperate hardwood forest over 60 years of succession. *Forest Ecology and Management* **180**(1-3): 475-491.

Allan, J. D. (1975). Components of Diversity. *Oecologia* **18**(4): 359-367.

Amiti, M. (1999). Specialization Patterns in Europe. *Weltwirtschaftliches* **135**(4): 573-593.

Amrhein, C. G. (1995). Searching for the elusive aggregation effect: evidence from statistical simulations. *Environment and Planning A* **27**(1): 105-119.

Anselin, L. (1995). Local Indicators of Spatial Association -LISA. *Geographical Analysis* **27**: 93-115.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Kluwer, Dordrecht

Arbia, G. (2001). Modelling the Geography of Economic Activities on a Continuous Space. *Papers in Regional Science* **80**(4): 411-420.

Baddeley, A. J., Møller, J. et Waagepetersen, R. P. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54**(3): 329-350.

Baddeley, A. J. et Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* **12**(6): 1-42.

Balassa, B. (1965). Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School of Economic and Social Studies* **33**(2): 99-117.

Baraloto, C., Paine, C. E. T. P., Patiño, S., Bonal, D., Hérault, B. et Chave, J. (2010). Functional trait variation and sampling strategies in species rich plant communities. *Functional Ecology* **24**: 208-216.

Barot, S., Gignoux, J. et Menaut, J.-C. (1999). Demography of a savanna palm tree: predictions from comprehensive spatial pattern analyses. *Ecology* **80**(6): 1987-2005.

Barrios, S., Bertinelli, L., Strobl, E. et Teixeira, A. C. F. (2003). *Agglomeration economies and the location of industries: a comparison of three small European countries*. CORE Discussion Papers, **2003/67**.

Besag, J. E. (1977). Comments on Ripley's paper. *Journal of the Royal Statistical Society* **B 39**(2): 193-195.

Bongers, F., Charles-Dominique, P., Forget, P.-M. et Théry, M., Eds. (2001). *Nouragues: dynamics and plant-animal interactions in a neotropical rainforest*. Biological Monographs Series. Dordrecht, The Netherlands, Kluwer Academic Publisher.

Bonneu, F. (2007). Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process. *Case Studies in Business, Industry and Government Statistics* **1**(2): 139-152.

Bonneu, F. (2009). *Processus poctuels spatiaux pour l'analyse du positionnement optimal et de la concentration*. Ph.D. Thesis, Université de Toulouse I. Toulouse**:** 138.

Bourguignon, F. (1979). Decomposable Income Inequality Measures. *Econometrica* **47**(4): 901-920.

Box, G. E. P. et Muller, M. E. (1958). A Note on the Generation of Random Normal Deviates. *Annals of Mathematical Statistics* **29**: 610-611.

Brülhart, M. et Torstensson, J. (1996). *Regional Integration, Scale Economies an Industry Location in the European Union*. Research Paper, **1435**. Centre for Economic Policy Research, London.

Brülhart, M. et Traeger, R. (2005). An Account of Geographic Concentration Patterns in Europe. *Regional Science and Urban Economics* **35**(6): 597-624.

Chiu, S. N. (2007). Correction to Koen's critical values in testing spatial randomness. *Journal of Statistical Computation and Simulation* **77**(11-12): 1001-1004.

Clapham, A. R. (1936). Over-dispersion in grassland communities and the use of statistical methods in plant ecology. *Journal of Ecology* **24**(1): 232-251.

Clark, P. J. et Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* **35**(4): 445-453.

Cole, R. G. et Syms, C. (1999). Using spatial patterns analysis to distinguish causes of mortality: an example from kelp in north-eastern New Zealand. *Journal of Ecology* **87**(6): 963-972.

Coleman, B. D. (1981). Random placement and species-area relations. *Mathematical Biosciences* **54**: 191-215.

Combes, P.-P. et Overman, H. G. (2004). The spatial distribution of economic activities in the European Union. *in* J. V. Henderson et J.-F. Thisse, (Eds), *Handbook of Urban and Regional Economics*. Elsevier. North Holland, Amsterdam. **4**.

Conceição, P. et Ferreira, P. (2000). *The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications*. UTIP Working Paper, **14**, Austin, Texas**:** 54 p.

Condit, R., Pitman, N., Leigh, E. G., Chave, J., Terborgh, J., Foster, R. B., Nunez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H. C., Losos, E. et Hubbell, S. P. (2002). Beta-diversity in tropical forest trees. *Science* **295**(5555): 666-669.

Coomes, D. A., Rees, M. et Turnbull, L. (1999). Identifying aggregation and association in fully mapped spatial data. *Ecology* **80**(2): 554-565.

Coste, S., Roggy, J.-C., Garraud, L., Heuret, P., Nicolini, E. et Dreyer, E. (2009). Does ontogeny modulate irradiance-elicited plasticity of leaf traits in saplings of rain-forest tree species? A test with Dicorynia guianensis and Tachigali melinonii (Fabaceae, Caesalpinioideae). *Annals of Forest Science* **66**(7): 709.

Cox, D. R. (1955). Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society* **B 17**(2): 129-164.

Cressie, N. A. (1993). *Statistics for spatial data*. John Wiley & Sons, New York. 900 p

Crist, T. O., Veech, J. A., Gering, J. C. et Summerville, K. S. (2003). Partitioning species diversity across landscapes and regions: A hierarchical analysis of alpha, beta, and gamma diversity. *The American Naturalist* **162**(6): 734-743.

Cutrini, E. (2009). Using entropy measures to disentangle regional from national localization patterns. *Regional Science and Urban Economics* **39**(2): 243-250.

Cuzick, J. et Edwards, R. (1990). Spatial Clustering for Inhomogeneous Populations. *Journal of the Royal Statistical Society* **B 52**(1): 73-104.

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal* **30**(119): 348-361.

Davis, H. T. (1941). *The theory of econometrics*. The Principia Press, Bloomington, Indiana

Diggle, P. J. (1976). Note on the Clark and Evans test of spatial randomness. *in* I. Hodder et C. Orton, (Eds), *Spatial Analysis in Archaeology*. Cambridge University Press, London.

Diggle, P. J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics* **35**: 87-101.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London. 148 p.

Diggle, P. J. (1985). A Kernel Method for Smoothing Point Process Data. *Applied Statistics* **34**(2): 138-147.

Diggle, P. J. (1990). Discussion of the paper by Cuzick and Edwards. *Journal of the Royal Statistical Society* **B 52**(1): 101.

Diggle, P. J. et Chetwynd, A. G. (1991). Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations. *Biometrics* **47**: 1155-1163.

Diggle, P. J. et Rowlingson, B. S. (1994). A Conditional Approach to Point Process Modeling of Elevated Risk. *Journal of the Royal Statistical Society* **A 157**(3): 433-440.

Duncan, R. P. (1991). Competition and the coexistence of species in a mixed podocarp stand. *Journal of Ecology* **79**(4): 1073-1084.

Duranton, G. et Overman, H. G. (2005). Testing for Localisation Using Micro-Geographic Data. *Review of Economic Studies* **72**(4): 1077-1106.

Eaton, B. C. et Lipsey, R. G. (1982). An Economic Theory of Central Places. *Economic Journal* **92**: 56-72.

Ellison, G. et Glaeser, E. L. (1997). Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy* **105**(5): 889-927.

Fehmi, J. S. et Bartolome, J. W. (2001). A grid-based method for sampling and analysing spatially ambiguous plants. *Journal of Vegetation Science* **12**(4): 467-472.

Feller, W. (1943). On a general class of contagious distributions. *Annals of Mathematical Statistics* **14**: 389-400.

Fernandez-Gonzalez, R., Barcellos-Hoff, M. H. et de Solorzano, C. O. (2005). A Tool for the Quantitative Spatial Analysis of Complex Cellular Systems. *IEEE Transactions on Image Processing* **14**(9): 1300-1313.

Feser, E. J. et Sweeney, S. H. (2000). A test for the coincident economic and spatial clustering of business enterprises. *Journal of Geographical Systems* **2**(4): 349-373.

Feser, E. J. et Sweeney, S. H. (2002). Theory, methods, and a cross-metropolitan comparison of business clustering. *in* P. McCann, (Eds), *Industrial Location Economics*. Edward Elgar, Cheltenham.

Flores, O., Gourlet-Fleury, S. et Picard, N. (2006). Local disturbance, forest structure and dispersal effects on sapling distribution of light-demanding and shade-tolerant species in a French Guianian forest. *Acta Oecologica* **29**(2): 141-154.

Fortin, M.-J. et Dale, M. R. T. (2005). *Spatial Analysis. A guide for ecologists.* Cambridge University Press, Cambridge

Fotheringham, A. S. et Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* **23**: 1025-1044.

Fulé, P. Z. et Covington, W. W. (1998). Spatial patterns of Mexican pine-oak forests under different recent fire regimes. *Plant Ecology* **134**: 197-209.

Gatrell, A. C. et Bailey, T. C. (1996). Interactive Spatial Data Analysis in Medical Geography. *Social Science & Medicine* **42**(6): 843-855.

Gatrell, A. C., Bailey, T. C., Diggle, P. J. et Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers* **21**: 256-274.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statisitician* **5**(3): 115-145.

Getis, A. (1984). Interaction modeling using second-order analysis. *Environment and Planning A* **16**: 173-183.

Getis, A. et Franklin, J. (1987). Second-order neighborhood analysis of mapped point patterns. *Ecology* **68**: 473-477.

Gignoux, J., Duby, C. et Barot, S. (1999). Comparing the performances of Diggle's test of spatial randomness for small samples with or without edge effect correction: application to ecological data. *Biometrics* **55**: 156-164.

Gini, C. (1913). *Sulla misura della concentrazione e della variabilità dei caratteri*. Atti del R. Istituo Veneto, **73**.

Goreaud, F. (2000). *Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes*. Ph.D. Thesis, ENGREF. Nancy.

Goreaud, F., Courbaud, B. et Collinet, F. (1997). Spatial structure analysis applied to modelling of forest dynamics: a few examples. *in* A. Amaro et M. Tomé, (Eds), *IUFRO Workshop: Empirical and process-based models for forest tree and stand growth simulation*. Novas Tecnologias, Oeiras, Portugal**:** 155-172.

Goreaud, F., Loreau, M. et Millier, C. (2002). Spatial structure and the survival of an inferior competitor: a theoretical model of neighbourhood competition in plants. *Ecological Modelling* **158**(1-2): 1-19.

Goreaud, F., Loussier, B., Ngo Bieng, M.-A. et Allain, R. (2004). *Simulating realistic spatial structure for forest stands: a mimetic point process*. Interdisciplinary Spatial Statistics Workshop 2004, Paris, December 2-3, 2004.

Goreaud, F. et Pélissier, R. (1999). On explicit formulas of edge-effect correction for Ripley's K-function. *Journal of Vegetation Science* **10**(3): 433-438.

Goreaud, F. et Pélissier, R. (2000). Spatial Structure Analysis of Heterogeneous Point Patterns: examples of application to forest stands. *ADS in ADE-4***:** 49 p.

Goreaud, F. et Pélissier, R. (2003). Avoiding misinterpretation of biotic interactions with the intertype K_{12} fonction: population independence *vs* random labelling hypotheses. *Journal of Vegetation Science* **14**: 681-692.

Gourlet-Fleury, S., Comu, G., Jesel, S., Dessard, H., Jourget, J. G., Blanc, L. et Picard, N. (2005). Using models to predict recovery and assess tree species vulnerability in logged tropical forests: A case study from French Guiana. *Forest Ecology and Management* **209**(1-2): 69-86.

Gourlet-Fleury, S., Guehl, J. M. et Laroussinie, O., Eds. (2004). *Ecology & management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana*. Paris, Elsevier.

Greig-Smith, P. (1952). The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany* **16**(62): 293-316.

Haase, P. (1995). Spatial pattern analysis in ecology based on Ripley's K function: Introduction and methods of edge correction. *Journal of Vegetation Science* **6**(4): 575-582.

Haase, P. (2001). Can isotropy vs. anisotropy in the spatial association of plant species reveal physical vs. biotic facilitation? *Journal of Vegetation Science* **12**(1): 127-136.

Haase, P., Pugnaire, F. I., Clark, S. C. et Incoll, L. D. (1996). Spatial patterns in a two-tiered semi-arid shrubland in southeastern Spain. *Journal of Vegetation Science* **7**(4): 527-534.

Haase, P., Pugnaire, F. I., Clark, S. C. et Incoll, L. D. (1997). Spatial pattern in Anthyllis cytisoides shrubland on abandoned land in southeastern Spain. *Journal of Vegetation Science* **8**(5): 627-634.

Haggett, P., Cliff, A. D. et Frey, A. E. (1977). *Locational analysis in human geography*. E. Arnold, London. 605 p

Hanus, M. L., Hann, D. W. et Marshall, D. D. (1998). Reconstructing the spatial pattern of trees from routine stand examination measurements. *Forest Science* **44**(1): 125-133.

Hart, P. E. (1971). Entropy and Other Measures of Concentration. *Journal of the Royal Statistical Society* **A 134**(1): 73-85.

He, F. et Duncan, R. P. (2000). Density-dependent effects on tree survival in an old-growth Douglas fir forest. *Journal of Ecology* **88**(4): 676-688.

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**(2): 427-432.

Hoel, P. G. (1943). On Indices of Dispersion. *The Annals of Mathematical Statistics* **14**(2): 155-162.

Holmes, T. J. et Stevens, J. J. (2004). Spatial Distribution of Economic Activities in North America. *in* J. V. Henderson et J.-F. Thisse, (Eds), *Handbook of Urban and Regional Economics*. Elsevier. North Holland, Amsterdam.

Hoover, E. M. (1936). The Measurement of Industrial Localization. *The Review of Economic Statistics* **18**(4): 162-171.

Houdebine, M. (1999). Concentration Géographique des Activités et Spécialisation des Départements Français. *Economie et Statistique* **326-327**(6-7): 189-204.

Hubert, L. J., Golledge, R. G. et Costanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis* **13**: 224-233.

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology* **52**(4): 577-586.

Ihaka, R. et Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**: 299-314.

Illian, J., Penttinen, A., Stoyan, H. et Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley-Interscience, Chichester. 534

Jaouen, G., Fournier, M. et Almeras, T. (2010). Thigmomorphogenesis versus light in biomechanical growth strategies of saplings of two tropical rain forest tree species. *Annals of Forest Science* **67**(2): 211.

Jensen, P. et Michel, J. (in press). Measuring spatial dispersion: exact results on the variance of random spatial distributions. *The Annals of Regional Science*.

Jolles, A. E., Sullivan, P. J., Alker, A. P. et Harvell, C. D. (2002). Disease transmission of aspergillosis in sea fans: Inferring process from spatial pattern. *Ecology* **83**(9): 2373-2378.

Jones, A. P., Langford, I. H. et Bentham, G. (1996). The Application of K-Function Analysis to the Geographical Distribution of Road Traffic Accident Outcomes in Norfolk, England. *Social Science & Medicine* **42**(6): 879-885.

Jost, L. (2006). Entropy and diversity. *Oikos* **113**(2): 363-375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* **88**(10): 2427-2439.

Jost, L., DeVries, P., Walla, T., Greeney, H., Chao, A. et Ricotta, C. (2009). Partitioning diversity for conservation analyses. *Diversity and Distributions* **16**(1): 65-76.

Jurasinski, G., Retzer, V. et Beierkuhnlein, C. (2009). Inventory, differentiation, and proportional diversity: a consistent terminology for quantifying species diversity. *Oecologia* **159**(1): 15-26.

Kelly, M. et Meentemeyer, R. K. (2002). Landscape dynamics of the spread of Sudden Oak Death. *PE&RS, Photogrammetric Engineering & Remote Sensing* **68**(10): 1001-1009.

Kempton, R. A. et Taylor, L. R. (1976). Models and statistics for species diversity. *Nature* **262**(5571): 818-820.

Kenkel, N. C. (1988). Pattern of Self-Thinning in Jack Pine: Testing the Random Mortality Hypothesis. *Ecology* **69**(4): 1017-1024.

Kenkel, N. C. (1993). Modeling Markovian Dependence in Populations of *Aralia Nudicaulis*. *Ecology* **74**(6): 1700-1706.

Keylock, C. J. (2005). Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. *Oikos* **109**(1): 203-207.

Kim, S. (1995). Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U.S. Regional Manufacturing Structure, 1860-1987. *The Quarterly Journal of Economics* **110**(4): 881-908.

Kim, Y., Barkley, D. L. et Henry, M. S. (2000). Industry characteristics linked to establishment concentrations in nonmetropolitan areas. *Journal of Regional Science* **40**(2): 231-259.

Kingham, S. P., Gatrell, A. C. et Rowlingson, B. S. (1995). Testing for Clustering of Health Events within a Geographical Information System Framework. *Environment and Planning A* **27**(5): 809-821.

Koen, C. (1991). Approximate confidence bounds for Ripley's statistic for random points in a square. *Biometrical Journal* **33**: 173-177.

Krugman, P. (1991). *Geography and Trade*. MIT Press, London

Kullback, S. et Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**(1): 79-86.

Kuuluvainen, T., Penttinen, A., Leinonen, K. et Nygren, M. (1996). Statistical opportunities for comparing stand structural heterogeneity in managed and primeval forests: an example from boreal spruce forest in southern Finland. *Silva Fennica* **30**(2-3): 315-328.

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oïkos* **76**: 5-13.

Lang, G. et Marcon, E. (2010). Testing randomness of spatial point patterns with the Ripley statistic. *ArXiv e-prints* **1006.1567**.

Law, R., Illian, J., Burslem, D., Gratzer, G., Gunatilleke, C. V. S. et Gunatilleke, I. (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology* **97**(4): 616-628.

Loreau, M. (2000). Are communities saturated? On the relationship between alpha, beta and gamma diversity. *Ecology Letters* **3**(2): 73-76.

Lorenz, M. O. (1905). Methods of Measuring the concentration of Wealth. *Quarterly Publications of the American Statistical Association* **9**: 209-219.

Lotwick, H. W. et Silverman, B. W. (1982). Methods for Analysing Spatial Processes of Several Types of Points. *Journal of the Royal Statistical Society* **44**(3): 406-413.

Ludovisi, A. et Taticchi, M. I. (2006). Investigating beta diversity by Kullback-Leibler information measures. *Ecological Modelling* **192**(1-2): 299-313.

Maasoumi, E. (1993). A compendium to information theory in economics and econometrics. *Econometric Reviews* **12**(2): 137-181.

MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews* **40**(4): 510-533.

Marcon, E. (2003). A note on "Industry characteristics linked to establishment concentrations in nonmetropolitan areas" confirming the validity of Ellison and Glaeser's index. *mimeo*.

Marcon, E., Hérault, B., Baraloto, C. et Lang, G. (in prep). The Decomposition of Shannon's Entropy and a Test for Beta Diversity.

Marcon, E. et Puech, F. (2002). *A New Method to Evaluate Spatial Economic Activity and Its Application to Two French Areas*. Cahiers de la MSE, **2002.37:** 22 p.

Marcon, E. et Puech, F. (2003). Evaluating the Geographic Concentration of Industries Using Distance-Based Methods. *Journal of Economic Geography* **3**(4): 409-428.

Marcon, E. et Puech, F. (2009). Generalizing Ripley's K function to inhomogeneous populations. *HAL* **halshs-00372631**.

Marcon, E. et Puech, F. (2010). Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods. *Journal of Economic Geography* **10**(5): 745-762.

Marcon, E., Traissac, S. et Lang, G. (in prep.). A Global Test for Ripley's K Function Poisson Null Hypothesis Rejection.

Maré, D. C. (2005). *Concentration, Specialisation and Agglomeration of Firms in New Zealand*. Motu Working Paper, **05-12**.

Marshall, A. (1890). *Principle of Economics*. Macmillan, London

Martens, S. N., Breshears, D. D., Meyer, C. W. et Barnes, F. J. (1997). Scales of above-ground and below-ground competition in a semi-arid woodland detected from spatial pattern. *Journal of Vegetation Science* **8**(5): 655-664.

Matérn, B. (1960). Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut* **49**(5): 1-144.

Mateu Mahiques, J., Albert Ortiz, J. M., Comas Rodríguez, C., Orts Ríos, V., Pernias Cerrillo, J. C. et Porcu, E. (2009). *Stochastic Processes for Spatial Econometrics*. Netbiblo, Oleiros, Spain

Matheron, G. (1970). *La théorie des variables régionalisées et ses applications*. Cahiers du centre de morphologie mathématique de Fontainebleau, **Fascicule 5**, Fontainebleau, France**:** 212 p.

McDonald, R. I., Peet, R. K. et Urban, D. L. (2003). Spatial pattern of *Quercus* regeneration limitation and *Acer rubrum* invasion in a Piedmont forest. *Journal of Vegetation Science* **14**(3): 441-450.

Moeur, M. (1993). Characterizing spatial patterns of trees using stem-mapped data. *Forest Science* **39**(4): 756-775.

Møller, J. et Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes 100*. Chapman and Hall. 300 p

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37**: 17-23.

Mori, T., Nishikimi, K. et Smith, T. E. (2005). A Divergence Statistic for Industrial Localization. *The Review of Economic Statistics* **87**(4): 635-651.

Morphet, C. S. (1997). A statistical method for the identification of spatial clusters. *Environment and Planning A* **29**(6): 1039-1055.

Nanami, S., Kawaguchi, H. et Yamakura, T. (1999). Dioecy-Induced Spatial Patterns of Two Codominant Tree Species, *Podocarpus nagi* and *Neolitsea aciculata*. *Journal of Ecology* **87**(4): 678-687.

Neyman, J. et Scott, E. L. (1958). Statistical Approach to Problems of Cosmology. *Journal of the Royal Statistical Society* **B 20**(1): 1-43.

Ngo Bieng, M.-A. (2007). *Construction de modèles de structure spatiale permettant de simuler des peuplements virtuels réalistes. Application aux peuplements mélangés chêne sessile – pin sylvestre de la région Centre.* PhD Thesis, ENGREF. Paris, France.

Ngo Bieng, M.-A., Ginisty, C. et Goreaud, F. (In press). Point process models for mixed sessile forest stands. *Annals of Forest Science*.

Openshaw, S. et Taylor, P. J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *in* N. Wrigley, (Eds), *Statistical Applications in the Spatial Sciences*. Pion, London**:** 127-144.

Pancer-Koteja, E., Szwagrzyk, J. et Bodziarczyk, J. (1998). Small-scale spatial pattern and size structure of *Rubus hirtus* in a canopy gap. *Journal of Vegetation Science* **9**(6): 755-762.

Patil, G. P. et Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association* **77**(379): 548-561.

Pélissier, R. et Couteron, P. (2007). An operational, additive framework for species diversity partitioning and beta-diversity analysis. *Journal of Ecology* **95**(2): 294-300.

Pélissier, R. et Goreaud, F. (2001). A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation. *Journal of Vegetation Science* **12**(1): 99-108.

Penttinen, A., Stoyan, D. et Henttonen, H. M. (1992). Marked Point Processes in Forest Statistics. *Forest Science* **38**(4): 806-824.

Perry, J. N., Liebhold, A. M., Rosenberg, M. S., Dungan, J., Miriti, M., Jakomulska, A. et Citron-Pousty, S. (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography* **25**(5): 578-600.

Pielou, E. C. (1959). The Use of Point-to-Plant Distances in the Study of the Pattern of Plant Populations. *Journal of Ecology* **47**(3): 607-613.

Pielou, E. C. (1960). A single mechanism to account for regular, random and aggregated populations. *Journal of Ecology* **48**: 575-584.

Pielou, E. C. (1962). The use of plant-to-neighbour distances for the detection of competition. *Journal of Ecology* **50**: 357-367.

Plotkin, J. B., Potts, M. D., Leslie, N., Manokaran, N., LaFrankie, J. V. et Ashton, P. S. (2000). Species-area curves, spatial aggregation, and habitat specialization in tropical forests. *Journal of Theoretical Biology* **207**: 81-99.

Podani, J. et Czárán, T. (1997). Individual-centered analysis of mapped point patterns representing multi-species assemblages. *Journal of Vegetation Science* **8**(2): 259-270.

Porter, P. W. (1960). Earnest and the Orephagians: A Fable of the Instruction of Young Geographers. *Annals of the Association of American Geographers* **50**: 297-299.

Qi, W., Fang, C. et Song, J. (2008). Measurement and spatial distribution of urban agglomeration industrial compactness in China. *Chinese Geographical Science* **18**(4): 291-299.

Qian, H., Ricklefs, R. E. et White, P. S. (2005). Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. *Ecology Letters* **8**(1): 15-22.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org.

Rebertus, A. J., Williamson, G. B. et Moser, E. B. (1989). Fire-Induced Changes in *Quercus Laevis* Spatial Pattern in Florida Sandhills. *Journal of Ecology* **77**(3): 638-650.

Rényi, A. (1961). *On Measures of Entropy and Information*. 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, University of California Press.

Ricotta, C. et Avena, G. (2003). An information-theoretical measure of β-diversity. *Plant Biosystems* **137**(1): 57 - 61.

Ripley, B. D. (1976). The Second-Order Analysis of Stationary Point Processes. *Journal of Applied Probability* **13**: 255-266.

Ripley, B. D. (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society* **B 39**(2): 172-212.

Ripley, B. D. (1979). Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society* **B 41**(3): 368-374.

Ripley, B. D. (1981). *Spatial statistics*. John Wiley & Sons, New York. 255 p.

Rowlingson, B. S. et Diggle, P. J. (1993). SPLANCS: Spatial Point Pattern Analysis Code in S-Plus. *Computers & Geosciences* **19**(5): 627-655.

Rysman, M. et Greenstein, S. (2003). A Note on Testing for Agglomeration and Dispersion. *mimeo*.

Schladitz, K. et Baddeley, A. J. (2000). A third order point process characteristic. *Scandinavian Journal of Statistics* **27**(4): 657-671.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* **27**: 379–423, 623–656.

Shannon, C. E. et Weaver, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press

Shorrocks, A. et Wan, G. (2005). Spatial decomposition of inequality. *Journal of Economic Geography* **5**(1): 59-81.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall. 175 p.

Simpson, E. H. (1949). Measurement of diversity. *Nature* **163**(4148): 688.

Skellam, J. G. (1952). Studies in statistical ecology. I, Spatial pattern. *Biometrika* **39**: 346-362.

Steinitz, O., Heller, J., Tsoar, A., Rotem, D. et Kadmon, R. (2005). Predicting regional patterns of similarity in species composition for conservation planning. *Conservation Biology* **19**(6): 1978-1988.

Sterner, R. W., Ribic, C. A. et Schatz, G. E. (1986). Testing for Life Historical Changes in Spatial Patterns of Four Tropical Tree Species. *Journal of Ecology* **74**(3): 621-633.

Stoyan, D., Kendall, W. S. et Mecke, J. (1987). *Stochastic Geometry and its Applications*. John Wiley & Sons, New York. 345 p.

Stoyan, D. et Stoyan, H. (2000). Improving ratio estimators of second order point process characteristics. *Scandinavian Journal of Statistics* **27**(4): 641-656.

Sweeney, S. H. et Feser, E. J. (1998). Plant Size and Clustering of Manufacturing Activity. *Geographical Analysis* **30**(1): 45-64.

Szwagrzyk, J. (1990). Natural regeneration of forest related to the spatial structure of trees: A study of two forest communities in Western Carpathians, southern Poland. *Vegetatio* **89**: 11-22.

Szwagrzyk, J. et Czerwczak, M. (1993). Spatial patterns of trees in natural forests of East-Central Europe. *Journal of Vegetation Science* **4**(4): 469-476.

Theil, H. (1967). *Economics and Information Theory*. Rand McNally and Company, Chicago

Thioulouse, J., Chessel, D., Dolédec, S. et Olivier, J.-M. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing* **7**(1): 75-83.

Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika* **36**: 18-25.

Tomppo, E. (1986). *Models and methods for analysing spatial patterns of trees 138*. The Finnish forest research institute, Helsinki, Finland. 65 p.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* **52**(1): 479-487.

Tuomisto, H. (2010). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* **33**(1): 2-22.

Upton, G. J. G. et Fingleton, B. (1985). *Spatial Data Analysis by Example, volume 1: Point Pattern and Quantitative Data 1*. John Wiley & Sons, New York. 410 p.

Van Lieshout, M. N. M. et Baddeley, A. J. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica* **50**(3): 344-361.

Veech, J. A., Summerville, K. S., Crist, T. O. et Gering, J. C. (2002). The additive partitioning of species diversity: recent revival of an old idea. *Oikos* **99**(1): 3-9.

Wang, Z.-F., Peng, S.-L., Liu, S.-Z. et Li, Z. (2003). Spatial pattern of *Cryptocarya chinensis* life stages in lower subtropical forest, China. *Botanical Bulletin of Academia Sinica* **44**(2): 159-166.

Ward, J. S. et Ferrandino, F. J. (1999). New derivation reduces bias and increases power of Ripley's L index. *Ecological Modelling* **116**(2-3): 225-236.

Watt, A. S. (1947). Pattern and process in the plant community. *Journal of Ecology* **35**: 1-22.

Weber, A. (1909). *Uber den Standort der Industrien*. Tübingen. English translation edited in 1971, "Theory of the location of industries", Russell & Russell.

Wei, Z. et Skarpe, C. (1995). Small-scale species dynamics in semi-arid steppe vegetation in Inner Mongolia. *Journal of Vegetation Science* **6**(4): 583-592.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* **30**(3): 279-338.

Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon* **21**(2/3): 213-251.

Wiegand, T. et Moloney, K. A. (2004). Rings, circles, and null-models for point pattern analysis in ecology. *Oikos* **104**(2): 209-229.

Wiegand, T., Moloney, K. A. et Milton, S. J. (1998). Population dynamics, disturbance, and pattern evolution: Identifying the fundamental scales of organization in a model ecosystem. *The American Naturalist* **152**(3): 321-337.

Wiegand, T., Moloney, K. A., Naves, J. et Knauer, F. (1999). Finding the Missing Link between Landscape Structure and Population Dynamics: A Spatially Explicit Perspective. *The American Naturalist* **154**(6): 605-627.

Zhu, S. et Chen, X. (2007). Optimizing Urban Spatial Structure of Lanzhou Based on Geographic Concentration Method of Industries. *Chinese Journal of Population, Resources and Environment* **5**(1): 58-62.

TABLE DES FIGURES

ANNEXE 1 : PROCESSUS PONCTUELS

Les processus ponctuels fournissent le cadre mathématique nécessaire à l'étude des structures spatiales. Une définition mathématique est nécessaire avant de présenter les processus les plus courants, dont le processus de Poisson. L'approche utilisée classiquement par les non-mathématiciens est locale : les propriétés d'un processus sont définies autour de chaque point. Elle a l'avantage d'être concrète et facilement compréhensible. Son inconvénient est de laisser un certain flou sur le comportement global du processus. Une définition globale est nécessaire pour aller plus loin : c'est celle que nous présenterons en première partie de ce chapitre. La définition locale est présentée ensuite pour les lecteurs rebutés par les équations.

Ensuite, nous analyserons en détail quelques processus classiques, choisis pour leur utilité en écologie puis les méthodes nécessaires à leur simulation.

Définitions

Processus ponctuel

Nous nous placerons dans une partie du plan, typiquement une placette forestière, notée A, $A \subset \mathbb{R}^2$. Les définitions données ici et la plupart des résultats sont valables dans un espace de dimension quelconque finie mais on se limitera en pratique à un espace à deux dimensions. D'autre part, travailler sur une placette de taille infinie n'a pas d'intérêt pratique, mais rien ne s'y opposerait sur le plan mathématique, sauf dans certains cas qui seront précisés dans le texte. A sera appelée aire ou zone ou fenêtre d'étude.

Nous nous intéresserons à des ensembles dénombrables de points (on dira aussi « semis de points »). Les points seront notés en minuscules, les ensembles en majuscules : $X \subset \mathbb{R}^2$, $x \in X$. Le semis de point X sera généralement défini sur tout le plan, et son nombre de points sera infini. À l'intérieur de l'aire d'étude A, un sous ensemble de X noté X_A sera observé et cartographié : $X_A = X \cap A$.

Nous ne nous intéresserons qu'à des ensembles de points localement finis, c'est-à-dire tels que leur nombre de points dans A soit fini : $n(A) < \infty$ pour A borné. Cette restriction n'a pas de conséquences pratiques.

Il est impossible de définir directement une fonction qui attribuerait à chaque semis la probabilité de le tirer, parce que la probabilité de chaque semis est nulle. On passe donc par des ensembles de semis de points, dont la probabilité n'est pas nulle. L'ensemble des semis de points localement finis est noté N_f. N_f peut être équipé d'une σ-algèbre $\mathcal{N}_f$.

Les éléments F de $\mathcal{N}_f$ sont des ensembles de semis de points localement finis. $\mathcal{N}_f$ est une tribu (ou σ-algèbre) de N_f :

- l'ensemble vide appartient à $\mathcal{N}_f$,
- pour tout $F \in \mathcal{N}_f$, le complémentaire de F dans $\mathcal{N}_f$ appartient aussi à $\mathcal{N}_f$,
- L'union de $(F_1, F_2, F_3, \dots)$, une suite dénombrable d'éléments de $\mathcal{N}_f$appartient aussi à $\mathcal{N}_f$.

$(N_f, \mathcal{N}_f)$ est un espace probabilisable. Il reste à le doter d'une mesure P, la probabilité de tirer un ensemble de semis de points particulier F. Cet ensemble est défini par le nombre de points tirés, $N(A)$, pour tout A borné.

P est une mesure car :

- $P(\emptyset) = 0$
- P est σ-additive : $P(\bigcup_{i=1}^{n} F_i) = \sum_{i=n}^{n} P(F_i)$, où $\bigcup_{i=1}^{n} F_i$ est une union disjointe.

P est une probabilité car on fixe $P(N_f) = 1$.

Un processus ponctuel est défini comme une application d'un espace de probabilité $(\Omega, \mathcal{F}, \mathcal{P})$ vers $(N_f, \mathcal{N}_f)$ tel que $N(A)$ est une variable aléatoire pour tout A borné. La définition précise de $(\Omega, \mathcal{F}, \mathcal{P})$ importe peu puisqu'il est impossible à utiliser directement.

Les processus sont notés en lettres grecques majuscules, par exemple Ξ. Un semis de points X est une réalisation de Ξ. On note $P(X \in F)$ la probabilité que le tirage de Ξ soit un élément d'un ensemble de semis de points F particulier, par exemple défini par son nombre de points.

Enfin, les surfaces comme $\|A\|$ seront notées abusivement A pour alléger les équations.

Propriété de premier ordre

Soit S une partie de A. La propriété de premier ordre $\mu(S)$, appelée également *mesure d'intensité* du processus Ξ, est l'espérance du nombre de points dans S :

$$\boxed{\mu(S) = \mathbb{E}\big(N(S)\big)} \tag{93}$$

(93) : Mesure d'intensité d'un processus ponctuel

Dans tous les cas que nous traiterons, la mesure d'intensité pourra être écrite comme l'intégrale d'une *fonction d'intensité* λ :

$$\boxed{\mu(S) = \int_S \lambda(x)dx} \tag{94}$$

(94) : Fonction d'intensité d'un processus ponctuel

Propriété de second ordre

La *mesure du moment factoriel de second ordre* de deux parties de A, S_1 et S_2, est l'espérance du nombre de paires de points du processus Ξ se trouvant respectivement dans S_1 et S_2 :

$$\boxed{\mu_2(S_1, S_2) = \mathbb{E}\left(\sum_{x_1, x_2 \in X,\, x_1 \neq x_2} \mathbf{1}(x_1 \in S_1,\, x_2 \in S_2)\right)} \tag{95}$$

(95) : Mesure du moment factoriel de second ordre d'un processus ponctuel

De même, cette mesure pourra être écrite comme l'intégrale de λ_2, appelée *densité du produit de second ordre* :

$$\boxed{\mu_2(S_1, S_2) = \iint_{\mathbb{R}^2 \mathrm{x} \mathbb{R}^2} \mathbf{1}(x_1 \in S_1,\, x_2 \in S_2)\boldsymbol{\lambda_2}(x_1,\, x_2)dx_1 dx_2} \tag{96}$$

(96) : Densité du produit de second ordre d'un processus ponctuel

On peut démontrer (Møller et Waagepetersen, 2004) que :

$$\mathbb{E}\big(N(S_1)N(S_2)\big) = \lambda_2(S_1, S_2) + \lambda(S_1 \cap S_2) \quad \textbf{(97)}$$

Définition locale

Un processus ponctuel est l'équivalent d'une variable aléatoire dont le résultat est un ensemble de points noté X, dans un ensemble de réalisations possibles, qui sera toujours ici une surface connue et délimitée notée A.

On utilise les processus ponctuels comme outils mathématiques pour caractériser et éventuellement modéliser des événements dont on connaît la répartition spatiale, par exemple les arbres dans une forêt.

Une façon intéressante de décrire un processus ponctuel dont on ne connaît pas la loi consiste à utiliser ses propriétés de premier ordre et de second ordre.

Propriété de premier ordre

Définition

Considérons une surface A dans laquelle on observe une réalisation d'un processus ponctuel. Chaque point est noté x. On note $N(S)$ le nombre de points situés dans une surface S donnée.

La propriété de premier ordre du processus ponctuel est son intensité, notée $\lambda(x)$. Elle est définie par :

$$\boxed{\lambda(x) = \lim_{dx \to 0} \left(\frac{\mathbb{E}\big(N(dx)\big)}{dx} \right)} \quad \textbf{(98)}$$

(98) : Propriété de premier ordre d'un processus ponctuel

dx est la surface élémentaire définie autour du point x.

Si $\lambda(x)$ est constante, on parlera de processus ponctuel *homogène* et on notera l'intensité simplement λ. Un processus est *stationnaire* s'il est invariant par translation et *isotrope* s'il est invariant par rotation. Un processus homogène est donc à la fois stationnaire et isotrope.

Probabilité de trouver un point dans une surface élémentaire

On ne s'intéressera ici qu'à des processus ponctuels *ordonnés* (Diggle, 1983 page 47, Annexe C), c'est-à-dire dont la probabilité de trouver plusieurs points sur une surface élémentaire dx est d'un ordre de grandeur plus petit que dx. En d'autres termes, on pourra écrire que la probabilité de trouver plusieurs points sur dx est à peu près égale à la probabilité de n'en trouver qu'un.

Cette hypothèse n'est pas contraignante. En pratique, elle élimine les processus coalescents (dans lesquels tous les points se trouveraient superposés parce qu'ils s'attirent entre eux) ou des processus qui généreraient par exemple pour chaque point un deuxième superposé. En pratique, tous les processus ponctuels que l'on rencontrera seront ordonnés.

Cette propriété permet de lier la probabilité à l'intensité. L'existence d'un point dans la surface dx suit une loi de Bernoulli de paramètre P_{dx}, qui est à la fois sa probabilité de succès et son espérance. Cette espérance est, d'après l'équation (98), $\lambda(x)dx$.

La probabilité de trouver un point dans la surface élémentaire dx autour du point x est par conséquent :

$$\boxed{P(N(dx) = 1) = \lambda(x)dx} \quad \textbf{(99)}$$

(99) : Probabilité de trouver un point dans une surface élémentaire

Cette relation est vérifiée tant que dx est suffisamment petite pour que la probabilité d'apparition de deux points reste négligeable.

Propriété de second ordre

Définition

La propriété de second ordre d'un processus ponctuel, notée $\lambda_2(x_1, x_2)$, est définie par :

$$\boxed{\lambda_2(x_1, x_2) = \lim_{dx_1 \to 0,\ dx_2 \to 0} \left(\frac{\mathbb{E}\big(N(dx_1)N(dx_2)\big)}{dx_1 dx_2} \right)} \quad \textbf{(100)}$$

(100) : Propriété de second ordre d'un processus ponctuel

λ_2 est aussi appelée densité de paires de points (Law *et al.*, 2009).

Probabilité de trouver deux points dans deux surfaces élémentaires

La probabilité jointe de la présence d'au moins un point dans chaque surface élémentaire centrée sur x_1 et x_2 est notée $P_{dx_1 dx_2}$. Ici encore, la probabilité de trouver plus d'un point dans une surface élémentaire est négligeable. L'événement « trouver à la fois un point dans dx_1 et dans dx_2 » réalise une épreuve de Bernoulli de paramètre $P_{dx_1 dx_2}$. Selon le même raisonnement que précédemment, son espérance est $P_{dx_1 dx_2}$. Or cette espérance est connue (100), d'où :

$$P(N(dx_1)N(dx_2) = 1) = dS_1 dS_2 \lambda_2(x_1, x_2) \quad \textbf{(101)}$$

On peut rapporter dS_1 et dS_2 à la propriété de premier ordre pour obtenir :

$$P(N(dx_1)N(dx_2) = 1) = P(N(dx_1) = 1)P(N(dx_2) = 1)\frac{\lambda_2(x_1, x_2)}{\lambda(x_1)\lambda(x_2)} \quad \textbf{(102)}$$

La grandeur $\frac{\lambda_2(x_1, x_2)}{\lambda(x_1)\lambda(x_2)}$, rapport de la propriété de second ordre sur la propriété de premier ordre, est appelée fonction de distribution radiale (Diggle, 1983), ou fonction de corrélation des paires de points (Cressie, 1993). Nous suivrons (Ripley, 1977) et toute la littérature en découlant en la notant $g(x_1, x_2)$. L'usage (par exemple Ripley, 1977 ; Goreaud, 2000), a imposé g plutôt que λ_2 comme mesure de la propriété de second ordre des processus ponctuels. Nous nous y conformerons :

$$\boxed{g(x_1, x_2) = \frac{P(N(dx_1)N(dx_2) = 1)}{P(N(dx_1) = 1)P(N(dx_2) = 1)}} \quad \textbf{(103)}$$

(103) : Propriété de second ordre d'un processus ponctuel

Si le processus est isotrope, c'est-à-dire que ses propriétés sont les mêmes dans toutes les directions, et que sa propriété de second ordre est stationnaire, $g(\cdot)$ ne dépend que de la distance entre les deux points et on la notera simplement $g(r)$.

On voit immédiatement que dans le cas d'une distribution de points indépendants, la probabilité jointe est égale au produit des probabilités, et par conséquent $g(\cdot) = 1$.

Résumé et Vocabulaire

Un semis de point observable, ou distribution, est la réalisation d'un processus ponctuel.

Une distribution est dite *homogène* si son intensité est constante sur l'aire d'étude et si les relations de dépendance entre les points le sont aussi. Elle est *indépendante* si la position d'un point ne dépend pas de la position des autres. En cas de dépendance, celle-ci est toujours stationnaire dans les cas traités par la littérature.

Une distribution peut être *complètement aléatoire*, c'est-à-dire homogène (propriété de premier ordre) et indépendante (propriété de second ordre), si chaque point est distribué avec une probabilité indépendante du lieu et indépendamment des autres. En d'autres termes, il s'agit d'une distribution de Poisson homogène.

Les points peuvent s'attirer, donnant des *agrégats* (voir Figure 5, page 27). On pourra parler de *concentration spatiale* ou d'*agglomération*. L'intensité de points sera localement plus grande, sans que ce ne soit contradictoire avec l'hypothèse éventuelle d'homogénéité : une autre réalisation du même processus aurait donné des agrégats à d'autres emplacements, et l'intensité locale mesurée sur plusieurs réalisations du processus aurait été constante.

Les points peuvent se repousser, générant la *dispersion* (voir Figure 7, page 29). Les forces de dispersion créent des distributions *régulières* dans lesquelles les points ont tendance à se situer à égale distance les uns des autres.

On notera avant d'aller plus loin que les irrégularités dans une distribution (par exemple des agrégats) peuvent être dues à sa propriété de premier ordre (les agrégats sont la réalisation d'un processus ponctuel dont l'intensité est plus grande) ou de deuxième ordre (les points s'attirent). Il est impossible de trancher à partir d'un jeu de données (Feller, 1943 ; Ellison et Glaeser, 1997). Pour trancher, plusieurs réalisations du processus ponctuel sont nécessaires, mais rarement disponibles dans la réalité.

Processus utilisés

Trois types de processus seront utilisés ici. Les processus de Poisson consistent à tirer des points indépendamment les uns des autres, éventuellement avec une intensité variable. Les processus de Cox en sont une extension, dans laquelle l'intensité est non seulement variable mais aléatoire (l'intensité est un champ

aléatoire). Parmi ceux-ci, les processus de Neyman-Scott permettent de simuler l'agrégation de façon simple : une série de points de base est tirée dans un premier temps, puis les points du processus sont tirés autour des premiers, indépendamment les uns des autres. Enfin, les processus de Gibbs (ou de Markov) sont les plus complexes : les points ne sont pas tirés indépendamment mais selon des fonctions d'interaction.

Le processus de Poisson homogène

Le processus de Poisson est un processus stationnaire et isotrope, dont la réalisation donne des points à la position complètement aléatoire. Inversement, un processus ponctuel complètement aléatoire est un processus de Poisson (démonstration : Diggle, 1983, pages 51-52).

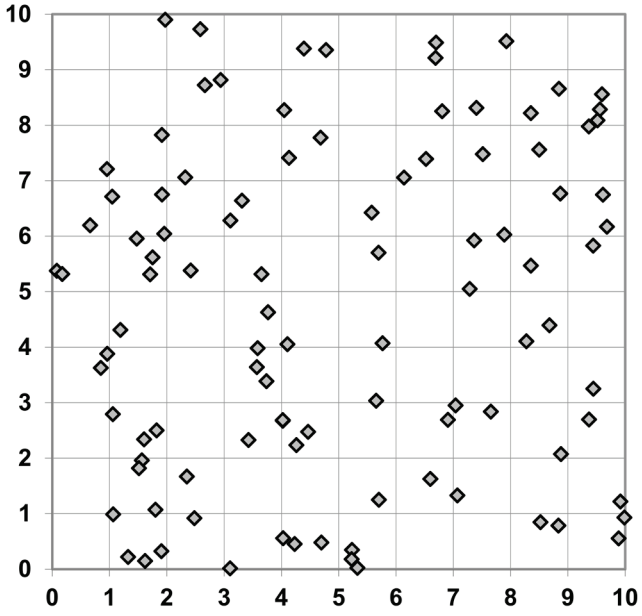


Figure 42 : Simulation d'un processus de Poisson homogène d'intensité 1

Le processus de Poisson joue un rôle central en statistiques spatiales, à la fois parce que c'est le plus simple donc celui dont les propriétés ont été le mieux étudiées, et aussi parce qu'il constitue généralement le modèle nul contre lequel des semis de points peuvent être testés. C'est un processus à accroissements indépendants, il joue le rôle des marches aléatoires pour les séries temporelles à temps discret et du mouvement Brownien pour les séries à temps continu. Ces propriétés sont utilisées pour le calcul de l'intervalle de confiance de la fonction K de Ripley (page 31).

Propriété de premier ordre

Une réalisation d'un processus de Poisson de paramètre λA sur l'aire d'étude A est un semis de points complètement aléatoire d'intensité λ. Le nombre de points suit une loi de Poisson de paramètre λA, c'est-à-dire que :

$$P(n(A) = k) = e^{-\lambda A} \frac{(\lambda A)^k}{k!} \quad \textbf{(104)}$$

Cette propriété reste valable pour une surface quelconque S choisie dans l'aire d'étude. Le nombre de points y suit une loi de Poisson :

- Son espérance est λS
- Sa variance est λS : une distribution complètement aléatoire n'est pas régulière.

On veillera à ne pas confondre l'intensité du processus ponctuel et la densité de points observée, c'est-à-dire le nombre de points divisé par la surface, qui ne correspond qu'à une réalisation du processus.

Comme on l'a vu plus haut (page 139), la probabilité de trouver un point dans une surface dx est égale à λdx.

Démonstration :

Considérons un domaine d'étude, noté A, une surface élémentaire dx et un processus de Poisson homogène d'intensité λ.

Le nombre de points localisés dans la surface dx suit une loi de Poisson de paramètre λdx. La probabilité de trouver k points dans cette surface est donc :

$$P(N(dx) = k) = e^{-\lambda dx}\frac{(\lambda dx)^k}{k!}$$

La probabilité de ne trouver aucun point est :

$$P(N(dx) = 0) = e^{-\lambda dx}$$

On choisit dx assez petit pour que λdx soit petit comparé à 1. Les premiers termes du développement limité de l'exponentielle sont :

$$\begin{aligned} P(N(dx) = 0) &= 1 - \lambda dx + \frac{(\lambda dx)^2}{2!} - \frac{(\lambda dx)^3}{3!} + \cdots \\ &\approx 1 - \lambda dx \end{aligned}$$

On retiendra l'approximation de premier ordre :

$$P(N(dx) = 0) = 1 - \lambda dx$$

La probabilité de trouver au moins un point dans dx est :

$$P(N(dx) > 0) = 1 - P(N(dx) = 0) = \lambda dx$$

Comme la probabilité de trouver plus d'un point dans dx est nulle (le processus est ordonné) :

$$\mathrm{P}(N(dx) = 1) = \lambda dx$$

Propriété de second ordre

Les points étant distribués indépendamment les uns des autres $g(\cdot) = 1$,

Remarque

Le processus de Poisson sera utilisé par la suite comme référence, en tant que générateur d'une distribution complètement aléatoire (*Complete Spatial Randomness*), à laquelle on comparera les distributions réelles. Diggle (1983 p. 51), le qualifie de pierre angulaire autour de laquelle est construite la théorie des processus ponctuels.

Le processus de Poisson hétérogène

Le processus de Poisson hétérogène est une extension du processus de Poisson homogène dans laquelle l'intensité n'est pas constante. Tout processus dont les points sont indépendants et dont le nombre de points rencontrés sur une surface dx centrée sur x a une intensité $\lambda(x)$ est un processus de Poisson hétérogène (Diggle, 1983).

La Figure 43 présente un exemple de processus de Poisson hétérogène simulé page 153.

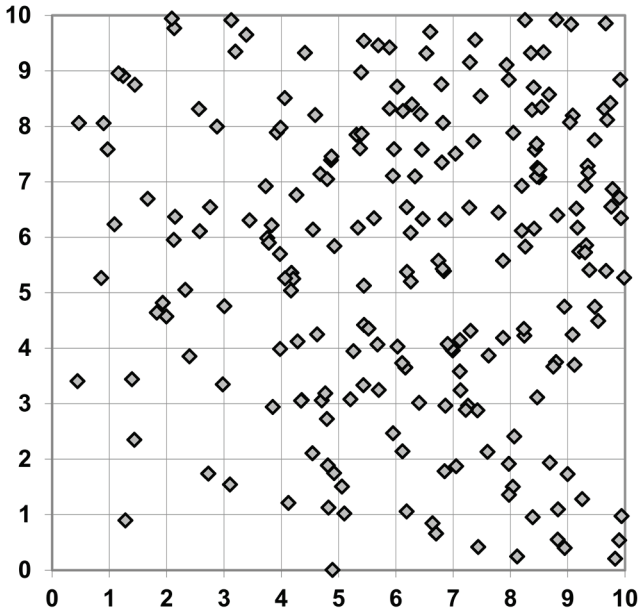


Figure 43 : Simulation d'un processus de Poisson hétérogène

La propriété de second ordre du processus, $g(\cdot)$, vaut toujours 1 puisque les points sont distribués indépendamment. Ce processus est généralement l'hypothèse nulle des modèles traitant des semis de points non homogènes.

Les processus de Cox

Les processus de Cox sont des processus de Poisson hétérogènes dont l'intensité en chaque point de l'espace est une variable aléatoire notée $Z(x)$. L'ensemble des variables $Z(x)$ pour tout $x \in \mathbb{R}^2$, $\{Z(x)\}$, est appelé un champ aléatoire. Ces processus ont été introduits par Cox (1955). On les appelle parfois *processus de Poisson doublement aléatoires*.

Le processus de Poisson agrégé ou Neyman-Scott

Les processus de Neyman-Scott (1958) sont la somme de processus indépendants définis autour de points centraux tirés préalablement. Tous ne sont pas des processus de Cox : ce n'est le cas que si les processus secondaires sont poissonniens. Nous ne nous intéresserons qu'à ceux-là.

La dénomination de cette famille de processus est assez fluctuante, et leur définition souvent donnée à partir d'un algorithme de simulation. Pour Diggle (1983), le *processus de Poisson agrégé* se déroule en trois temps:

- On génère un certain nombre de points N_1 aléatoirement, par un processus de Poisson de intensité ρ. Ces points parents ne sont pas retenus dans la distribution finale. Souvent, le nombre de points pères est fixé à ρA plutôt que tiré par une loi de Poisson.
- Autour de chaque parent, un nombre de points aléatoire est tiré selon une distribution de probabilité. Souvent (Plotkin *et al.*, 2000 par exemple), le nombre total de points fils N_2 est choisi arbitrairement et chacun est attribué aléatoirement à un point parent.
- La position des points fils autour des points pères est déterminée par une intensité de probabilité bivariée (dans le plan ; multivariée dans un espace multidimensionnel). Souvent, l'intensité ne dépend que de la distance au point, pour que le processus soit isotrope, et ne dépend que d'un seul paramètre correspondant à la taille moyenne des agrégats.

Goreaud (2000) utilise des processus de Neyman-Scott en fixant le nombre de points pères à N_1, puis un nombre constant N_2/N_1 de points fils autour de chaque point père, eux-mêmes distribués selon un processus binomial dans un cercle de rayon r. N_1, N_2/N_1et r sont les trois paramètres du processus. On obtient de cette façon N_1 agrégats de rayon r contenant chacun N_2/N_1 points (voir Figure 44). Contrairement aux apparences, il s'agit bien d'un processus homogène d'intensité N_2/A : si on répète le processus un grand nombre de fois, les agrégats seront disposés à des endroits différents et la densité locale, en

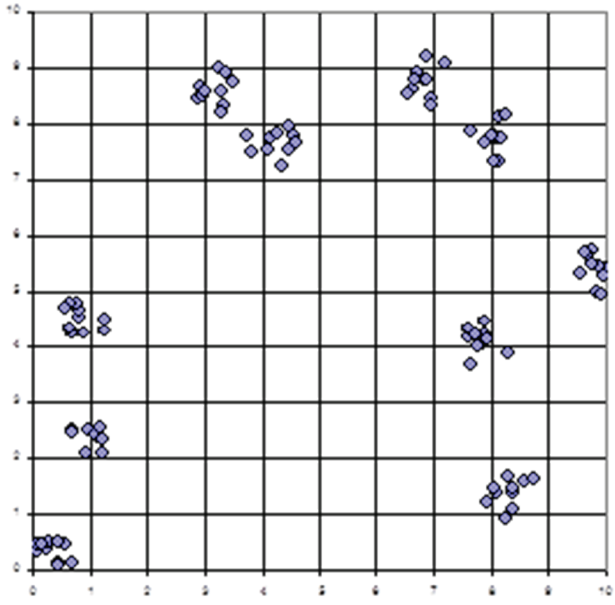


Figure 44 : Réalisation d'un processus de Poisson agrégé 10 agrégats de 10 points distribués complètement aléatoirement dans un cercle de rayon 0,5 (simulation simplifiée d'un processus de Matérn)

moyenne sur toutes les réalisations du processus, sera partout la même. Si le nombre de points (pères et fils) n'est pas fixé mais suit une loi de Poisson (respectivement d'espérance N_1et N_2/N_1), on a affaire à un processus de Matérn (1960 ; in Stoyan *et al.*, 1987).

Diggle (1983 p. 57) et Plotkin *et al.* (2000) utilisent une distribution normale des points fils autour des points pères dont la densité de probabilité est $h(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{\|x\|}{2\sigma^2}}$. La densité de probabilité de x est celle d'une loi normale à deux dimensions. La distance moyenne au centre de l'agrégat est donc $\sigma\sqrt{2}$. Contrairement au processus de Matérn, la taille des agrégats n'est pas limitée, mais on peut calculer la probabilité qu'un point se situe dans le cercle de rayon σ autour du centre de son agrégat. La probabilité de se trouver dans le cercle de dispersion d'ordre k, c'est-à-dire le cercle de rayon $k\sigma$ est $1 - e^{-\frac{k^2}{2}}$, ce qui donne les valeurs numériques suivantes : environ 39% des points se trouvent dans le cercle de rayon σ autour du centre de leur agrégat, 86% à moins de 2σ du centre et 99% à moins de 3σ.

Il s'agit du processus de Thomas (1949).

Les deux processus de Neyman-Scott que nous utiliserons par la suite sont ceux de Matérn et de Thomas.

Shot noise Cox

Les processus dits *shot noise Cox* sont définis par un champ aléatoire qui dépend d'un tirage préalable d'un processus de Poisson marqué Υ, c'est-à-dire un processus dans lequel chaque point tiré reçoit une valeur (positive) qui le décrit (par exemple, la hauteur si les points représentent des arbres). Le tirage de Υ fournit des points marqués, notés $(y, m) \in \Upsilon$.

Alors :

$$Z(x) = \sum_{(y,m)\in\Upsilon} mk(y,x) \qquad \textbf{(105)}$$

$k(y,x)$ est une fonction de noyau : sa valeur dépend de la position du point x par rapport à la référence y issue du processus père, et $k(y,\cdot)$ est une densité de probabilité sur $\mathbb{R}^2$. Généralement, $k(y,x)$ décroît avec la distance entre x et y et $Z(x)$ décrit l'influence des points de Υ (les effets de tous les points s'ajoutent mais diminuent avec la distance).

Ce modèle convient particulièrement bien pour la description d'arbres parents de position aléatoire (tirée dans Υ) disséminant leurs graines proportionnellement à leur taille m et en fonction de la distance (selon $k(y,\cdot)$).

Les processus de Gibbs

Les processus de Gibbs, également appelés de Markov, sont des processus dans lesquels les points ne sont pas indépendants, contrairement aux deux cas précédents. Ils peuvent être définis de façon globale, rigoureuse mais peu intuitive, ou de façon locale, plus simple à comprendre. Nous présentons ici la définition globale de Møller et Waagepetersen (2004 chap. 6) et celle, locale, de Tomppo (1986), utilisée par Goreaud (2000), puis une synthèse.

Définition globale

Fonction de densité d'un processus par rapport à un autre

Soit $F \in N_f$, un ensemble de semis de points localement fini, et deux processus ponctuels Ξ_1 et Ξ_2. Ξ_2 est dit *absolument continu* par rapport à Ξ_1 si et seulement si $P(X_1 \in F) = 0 \Longrightarrow P(X_2 \in F) = 0$, c'est-à-dire que tout tirage possible de Ξ_2 l'est aussi pour Ξ_1. Cette définition est équivalente (Møller et Waagepetersen, 2004, 3.2.4) à l'existence d'une fonction de densité de Ξ_2 par rapport à Ξ_1, notée f, telle que

$$P(X_2 \in F) = \mathbb{E}\big(\mathbf{1}(X_1 \in F) f(X_1)\big) \qquad \textbf{(106)}$$

Une fonction de densité permet de définir un processus ponctuel à partir d'un autre (plus simple) à la seule condition qu'il soit absolument continu par rapport à ce dernier.

Densité par rapport à un processus de Poisson

Nous nous intéresserons à des processus définis par une densité chargeant le processus de Poisson standard (c'est-à-dire d'intensité 1) sur A. Dans la majorité des cas, la fonction de densité ne sera connue que proportionnellement à une constante, appelée constante de normalisation, qui sera incalculable. Cette fonction sera notée h, $h \propto f$.

Voisinage

Une relation de voisinage est une relation entre deux points, réflexive et symétrique, notée $\sim$: $x \sim x$, si $x_1 \sim x_2$ alors $x_2 \sim x_1$. L'ensemble des voisins de x constitue son voisinage.

Le voisinage le plus couramment utilisé par la suite sera l'ensemble des points situés à une distance inférieure à une valeur R choisie.

Processus de Markov

Un processus ponctuel est dit Markovien s'il a une densité (par rapport au processus de Poisson standard) et si cette densité ne dépend que du voisinage de chaque point.

Intensité conditionnelle de Papangelou

L'intensité conditionnelle de Papangelou pour un processus ponctuel Ξ de densité f est

$$\lambda^*(X, x) = \frac{f(X \cup \{x\})}{f(X)} \quad \textbf{(107)}$$

où $X \in N_f$ et $x \in A \backslash X$, et $f(X) > 0$; $\lambda^*(X, x) = 0$ si $f(X) = 0$.

λ^* est le rapport des densité du processus quand on ajoute un point à un semis existant. λ^* présente l'avantage sur f d'éliminer la constante de normalisation, c'est pourquoi son usage sera central par la suite.

Si $\lambda^*(X_1, x) \leq \lambda^*(X_2, x)$ pour tout $X_1 \subset X_2$, le processus est est dit *attractif* , il est dit *répulsif* dans le cas contraire.

Fonctions d'interaction

Dans de nombreux cas intéressants, la fonction de densité peut être écrite comme un produit de fonctions d'interaction. On note ϕ_n les fonctions dites d'interactions entre les n-uplets de points telles que :

$$f(X) \propto \prod_{x \in X} \phi_1(x) \prod_{x_1, x_2 \in X} \phi_2(x_1, x_2) \dots \prod_{x_1, \dots, x_n \in X} \phi_n(x_1, \dots, x_n) \quad \textbf{(108)}$$

Généralement, on se limite aux interactions entre paires de points : seules ϕ_1 et ϕ_2 sont différentes de 1. La relation de proportionnalité est due à la constante de normalisation. On peut également noter les fonctions ϕ sans indice :

$$f(X) \propto \prod_{Z \subseteq X} \phi(Z) \quad \textbf{(109)}$$

Un processus est markovien si et seulement si $\phi(Z) = 1$ dès que deux points de Z ne sont pas voisins (c'est-à-dire le plus souvent : sont distants de plus de R).

Pour éliminer la constante de normalisation, on utilise l'intensité conditionnelle de Papangelou :

$$\lambda^*(X,x) = \prod_{Z \subseteq X} \phi(Z \cup \{x\}) \quad \textbf{(110)}$$

où $X \in N_f$ et $x \in A \backslash X$.

Définition locale

Un semis de points est défini par son cardinal, fixé, $n(X_A)$ et une fonction d'interaction entre les paires de points, appelée le *potentiel de paire* :

$$u(x_i, x_j) = u(\|x_i - x_j\|) = u(r) \quad \textbf{(111)}$$

Le potentiel de paire ne dépend que de la distance entre x_i et x_j. L'*énergie totale* du modèle est la somme de tous les potentiels de paires :

$$U(X) = \sum_{i=1}^{n(X_A)-1} \sum_{j=i+1}^{n(X_A)} u(\|x_i - x_j\|) \quad \textbf{(112)}$$

L'idée est qu'il est possible de trouver des configurations de points minimisant l'énergie totale, en partant d'un semis de points quelconque respectant seulement $n(X_A)$ (par exemple tiré dans un processus binomial). Chaque point peut être déplacé aléatoirement et l'énergie totale ensuite recalculée : si elle a diminué, le nouveau point est conservé, sinon on revient à l'état précédent.

Le choix de la forme du potentiel de paire définit le comportement du modèle (Figure 45) :

- Au-delà d'un seuil R, $u(r) = 0$. Seuls les points voisins interagissent.
- $u(r) = \infty$ pour $r \leq R$ interdit l'existence de paires de points à distance inférieure au seuil choisi R (processus hard-core).
- $u(r) = \beta, \beta \in \mathbb{R}_+^*$ pour $r \leq R$ crée une répulsion entre les points voisins. Au final, le semis de point est régulier.
- $u(r) = \beta, \beta \in \mathbb{R}_-^*$ pour $r \leq R$ crée une attraction entre les points voisins. Sans précaution, tous les points vont se superposer.
- La fonction $u(r)$ peut prendre diverses valeurs qui changent à plusieurs reprises, définissant plusieurs seuils R_i.

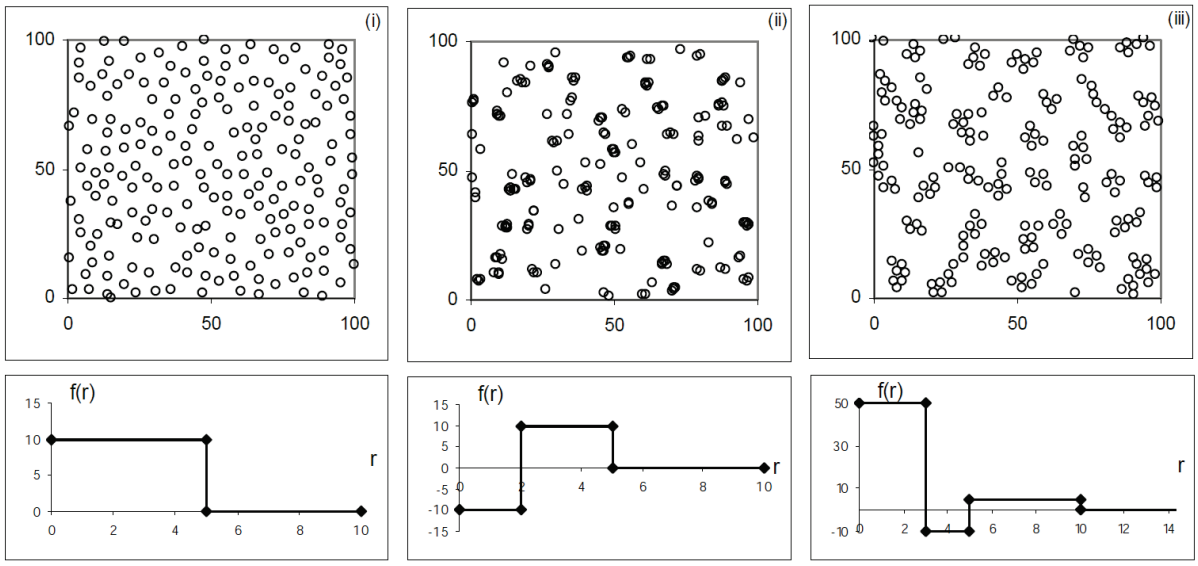


Figure 45 : Simulations de semis de points dans un processus de Gibbs.
Trois semis de points (en haut) obtenus par trois fonctions différentes $u(r)$ - notées $f(r)$ en bas : (i) un semis de points régulier, (ii) un semis attractif à courte distance, puis répulsif, (iii) un semis constitué d'agrégats régulièrement répartis. La figure est extraite de Goreaud (2000), fig. 113c, p. 42.

Illian *et al.* (2008) définissent, pour le semis de points X_A (tiré aléatoirement mais dont le nombre de points est fixé), la fonction de densité de localisation :

$$f'(X), X = \{x_1, x_2, \dots, x_{n(X_A)}\} \quad \textbf{(113)}$$

Pour un processus de Gibbs à nombre de points fixé :

$$f'(X) = e^{-\sum_{i=1}^{n(X_A)-1}\sum_{j=i+1}^{n(X_A)} u(\|x_i - x_j\|)} / K \quad \textbf{(114)}$$

K est une constante de normalisation qui assure que $f'(X)$ est bien une densité de probabilité. La constante est presque toujours impossible à calculer.

Ces fonctions ont leur origine en physique statistique. Le semis de point optimal minimise l'entropie $\int_A \dots \int_A f'(X)\ln f'(X) dx_1 \dots dx_{n(X_A)}$ pour une énergie totale donnée $\int_A \dots \int_A U(X) f'(X) dx_1 \dots dx_{n(X_A)}$.

Synthèse

La fonction d'interaction ϕ_2 et la fonction de potentiel de paire sont deux formes équivalentes :

$$\phi_2(\|x_i - x_j\|) = \mathrm{e}^{-u(\|x_i - x_j\|)} \quad \textbf{(115)}$$

Aux constantes de normalisation près, la fonction de densité par rapport au processus de Poisson et la fonction de densité de localisation sont identiques : $f \propto f'$, à la seule condition que seules les interactions d'ordre 2 interviennent.

Les deux approches sont donc à peu près équivalentes. Les principales différences sont :

- L'approche locale a été construite à partir de la physique statistique, pour un nombre de points fixés dans une zone d'étude A. Les processus sont généralement appelés *processus de Gibbs finis, à nombre de points fixe*. Son algorithme de simulation, y compris dans Illian *et al.* (2008), est problématique.
- L'approche globale a été construite à partir de la théorie des processus ponctuels, pour un nombre de points variable dans $\mathbb{R}^2$. Les processus sont plutôt appelés *processus de Markov*. Ils ne sont pas généralement, pas conditionnés au nombre de points ; en pratique l'espérance de leur nombre de points n'est pas connue, ce qui limite leur usage. Enfin, ils ont souvent des fonctions d'interaction d'ordre différent de 2 différentes de 1 : $\phi_1 = \beta$, où β est une constante, pour les processus de Strauss ci-dessous par exemple.

Le processus de Strauss

Le processus de Strauss est le plus simple des processus de Markov. Il est défini par :

$$\phi_1 = \beta$$
$$\phi_2(x_i, x_j) = \phi_2(\|x_i - x_j\|) = \phi_2(r) = \gamma^{\mathbf{1}(r \le R)} \quad \textbf{(116)}$$

β est une constante positive. La fonction d'interaction n'est définie qu'entre paires de points. Elle vaut γ, $\gamma \in [0; 1]$, si les deux points sont voisins (leur distance $r \le R$), 1 sinon (on pose $0^0 = 1$ pour le cas particulier $\gamma = 0$).

γ ne peut pas être supérieur à 1, sinon l'attractivité du processus amènerait tous les points à se superposer. Si $\gamma = 1$, le processus est un Poisson homogène. Si $0 < \gamma < 1$, le processus est répulsif et on obtient une distribution régulière de points. Si $\gamma = 0$, aucune paire de point ne peut se trouver à une distance inférieure à R. Le processus est dit *hard-core*.

La densité s'écrit :

$$f(X) \propto \beta^{n(X)} \gamma^{n(b(x,r))} \quad \textbf{(117)}$$

Au final, il n'est pas simple de prédire le nombre de points obtenu même approximativement à partir des paramètres. Le processus de Strauss conditionnel (au nombre de points $n(X) = N$) est en revanche simple à utiliser :

$$f(X) \propto \gamma^{n(b(x,r))} \quad \textbf{(118)}$$

Il s'agit d'un processus de Gibbs vu plus haut.

Le processus de Strauss multi-étages

Plusieurs seuils R_k peuvent être utilisés pour définir plusieurs niveaux d'interactions dépendant de la distance. Soit $n(R)$ le nombre de seuils, et on pose $R_0 = 0$ et $0^0 = 1$. Un processus de Strauss multi-étages est défini par :

$$\phi_2(x_i, x_j) = \prod_{k=1}^{n(R)} \gamma_k^{\mathbf{1}(R_{k-1} < r \leq R_k)} \quad \textbf{(119)}$$

À chaque intervalle de distances :

- $\gamma_k = 0$ interdit la présence de paires de points éloignés de $R_{k-1} < r \leq R_k$. Si $\gamma_1 = 0$, aucune paire de points distante de moins de R_1 n'est possible : on parle de processus hard-core.
- $0 < \gamma_k < 1$ défavorise les paires de points entre R_{k-1} et R_k.
- $\gamma_k = 1$ n'a pas d'effet.
- $\gamma_k > 1$ favorise les paires de points entre R_{k-1} et R_k.

On peut facilement rapprocher ces processus de ceux de la Figure 45, en convertissant les interactions en potentiels de paires : $u(r) = -\ln(\gamma_k)$ et en fixant le nombre de points.

Simulation

Les méthodes de simulations diffèrent entre les trois grandes familles de processus. Celle des processus de Markov est la plus complexe puisque les points ne sont pas tirés indépendamment.

Pour permettre la mise en œuvre des outils présentés, leur implémentation dans R (Ihaka et Gentleman, 1996) sera systématiquement présentée si elle existe, notamment dans le module Spatstat (Baddeley et Turner, 2005).

Simulation des processus de Poisson

Poisson homogène

La simulation la plus simple du processus consiste à :

- Tirer aléatoirement selon une loi de Poisson de paramètre λA le nombre N de points. Très souvent, cette étape est négligée et le nombre de points est fixé à λA. Le processus est alors un processus binomial (Stoyan *et al.*, 1987 p. 36), peu différent en pratique d'un véritable processus de Poisson. La différence majeure est que les nombres de points dans deux surfaces du domaine ne sont pas indépendants puisque le nombre total de points dans le domaine d'étude est fixé (Stoyan *et al.*, 1987 p. 37).
- Pour chacun de ses points, tirer chaque coordonnée selon une loi uniforme sur chacun des côtés du domaine d'étude s'il est rectangulaire. Si le domaine a une forme non rectangulaire, on tirera les points dans un domaine rectangulaire en éliminant les points hors du domaine, jusqu'à obtenir le nombre de points N dans le domaine.

Poisson hétérogène

La méthode de simulation la plus simple consiste à simuler un processus de Poisson homogène, dit *dominant*, de densité λ_{max} supérieure ou égale à l'intensité maximale du processus hétérogène, puis à supprimer chaque point x_i avec une probabilité $\frac{\lambda_{x_i}}{\lambda_{max}}$. En pratique, on tire pour chaque point une valeur u_i dans la loi uniforme sur $[0;1]$ et on élimine le point x_i si $u_i \geq \frac{\lambda_{x_i}}{\lambda_{max}}$. L'inconvénient de cet algorithme est d'être lent si le processus est très hétérogène : si λ_{max} est grand, on tire beaucoup de points dont une grande partie sera supprimée ensuite.

Simulation dans R

La fonction `rpoispp(lambda, lmax, win = owin(c(0,1),c(0,1)), ...)` du module Spatstat simule des processus de Poisson, homogènes ou non, dans des fenêtres de forme quelconque. Paramètres :

- `lambda` : densité du processus,
- `lmax` : densité du processus dominant, facultative (estimée si elle n'est pas saisie),

- `win` : fenêtre de simulation, par défaut le domaine carré de côté 1,
- … : paramètres supplémentaires, passés à la fonction *lambda*.

Le code ci-dessous simule un processus de densité $\lambda(x,y) = \frac{0{,}3xy}{ln(x+1)ln(y+1)}$ dans une fenêtre carrée de 10x10 :

```
>#  Définition de la fonction de densité lambda
>lamda <- function(x,y) {.3 * x * y / log(x+1) / log(y+1)}
>#Création d'un semis de point de densité lambda, dans une fenêtre carrée de 10x10
>pp <- rpoispp(lamda, win=owin(c(0,10), c(0,10)))
>#  Affichage
>plot(pp)
```

Le résultat de la fonction *plot* se trouve Figure 43, page 144. La simulation des processus hétérogènes est faite par éclaircie d'un processus homogène dominant.

Simulation des processus de Cox

Simulation approchée

La définition du processus donnée par Diggle (1983) est l'algorithme de simulation. Les seules précisions à apporter concernent des détails techniques :

- Pour le processus de Matérn, la position des points fils est obtenue en simulant leurs coordonnées par deux lois uniformes et en éliminant les points placés hors du cercle de rayon r. Les points éliminés sont tirés à nouveau.
- Pour le processus de Neyman-Scott gaussien, les coordonnées des points fils sont simulées par deux lois normales $\mathcal{N}(0,\sigma^2)$ puis ajoutées à celles du point père. Le domaine d'étude est traité comme un tore pour régler les problèmes d'effets de bord. La simulation de la loi normale peut être réalisée par la méthode de Box-Muller (Box et Muller, 1958) si le générateur de nombres aléatoires disponible ne fonctionne que dans la loi uniforme : si x et y suivent des lois uniformes sur $]0;1]$, alors z_1 et z_2 définies par les équations (120) suivent une loi normale centrée réduite :

$$z_1 = \sqrt{-2ln(x)}cos(2\pi y)$$
$$z_2 = \sqrt{-2ln(x)}sin(2\pi y) \quad \textbf{(120)}$$

Les variables aléatoires tirées sur une loi uniforme sont fournies par les générateurs de nombres pseudo-aléatoires et le calcul de $z_1$ou z_2 est très rapide, ce qui rend cette méthode plus performante que son alternative : le théorème central limite prévoit que la somme d'un grand nombre n (classiquement plus de 30) de

tirages d'une loi uniforme a une distribution normale de même espérance et de variance n fois la variance de la loi uniforme.

Si le domaine n'est pas traité comme un tore, la simulation est biaisée : certains centres situés hors du domaine ont des points fils situés à l'intérieur. La simulation les élimine. Une solution consiste à tirer les centres dans une zone plus grande, suffisamment loin des limites pour que la probabilité qu'un point fils d'un centre non tiré soit dans le domaine soit très faible. Pour le processus de Matérn, il suffit d'ajouter une zone tampon de largeur égale au rayon des cercles. Pour le processus de Thomas, une zone tampon de largeur égale à trois écart-types de la loi normale permet de limiter le risque à 1% des points fils des centres non simulés.

Simulation dans R

La fonction `rMatClust(kappa, r, mu, win = owin(c(0,1),c(0,1)))` du module Spatstat simule des processus de Matérn. Paramètres :

- `kappa` : densité du processus de Poisson pour les centres,
- `r` : rayon des cercles
- `mu` : nombre moyen de points par cercle
- `win` : fenêtre de simulation, par défaut le domaine carré de côté 1.

```
>#  Simulation d'un processus de
Matérn
>pp <- rMatClust(0.1, 0.5, 10, win
= owin(c(0,10),c(0,10)))
>plot(pp)
```

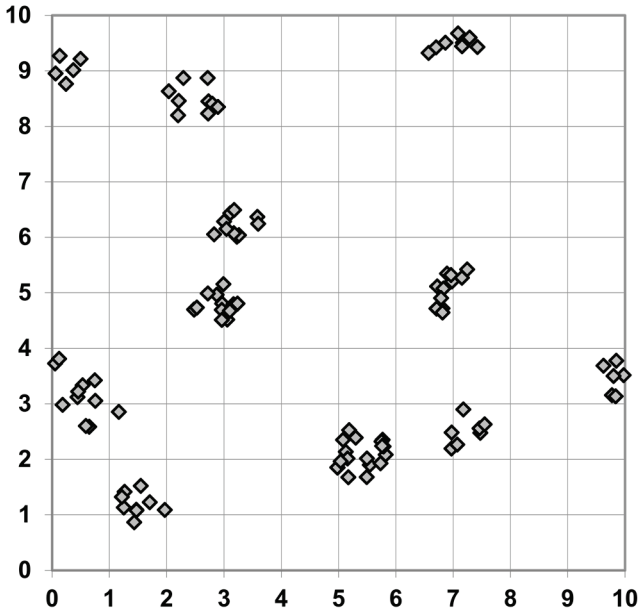


Figure 46 : Simulation d'un processus de Matérn

Le résultat de la simulation est représenté par la Figure 46. Les centres sont tirés dans une fenêtre rectangulaire contenant le domaine et une zone tampon de la taille du rayon des cercles : la simulation n'est donc pas biaisée. Le nombre de points fils autour de chaque centre est tiré dans une loi de Poisson d'espérance *mu*.

La fonction `rThomas(kappa, sigma, mu, win = owin(c(0,1),c(0,1)))` simule un processus de Thomas. Paramètres :

- `kappa` : densité du processus de Poisson pour les centres,
- `sigma` : écart-type de la loi de densité des points fils,
- `mu` : nombre moyen de points par cercle
- `win` : fenêtre de simulation, par défaut le domaine carré de côté 1.

```
>#  Simulation d'un processus de
Thomas
>pp <- rThomas (0.1, 0.5, 10, win =
owin(c(0,10),c(0,10)))
>plot(pp)
```

La Figure 47 montre le résultat de la simulation d'un processus de Thomas avec les mêmes paramètres que le processus de Matérn de la Figure 46. La zone tampon utilisée est de 4 écart-types.

Le module Spatstat fournit une fonction `rNeymanScott` qui permet de générer n'importe quel processus de Neyman-Scott en lui donnant comme paramètre la fonction de génération des points fils.

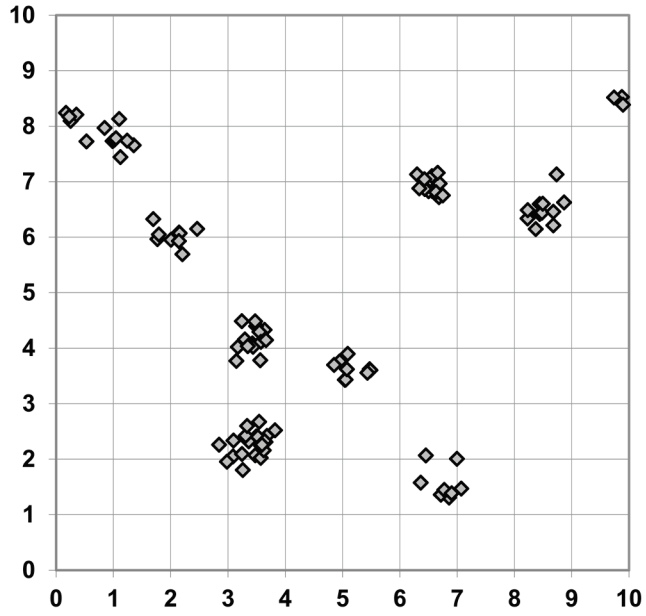


Figure 47 : Simulation d'un processus de Thomas

Simulation des processus de Gibbs

Méthode de Metropolis Hastings

L'algorithme de Metropolis-Hastings génère une chaîne de Markov dont les états successifs sont des semis de points et la limite le processus ponctuel à simuler. Il appartient à la famille des algorithmes MCMC (*Monte Carlo Markov Chains*) qui consistent à générer des suites Y_n d'approximations d'une distribution cible. Le caractère markovien de la suite vient du fait que chaque état Y_n ne dépend que du précédent. A condition que la suite soit *réversible* et *irréductible*, elle tend vers la distribution cible. La réversibilité signifie que relativement à la distribution cible la probabilité que la chaine aille d'un état A vers un état B est égale à la probabilité que la chaine aille de B vers A. L'irréductibilité signifie que tous les états possibles de Y_n peuvent être atteints, quel que soit le point de départ Y_0.

Le passage d'un état de la suite au suivant est réalisé selon l'algorithme suivant :

- Partant de l'état Y_n, un état candidat $\tilde{Y}$ est généré *selon un noyau de proposition* $q(Y_n, \tilde{Y})$ dont la densité de probabilité est connue.
- Le *ratio de Hastings* est le rapport :

$$\rho = \frac{p(\tilde{Y})q(\tilde{Y}, Y_n)}{p(Y_n)q(Y_n, \tilde{Y})} \quad \textbf{(121)}$$

- Pour assurer la réversibilité de la chaîne, la probabilité de passer de Y_n à $\tilde{Y}$ est $\min(1, \rho)$: le candidat $\tilde{Y}$ est retenu en tant que Y_{n+1} avec une probabilité d'autant plus grande que le ratio de Hastings est grand. S'il n'est pas retenu, $Y_{n+1} = Y_n$.

Les probabilités des différents semis de points ne sont pas connues mais leur rapport l'est dans certains cas :

- Si le passage de Y_n à $\tilde{Y}$ consiste à ajouter un point x au semis X, le rapport de probabilités est l'intensité conditionnelle de Papangelou, équation (107) : $\lambda^*(X, x)$. Le rapport est inverse si on supprime un point.
- Si le passage consiste à déplacer un point le rapport des fonctions d'interaction, équation (108), du nouveau et de l'ancien point sont simples à calculer : les fonctions d'interaction n'interviennent qu'entre les points et leurs voisins.

Algorithme pour un processus de Markov conditionné au nombre de points

On tire Y_0 l'état initial dans un processus binomial contenant le bon nombre de points $n(Y_0)$. Ensuite, partant de Y_n, on choisit un des points au hasard et on le remplace par un nouveau point tiré uniformément dans A.

On calcule $p(\tilde{Y})/p(Y_n)$ en prenant simplement en compte les interactions de l'ancien point et du nouveau point avec leurs voisins. Le noyau de proposition est symétrique ($q(\tilde{Y}, Y_n) = q(Y_n, \tilde{Y})$) : sa valeur est le produit de la probabilité de choisir un point particulier ($1/n(Y_n)$) pour le supprimer et de la densité de probabilité de le recréer à son nouvel emplacement ($1/A$). Le ratio de Hastings est donc simplement égal au rapport des probabilités, ce qui signifie qu'une configuration plus probable sera systématiquement retenue, comme dans l'algorithme de Tomppo, mais la réversibilité est assurée par la possibilité de retenir avec une probabilité égale à ρ une configuration moins probable (qui augmente l'énergie totale du modèle).

Algorithme de simulation

L'algorithme de naissance et mort de points est présenté ici. Il peut être combiné avec l'algorithme de déplacement (Møller et Waagepetersen, 2004, algorithme 7.5, page 115). Dans R, un tirage au sort entre déplacement et naissance-mort est fait avant un tirage éventuel entre naissance et mort si le déplacement n'a pas été retenu.

Y_0 l'état initial peut être quelconque, y compris égal à $\emptyset$.

On choisit une probabilité q arbitraire de réaliser une mort de point, $1-q$ étant la probabilité de naissance (par défaut dans R, $q = 0{,}5$).

Partant de Y_n, on tire u_{mort} dans une variable aléatoire uniforme sur $[0;1]$:

- Si $u_{mort} \leq q$, on tente de faire disparaître un point x :
  - Il est tiré uniformément dans le semis avec la probabilité $q_{mort}(Y_n, x) = 1/n(Y_n)$. Le candidat $\tilde{Y}$ est $Y_n \backslash \{x\}$.
  - Pour la réversibilité de la chaîne, il faut calculer la densité de probabilité de naissance de x dans le semis de points ne le contenant plus : $q_{naissance}(\tilde{Y}, x) = 1/A$ puisque le processus est homogène.
  - On peut donc écrire le ratio de Hastings :

$$\rho_{mort} = \frac{(1-p) q_{naissance}(Y_n \backslash \{x\}, x)}{\lambda^*(Y_n \backslash \{x\}, x) p \; q_{mort}(Y_n, x)} \quad \textbf{(122)}$$

  - on tire $u_{confirmation}$ dans une variable aléatoire uniforme sur $[0;1]$.
  - Si $u_{confirmation} < \rho_{mort}$, le point x est effectivement supprimé et $Y_{n+1} = \tilde{Y}$, sinon $Y_{n+1} = Y_n$.
- Si $u_{mort} > q$, on tente de faire apparaître un point x :
  - Il est tiré uniformément dans A avec la densité de probabilité $q_{naissance}(Y_n, x) = 1/A$. Le candidat $\tilde{Y}$ est $Y_n \cup \{x\}$.
  - Pour la réversibilité de la chaîne, il faut calculer la densité de probabilité de mort de x dans $\tilde{Y}$: $q_{mort}(\tilde{Y}, x) = 1/[n(Y_n) + 1]$.
  - On peut donc écrire le ratio de Hastings :

$$\rho_{naissance} = \frac{\lambda^*(Y_n, x) p \; q_{mort}(Y_n \cup \{x\}, x)}{(1-p) \; q_{naissance}(Y_n, x)} \quad \textbf{(123)}$$

  - on tire $u_{confirmation}$ dans une variable aléatoire uniforme sur $[0;1]$.
  - Si $u_{confirmation} < \rho_{naissance}$, le point x est effectivement créé, sinon $Y_{n+1} = Y_n$.

Pour améliorer l'efficacité de l'algorithme, la probabilité p peut varier selon l'état de la chaîne de Markov, par exemple en augmentant la probabilité de naissance quand le nombre de points diminue. En pratique, l'algorithme décrit ici est suffisant pour les cas courants. La question de la convergence n'a pas de réponse claire : R propose 500 000 itérations par défaut. En réalité, le temps d'obtention du résultat est aléatoire et non borné.

Simulation dans R

La fonction `rmh(model, start, control, verbose=TRUE, ...)` du module Spatstat simule des processus de Gibbs par déplacements, naissances et morts de points. Paramètres :

- `model` : modèle décrivant le processus à simuler,
- `start` : paramètres de départ, notamment le nombre de points,
- `control`: paramètres de contrôle du fonctionnement de l'algorithme,
- `verbose` : affichage ou non de la progression,
- … : paramètres supplémentaires passés au modèle.

La description du modèle est faite dans une liste, dont les éléments sont :

- `cif` : chaîne de caractère contenant le nom du modèle, par exemple `"Strauss"` ou `"Straush"` pour un processus de Strauss normal ou hardcore. La liste des modèles utilisables est dans l'aide de R et évolue à mesure des nouvelles implémentations.
- `par` : vecteur contenant les paramètres du modèle.

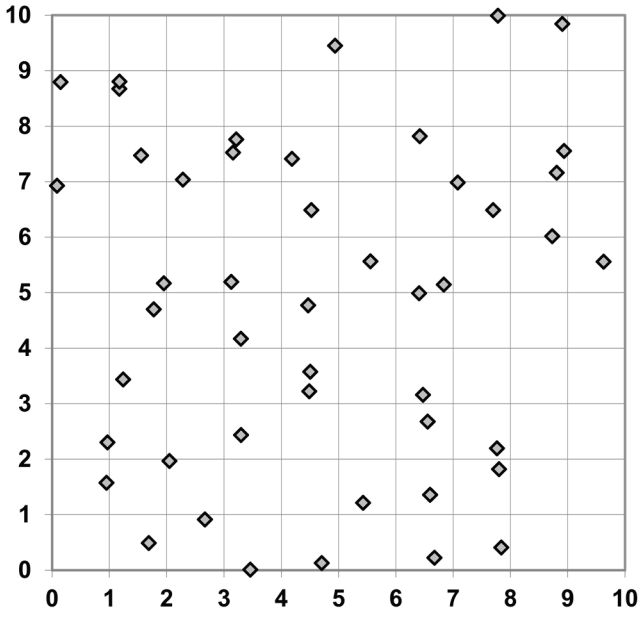


Figure 48 : Simulation d'un processus de Strauss

Le contrôle du fonctionnement de l'algorithme est décrit dans une liste dont les éléments utiles sont :

- `p` : la probabilité que l'algorithme propose un déplacement plutôt qu'une naissance ou mort. Fixer `p=1` et partir d'un semis de points tiré dans un processus binomial permet de simuler un processus conditionné au nombre de points,
- `q` : la probabilité que l'algorithme propose une mort de point plutôt qu'une naissance,
- `nrep` : le nombre d'itérations

```
># Description du modèle
>modStrauss <- list(cif="strauss",par=list(beta=1, gamma=0.5, r=1),
w=c(0,10,0,10));
># Simulation, 500 000 itérations
>pppStrauss <- rmh(model=modStrauss, start=list(n.start=80),
control=list(nrep=5e5));
>plot(pppStrauss)
```

Le résultat de la simulation est montré Figure 48.

Simulation parfaite

La simulation par MCMC n'apporte pas de garantie de convergence du résultat. En pratique, ce problème est peu gênant parce qu'il est possible de réaliser un très grand nombre d'itération en très peu de temps de calcul. Des algorithmes de simulation parfaite ont été développés récemment (Møller et Waagepetersen, 2004). Leur résultat est à coup sûr une réalisation du processus attendu.

La fonction `rStrauss(beta, gamma = 1, R = 0, W = owin())` simule un processus de Strauss par simulation exacte (Figure 49). Paramètres :

- `beta` : paramètre d'intensité,
- `gamma` : paramètre d'interaction,
- `R` : distance d'interaction,
- `W` : fenêtre de simulation.

```
># Simulation parfaite d'un
processus de Strauss
>pp <- rStrauss (1, 0.5, 1, W =
owin(c(0,10),c(0,10)))
>plot(pp)
```

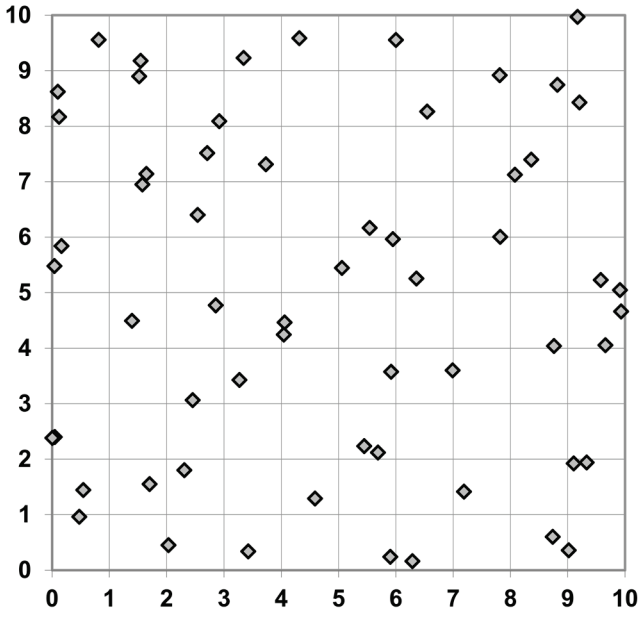


Figure 49 : Simulation parfaite d'un processus de Strauss

La simulation exacte est encore

« expérimentale » d'après la documentation de Spatstat. Le seul processus implémenté est celui de Strauss, la fenêtre est obligatoirement rectangulaire… Les développements mathématiques sont récents et les algorithmes compliqués. Actuellement, il est plus efficace de multiplier les itérations d'une simulation MCMC plutôt que de recourir à la simulation parfaite.

ANNEXE 2 : MÉTHODES ALTERNATIVES EN STATISTIQUES SPATIALES CONTINUES

Les méthodes de caractérisation de la structure spatiale d'un semis de points autres que celles traitées au premier chapitre sont passées en revue ici. Les premières sont les variantes de la fonction K de Ripley, développées pour certains objectifs particuliers. Les autres sont l'ensemble des méthodes dites « du plus proche voisin » et quelques outils plus confidentiels. Enfin, les fonctions inter-types à marques continues, inspirées de K également, sont mentionnées pour être complet.

Les variantes de K

Les variantes destinées à améliorer certaines propriétés de la fonction K sont nombreuses. Les principales sont présentées ici. La fonction L_q de Haase (2001) permet de détecter l'anisotropie. La fonction L_i de Getis et Franklin (1987) fournit des valeurs individuelles de L et permet la cartographie de l'agrégation. La fonction D standardisée est une variante de D dont les valeurs sont comparables d'un jeu de points à l'autre. Enfin, la méthode des grilles de Fehmi et Bartolome (2001) permet de travailler sur des points dont la position exacte n'est pas connue, mais seulement l'appartenance à une case d'un quadrillage.

La fonction L_q de Haase (2001)

Dans le but de distinguer les facteurs biotiques et du milieu dans la responsabilité des structures spatiales, Haase (2001) propose une variante de la fonction de Ripley calculée pour chaque direction (en pratique, les quatre secteurs NE, NW, SE et SW). Le principe est simple et consiste seulement à ne comptabiliser les voisins que dans le secteur concerné. La fonction K obtenue de cette façon est sous-estimée d'un facteur 4, la correction est apportée dans le calcul de L :

$$L_q = \sqrt{\frac{4K_q(r)}{\pi}} - r \qquad \textbf{(124)}$$

La correction des effets de bord est plus problématique : comme l'auteur utilise la correction de Ripley, qui prend en compte les points de la couronne ajoutée à chaque incrémentation de r, les points de tous les secteurs entrent dans le calcul. L'attribution du nombre de voisins résultant de la correction est réalisée proportionnellement à la contribution de chaque secteur. La correction de Besag aurait permis de ne comptabiliser que les voisins du secteur concerné. À cette petite réserve près, l'utilisation de la fonction L_q constitue une avancée puisqu'elle permet de détecter l'anisotropie.

La fonction L_i de Getis et Franklin (1987)

Les auteurs utilisent une variante individuelle de la fonction L. Alors que la fonction K de Ripley est estimée par le nombre moyen de voisins autour de tous les points, Getis et Franklin (1987) s'intéressent au nombre de voisins autour de chaque point et définissent pour chaque point x_i :

$$K_i(r) = \frac{A}{n(A)-1}\frac{2\pi r}{L_{ir}}\sum_{j=1,i\neq j}^{n(A)} \mathbf{1}\left(\|x_i - x_j\| \leq r\right) \qquad (125)$$

$$L_i(r) = \sqrt{\frac{K_i(r)}{\pi}}$$

(125) : La fonction L_i de Getis et Franklin

Le facteur $2\pi r/L_{ir}$ correspond à la correction des effets de bord, par la méthode de Ripley (*cf.* page 23). On notera au passage la correction du biais classique de K ($n(A)-1$ et non $n(A)$ au dénominateur, *cf.* page 49).

Cette approche ne remplace pas celle de Ripley : s'agissant de valeurs individuelles, les courbes de L_i peuvent être comparées à une hypothèse nulle locale. Les agrégats observés peuvent être dus à l'existence d'une structure spatiale significative ou être de simples fluctuations stochastiques d'un processus de Poisson homogène. Pour trancher, on peut générer un semis complètement aléatoire, y choisir un point x_i au hasard et calculer $L_i(r)$. On peut ensuite répéter l'opération un grand nombre de fois, éliminer les valeurs extrêmes selon la méthode de Monte-Carlo et on obtiendra un intervalle de confiance. Si la courbe $L_i(r)$ du point observé sort de l'intervalle de confiance, on pourra conclure que le point ne fait probablement pas partie d'une réalisation d'un processus de Poisson.

On peut remarquer que la valeur de $L_i(r)$ est très similaire à une densité locale calculée sans pondération sur la surface du cercle de rayon r centré sur le point

x_i. Des méthodes de calcul de densité plus élaborées peuvent être préférables (Diggle, 1985).

La fonction *D* standardisée

Cette évolution de la fonction D de Diggle et Chetwynd (1991) a été introduite par Feser et Sweeney (2000) pour permettre une meilleure interprétation des résultats.

La fonction $D(r)$ est la différence entre les valeurs de la fonction K de la population des cas et de celle des contrôles (page 53). Sa valeur peut donc être interprétée comme la différence de surface de deux cercles : celui qui contiendrait le nombre de cas observés et celui qui contiendrait le nombre de contrôles observé, dans le cas d'une distribution complètement aléatoire des points. Cette différence de surfaces est assez peu parlante et présente un inconvénient de taille : elle augmente avec la valeur de r. Les maxima de D ne correspondent donc pas à un maximum d'agrégation, quelle qu'en soit la définition retenue.

La normalisation de la fonction D consiste à diviser sa valeur par l'écart-type de la valeur de l'hypothèse nulle. Diggle et Chetwynd (1991) ont montré que sous l'hypothèse nulle d'étiquetage aléatoire, l'espérance de $D(r)$ est 0 et la valeur supérieure de l'intervalle de confiance à la distance r vaut approximativement 1,96 fois $\sigma(r)$, l'écart-type de $D(r)$. L'estimation de $\sigma(r)$ étant assez complexe (p. 1157), il est plus simple d'évaluer l'intervalle de confiance par la méthode de Monte-Carlo et d'en déduire l'écart-type. Cet écart-type augmente avec r. Diviser la valeur de $D(r)$ par $\sigma(r)$ est donc une forme de normalisation, justifiée par les auteurs par son analogie avec un score z, habituel en statistiques. Le score z est le rapport entre une valeur observée et l'écart-type de la distribution d'espérance nulle dont on cherche à déterminer si elle en est une réalisation plausible. Dans le cas d'une distribution normale, on compare classiquement le score z à la valeur de la fonction de Student (1,96 pour un nombre d'observation assez grand). Un grand score z signifie une faible probabilité que la valeur observée soit une réalisation de l'hypothèse nulle.

Les auteurs définissent ainsi implicitement l'agrégation maximale comme la plus grande probabilité de rejet de l'hypothèse d'étiquetage aléatoire. Cette définition permet de comparer les valeurs de la fonction pour différentes valeurs distances, mais reste assez éloignée de la notion intuitive d'agrégation. L'intervalle de confiance de l'hypothèse nulle de la fonction D standardisée est par construction $[-1{,}96; 1{,}96]$ quel que soit la distance.

Cette évolution de la fonction D ne règle pas le problème fondamental non résolu par la fonction originale : la différence de deux fonctions K ne permet que de

comparer la structure propre de chacune des deux populations mais ignore totalement leurs localisations relatives.

La méthode des grilles de Fehmi et Bartolome (2001)

Fehmi et Bartolome (2001) remarquent à juste titre que l'acquisition des données géographiques est coûteuse, particulièrement dans leur domaine d'étude (les plantes herbacées dans une prairie), et parfois imprécise : les plantes sont parfois difficiles à individualiser, les tiges multiples peuvent être attribuées à un ou plusieurs individus.

Ils proposent une simplification qui consiste à placer sur le domaine d'étude une grille de maille convenable (des carrés de 5 cm de côté pour les herbacées étudiées sur une placette de 50 cm x 50 cm) et simplement noter dans ces carrés la présence ou l'absence d'une plante, une tige multiple ou de diamètre important pouvant donc être présente sur plusieurs carrés. Les carrés contenant au moins une plante se voient affecter un point (et un seul) en leur centre, et la fonction de Ripley est appliquée à ce semis de points. L'hypothèse nulle est la répartition aléatoire du nombre de points observés sur les centres de tous les carrés.

Les auteurs concluent d'après leurs dix placettes que la structure spatiale ainsi détectée est très similaire à celle mise en évidence par la fonction K authentique, pour un effort de collecte de données très inférieur. Une différence notable est la détection systématique d'une concentration à une distance correspondant à la taille des plantes si elle est supérieure à celle des carrés (cette concentration apparente est due à la représentation d'une plante unique par plusieurs points contigus).

Les données des auteurs ne contiennent pas de semis de points très hétérogènes pour lesquels de nombreux points seraient remplacés par un seul dans quelques carrés, cas où la perte d'information serait beaucoup plus sensible, amenant à une sous-estimation importante du niveau de l'agrégation.

Autres Méthodes

Ces autres méthodes ne sont pas fondées sur la fonction K de Ripley. La plus ancienne, présentée pour son intérêt historique, est celle de Clark et Evans (1954). Les fonctions F et G de Diggle et J de Van Lieshout et Baddeley prennent en compte seulement le plus proche voisin.

Le test T_k de Cuzick et Edwards est un test non paramétrique utilisant l'ensemble des points.

La fonction de Podani et Czárán (1997) est la seule à traiter les interactions entre plus de deux types de points. Une extension de K aux interactions entre triplets de points a été développée par Schladitz (2000) mais n'est pas présentée ici.

Enfin, la méthode de reconnaissance des agrégats de Coomes *et al.* (1999) permet de détecter des agrégats par une suite de tests.

La méthode du plus proche voisin

Clark et Evans (1954) ont introduit la première méthode fondée sur la distance au plus proche voisin. Dans le cadre de l'hypothèse nulle d'une distribution de Poisson, la distance moyenne entre chaque point et son plus proche voisin, notée $\bar{r}_e$ est connue. En présence d'agrégation, la distance moyenne observée, notée $\bar{r}_a$, est inférieure ; elle est supérieure si la structure est régulière.

Formellement, on considère un semis de $n(A)$ points, on note r_i la distance du point x_i à son plus proche voisin et λ l'intensité :

$$R = \frac{\bar{r}_a}{\bar{r}_e} = \frac{\sum_{i=1}^{n(A)} r_i}{n(A)} \Big/ \frac{1}{2\sqrt{\lambda}} \quad \textbf{(126)}$$

L'intensité est estimée couramment par $n(A)/A$. Les auteurs notent qu'un estimateur non biaisé est plutôt $(n(A)-1)/A$, puisque le nombre de voisins est égal au nombre de points à l'exception du point de référence.

La valeur de référence R est 1, correspondant à une distribution complètement aléatoire. Les valeurs extrêmes sont 0 si tous les points sont superposés (agrégation extrême) et environ 2,15 pour une distribution hexagonale régulière, la plus dispersée possible.

Pour déterminer la significativité de la valeur de R, Clark et Evans proposent la statistique C :

$$C = \frac{\bar{r}_a - \bar{r}_e}{\sigma_{\bar{r}_e}} \quad \textbf{(127)}$$

où $\sigma_{\bar{r}_e}$ est l'écart-type de $\bar{r}_e$ égal à $\sqrt{\frac{(4-\pi)}{4\pi\lambda n(A)}} \approx \frac{0{,}26136}{\sqrt{\lambda n(A)}}$. C suit une loi normale centrée réduite, et sa valeur peut être comparée à la valeur critique de la loi de Student, au seuil considéré (1,96 au seuil de 5% pour $\hat{n}(A)$ assez grand).

La valeur de R comme celle de C ont été abondamment contestées dans la littérature. La valeur de $\bar{r}_e$ est calculée pour un domaine illimité. Clark et Evans prévoient une méthode de correction des effets de bord qui consiste à définir une zone tampon suffisante pour que le plus proche voisin de tous les points de la zone d'étude soit clairement localisé. Upton et Fingleton (1985) indiquent que cette précaution a été fréquemment négligée dans les applications. D'autre part, la normalité de C a été remise en cause, comme la formule de la variance, notamment par Diggle (1976). Diverses alternatives ont été proposées, par exemple dans Upton et Fingleton (1985), p. 74. Enfin, Porter (1960), cité par Haggett *et al.* (1977), a donné des contre-exemples correspondant à des distributions particulières, dans lesquels les résultats donnés par l'indice de Clark et Evans sont contradictoires à la construction du jeu de données : la valeur de R n'est que le reflet d'une distance moyenne et ne prend pas en compte sa dispersion.

Malgré toutes ces réserves, l'indice R a été le premier à prendre en compte explicitement les distances entre points, définir clairement la distribution poissonnienne comme référence, et donner un niveau de significativité aux résultats. Cette première approche ne prend en compte que le plus proche voisin et la seule distance moyenne. L'évolution suivante a été accomplie par les fonctions F et G de Diggle, qui prennent en compte la fonction de distribution de la distance au plus proche voisin plutôt qu'une seule valeur moyenne, avant la fonction K de Ripley qui prend en compte toutes les distances et tous les voisins.

Les fonctions F et G de Diggle

Les fonctions du plus proche voisin F et G ont été introduites par Diggle (1979) et détaillées dans Diggle (1983).

Considérons un semis de points de densité moyenne λ dans un domaine d'étude. On se place au choix en chaque point d'une maille prédéfinie placée sur le domaine d'étude (fonction F) ou sur chaque point du semis (fonction G). On construit les courbes $F(r)$ et $G(r)$ de la fraction des points dont le plus proche voisin est à une distance inférieure ou égale à r. La définition de G est assez intuitive : en présence d'agrégats, on s'attend à voir les courbes augmenter rapidement ; au contraire, si les points se repoussent, peu de points auront leur premier voisin à petite distance. La fonction G décrit le voisinage des points. La fonction F quant à elle mesure les espaces vides dans le jeu de données : $1 - F(r)$ est la proportion de points qui n'ont aucun voisin à la distance r. Si les points sont agrégés, ils laisseront des vides entre eux et la courbe F présentera des valeurs faibles.

Le choix de la maille pour le calcul de la fonction F est discuté par Ripley (1981) et Diggle (1983). Upton et Fingleton (1985) conseillent de choisir une grille régulière du même nombre de points que la distribution étudiée, limité à 400, alors

que Diggle conseille de se limiter à la racine du nombre de points. Il est clair que la précision de l'estimation augmente avec le nombre de points.

Les courbes de référence sont définies pour l'hypothèse nulle d'une distribution complètement aléatoire (processus de Poisson), et leurs équations sont :

$$F_0(x) = 1 - e^{-\pi\lambda x^2}$$
$$G_0(w) = 1 - e^{-\pi\lambda w^2} \quad \textbf{(128)}$$

Les courbes réelles pour le semis de points sont estimées et comparées aux courbes théoriques. Le test porte sur le plus grand écart entre la valeur observée et la valeur théorique :

$$dx = \sup\left|\hat{F}(x) - F_0(x)\right|$$
$$dw = \sup\left|\hat{G}(w) - G_0(w)\right| \quad \textbf{(129)}$$

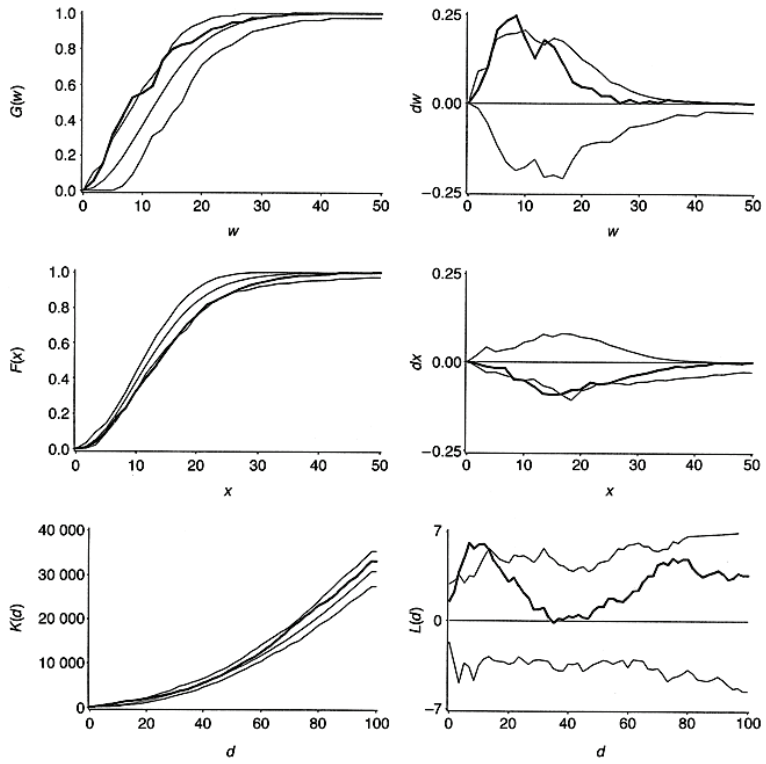


Figure 50 : Utilisation des fonctions F, G et K pour caractériser un semis de points agrégé. Figure extraite de Barot *et al.*(1999).

Le test de significativité est réalisé par la méthode de Monte-Carlo : on génère un grand nombre de jeux de données aléatoires correspondant à l'hypothèse nulle et on calcule dx et dw pour chacune d'elle. On élimine les valeurs extrêmes correspondant au seuil de confiance choisi, et on obtient ainsi un intervalle de confiance de l'hypothèse nulle. Si dx est significativement négatif, ou dw significativement

positif, on peut conclure à une agrégation, et inversement à la dispersion. La Figure 50, extraite de Barot *et al.* (1999), illustre le comportement de F et G dans le cas d'un semis de points présentant des agrégats. Les courbes en gras sont les fonctions calculées d'après le jeu de points, les courbes claires les limites de l'intervalle de confiance de l'hypothèse nulle d'une distribution de Poisson homogène, définies comme les valeurs extrêmes de 500 simulations (équivalent à un seuil de confiance de 99,8%). dw est significativement positif, comme la fonction L de Ripley présentée à titre de comparaison, alors que dx est négatif.

Les points proches de la limite du domaine posent problème : s'ils sont plus proches de la bordure que de leur plus proche voisin observé, rien ne garantit qu'un point à l'extérieur du domaine ne soit pas en réalité le plus proche voisin, la distance est donc probablement surestimée. La méthode de correction des effets de bord classique (Diggle, 1979) consiste à éliminer tous ces points pour ne retenir que ceux dont un voisin est rencontré avant la limite du domaine. Une conséquence est la diminution du nombre de points, et donc une perte de puissance du test pour les jeux de données de petite taille. Diggle (1983) estime que la méthode est efficace dès trente points. Gignoux *et al.* (1999) développent une méthode de calcul sans correction des effets de bord, applicable à partir d'une quinzaine de points : aucun point n'est éliminé, mais le prix à payer est un léger biais dans l'estimation de dx et dw car les effets de bord ne sont pas les mêmes pour le jeu de points observé s'il est agrégé ou régulier et la distribution de référence. Les auteurs estiment que ce biais est faible même pour des structures très prononcées et que le gain en puissance du test compense largement cet inconvénient.

Skellam (1952) avait établi que sous l'hypothèse d'une distribution complètement aléatoire, la densité de probabilité que le plus proche voisin de chaque point se trouve à la distance r suivait :

$$f(r) = 2\lambda r e^{-\lambda r^2} \quad \textbf{(130)}$$

λ est le nombre de points moyen dans un cercle de rayon 1, et non par unité de surface. Pielou (1959) propose d'utiliser la statistique α :

$$\alpha = \pi D \bar{\omega} \quad \textbf{(131)}$$

D est la densité de points observée (le nombre de points moyen par unité de surface), $\omega = r^2$ et $\bar{\omega}$ est la moyenne du carré de la distance au plus proche voisin de chaque point. Pielou montre, à partir de (130), que α vaut $(n(A) - 1)/n(A)$ pour un processus de Poisson homogène (et que $2n(A)\alpha$ suit une distribution de χ^2 à

$2n(A)$ degrés de liberté), plus en cas d'agrégation et moins en cas de dispersion. Les effets de bord sont ignorés. Ce test oublié est clairement l'ancêtre de *G*.

La fonction *K* de Ripley fournit plus d'information que les fonctions *F* et *G* puisqu'elle prend en compte toutes les distances, donc tous les voisins et non seulement le plus proche. Diggle (1983) défend l'usage de *F* et *G* pour leur plus grande puissance à détecter certaines structures. D'après lui, les tests peuvent être classés dans l'ordre décroissant de leur aptitude à détecter le non-respect de l'hypothèse nulle d'une distribution complètement aléatoire : pour l'agrégation, *F* puis *K* puis *G*, et pour la régularité, *K* puis *G* puis *F*. Ces résultats sont remis en cause par Gignoux *et al.* (1999) qui affirment que la puissance de *F* et *G* est similaire pour la détection de l'agrégation. Diggle conseille d'utiliser systématiquement les trois statistiques, qu'il estime complémentaires. En pratique, avec le développement de la puissance de calcul (nécessaire particulièrement pour *K*), les études utilisent très largement la fonction de Ripley et de plus en plus rarement *F* et *G*. Un rare exemple de l'utilisation conjointe des trois outils est donné par Barot *et al.* (1999).

La fonction *J* de Van Lieshout et Baddeley (1996)

Van Lieshout et Baddeley ont introduit la fonction *J* :

$$J(r) = \frac{1 - G(r)}{1 - F(r)} \qquad \textbf{(132)}$$

Cette fonction combine l'utilisation des deux fonctions de Diggle. Elle présente les caractéristiques suivantes :

- Une valeur inférieure à 1 indique l'agrégation, une valeur supérieure la régularité,
- $J(r) = 1$ pour tout r dans le cas d'un processus de Poisson homogène,
- Pour des processus markoviens dont la distance d'interaction est r_{max}, $J(r)$ est constant pour $r > r_{max}$.

Un exemple d'application est donné dans Illian *et al.* (2008, p.213).

Le test T_k de Cuzick et Edwards (1990)

Cuzick et Edwards (1990) introduisent un test non paramétrique permettant de détecter une structure spatiale (concentration ou dispersion) dans un semis de point non homogène, applicable à la caractérisation des maladies rares.

Le semis de points est divisé en cas (les points étudiés) et contrôles. Les contrôles forment la distribution de référence et le test est, comme le remarque Diggle (1990) dans la discussion qui suit l'article, un test d'étiquetage aléatoire : la question est de savoir si la population des cas diffère significativement de celle des contrôles. Les contrôles peuvent être un échantillon représentatif de la population totale pour des raisons pratiques : les auteurs étudient la répartition spatiale des maladies rares et la comparent à un échantillon de population le plus représentatif possible, notamment en ce qui concerne les variables relatives au sujet traité (âge et sex-ratio dans ce cas précis). La valeur du test T_k est la somme pour tous les cas du nombre de cas parmi les k plus proches voisins. Formellement, on numérote les $n(A)$ points et on les note x_i. On définit d_i^k comme le nombre de cas parmi les k plus proches voisins du point x_i. Pour k choisi, T_k est défini comme la somme des d_i^k pour tous les cas :

$$T_k = \sum_{i=1}^{\mathrm{n}_c(A)} d_i^k \qquad \textbf{(134)}$$

(133) : Test T_k de Cuzick et Edwards

Les auteurs définissent également plusieurs tests apparentés : T_{run}, le nombre de cas rencontrés autour de chaque cas avant la rencontre du premier contrôle et T_k^{inv}, le nombre de cas rencontrés jusqu'à l'apparition du $k^{\text{ième}}$ contrôle.

L'espérance de T_k est connue pour l'hypothèse nulle de l'étiquetage aléatoire. Son intervalle de confiance est calculé par les auteurs ou peut être établi par la méthode de Monte Carlo. Un exemple est donné (Figure 51) concernant les cas de leucémie dans une région anglaise. Les 62 cas détectés sont visuellement très concentrés, mais la question consiste à savoir si cette agrégation est seulement celle de la population dans une zone urbaine ou si une structure différente peut être mise en évidence. Les contrôles sont 141 personnes choisies au hasard dans le registre des naissances. Les résultats sont donnés pour différentes valeurs de k et permettent de conclure à une concentration significative au seuil de 99% pour T_1 à T_4 et T_2^{inv}.

La valeur de k la plus intéressante ne peut pas être établie *a priori*. Elle dépend évidemment de la proportion de cas par rapport aux contrôles (plus elle est faible, plus le nombre de voisins devra être grand pour trouver des cas parmi eux). Les auteurs proposent une combinaison de plusieurs T_k pour obtenir un test combiné (T_{comb}). Le choix de plusieurs valeurs de k n'est en fait pas plus évident que celui d'une seule et on peut préférer calculer T_k pour une large plage de valeurs de k et choisir a posteriori la plus significative. S'agissant d'un test d'écart à l'hypothèse nulle, il suffit qu'une valeur soit significative pour que le test soit concluant.

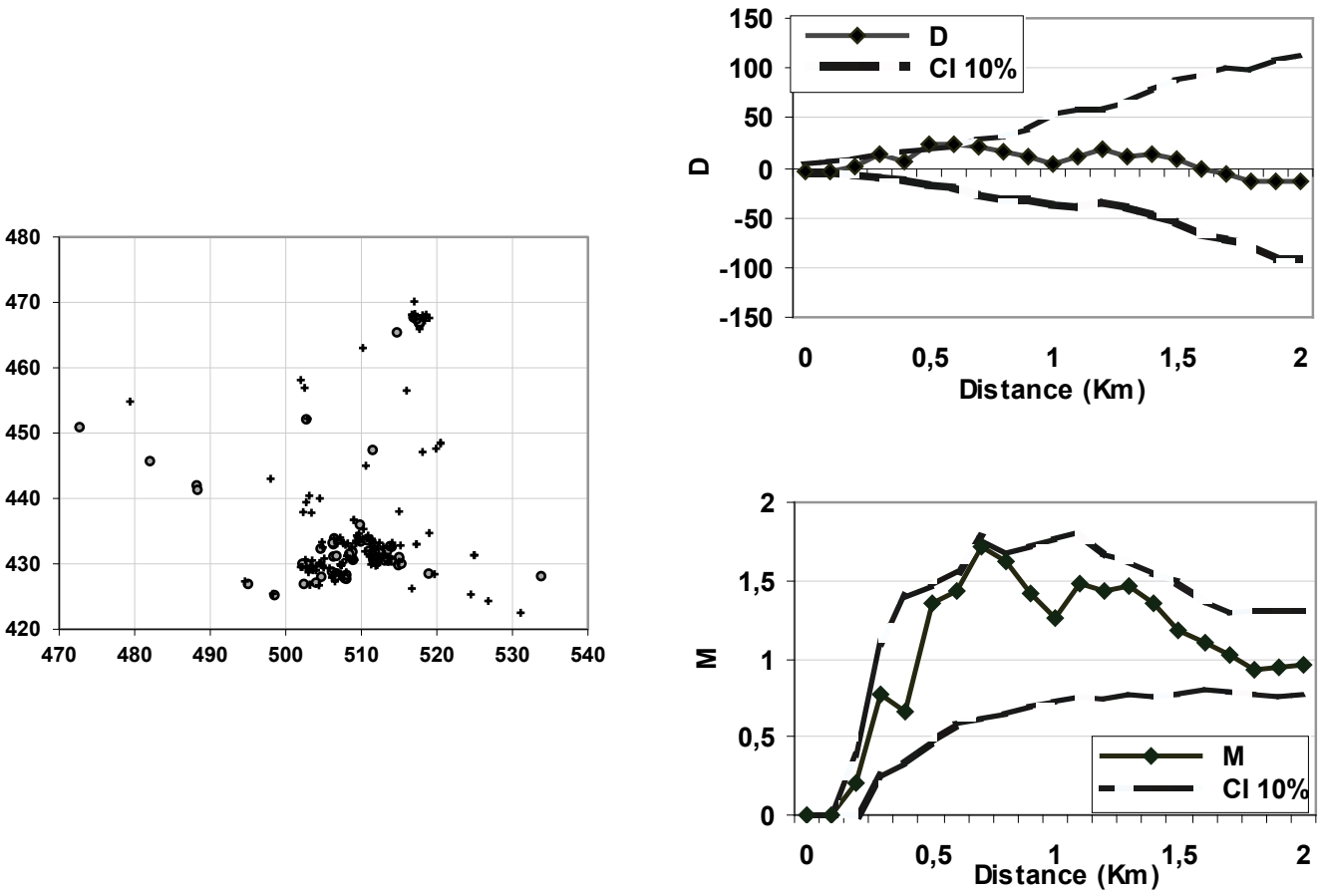


Figure 51 : Données de Cuzik et Edwards. Carte des points et fonctions *D* et M_{cas}. Les cas sont représentés par des cercles et les contrôles par de croix.

La conclusion du test porte seulement sur l'existence ou non d'une concentration significative (T_k supérieur à la valeur attendue : on trouve plus de cas parmi les k plus proches voisins que dans le cadre de l'hypothèse nulle) ou une dispersion significative (T_k inférieur à la valeur attendue). Diggle (1990) suggère que la fonction D (qu'il formalisera dans Diggle et Chetwynd, 1991) donne cette information, et permet en plus de préciser à quelle distance cette structure est présente. La courbe de D ainsi que celle de M (dans sa version cas-contrôles, page 62) sont présentées sur la figure : les deux donnent les mêmes résultats et détectent les agrégats, mais souffrent d'un manque de puissance en raison du faible nombre de contrôles (les intervalles de confiance locaux sont donnés au seuil de 10%).

La valeur de k permet de déterminer à quelle échelle la structure apparaît, mais en terme de nombre de voisins. Une représentation de la significativité de l'écart à l'hypothèse nulle de la valeur de T_k en fonction de k serait très similaire à la fonction D standardisée, le nombre de voisins remplaçant la distance.

La fonction Δ***MCI*** de Podani et Czárán (1997)

L'analyse intertype de nombreux types de points deux à deux pose des difficultés pratiques (l'analyse de $\frac{I(I-1)}{2}$ courbes pour I espèces) et théoriques : si toutes les précautions sont prises dans la construction des fonctions K et M intertypes pour que les structures détectées ne soient pas la conséquence des structures de chacun des types de points, rien ne garantit que la structure caractérisée par la fonction K_{23} ne soit pas due seulement aux valeurs de K_{12}et K_{13}, les fonctions n'étant pas indépendantes.

Une solution attrayante consiste à évaluer la structure spatiale intertype dans son ensemble, tous types de points confondus, quitte à se pencher sur chaque type en détail par la suite. C'est l'objet du travail de Podani et Czárán (1997).

Définition

On considère I espèces de points, notées i. On définit une *combinaison d'espèces* par le vecteur à I dimensions composé d'une suite de 0 et de 1, selon que l'espèce i est absente ou présente. Il existe donc 2^I combinaisons d'espèces différentes, notées F_j.

Sous l'hypothèse d'une répartition complètement aléatoire (CSR), les points de l'espèce i sont la réalisation d'un processus de Poisson dont l'intensité est estimée par $\hat{n}_i(A)/A$. La probabilité de ne trouver aucun point de l'espèce i dans la surface $S \subset A$ est $p_i^0(r) = e^{-\frac{\hat{n}_i(A)}{A}S}$. La probabilité de la présence d'une espèce dans un cercle de rayon r centré sur un point quelconque est donc $p_i^1(r) = 1 - e^{-\frac{n_i(A)}{A}\pi r^2}$. La probabilité de trouver une combinaison d'espèces F_j dans le cercle est le produit des probabilités de présence ou d'absence des espèces la composant. On la note $P(F_j, r)$.

Un exemple permet de clarifier ces notions. Supposons 3 espèces dont les probabilités de présence dans un cercle de rayon r donné sont respectivement 0,1, 0,2 et 0,7. La probabilité de trouver la combinaison d'espèces $F_1 = (0; 1; 1)$ est $P(F_1, r) = (1 - 0.1) \times 0{,}2 \times 0{,}7$.

La théorie de l'information (voir page 90) définit la quantité d'information apportée par l'observation d'un évènement E par la valeur $I(E) = -\ln P(E)$. Plus l'évènement est rare, plus sa survenue apporte de l'information. $I(F_k, r) = -\ln P(F_k, r)$ est la quantité d'information apportée par la combinaison d'espèces F_k observée autour du point x_k.

On se place maintenant autour de chacun des $n(A)$ points du système étudié. Pour une valeur de r donnée, on peut calculer la quantité d'information moyenne, notée *MCI* (*Mean Compositional Information*) : $MCI(r) = \frac{\sum_{k=1}^{n(A)} I(F_k,r)}{n(A)}$.

L'espérance sous l'hypothèse nulle d'indépendance complète de la quantité d'information des combinaisons d'espèces observées dans un cercle de rayon r est la somme pondérée par leur probabilité des quantités d'information des toutes les combinaisons : $\mathbb{E}(MCI(r)) = \sum_{j=1}^{2^n} P_r(F_j) I_r(F_j) = -\sum_{j=1}^{2^n} P_r(F_j) \ln P_r(F_j)$.

La fonction $\Delta MCI(r)$ est définie par la différence entre la quantité d'information observée et la quantité d'information attendue.

La significativité statistique de la valeur observée est assurée par le calcul d'un intervalle de confiance de l'hypothèse nulle (CSR) par la méthode de Monte-Carlo.

Interprétation

Les auteurs estiment (p. 262) :

- Qu'une valeur négative correspond à une répulsion : les combinaisons d'espèces autour des points sont « moins nombreuses et moins diverses » que dans le cas d'une distribution complètement aléatoire
- Qu'une valeur positive suggère l'agrégation : les combinaisons sont « plus nombreuses et plus rares ».

Ils nuancent cette interprétation en indiquant que les exemples montrent que l'interprétation n'est pas toujours simple.

Discussion

L'interprétation des auteurs peut être remise sérieusement en doute en analysant la construction de la fonction.

Si on fixe les effectifs $n_i(A)$ de chaque espèce, on fixe également $\mathbb{E}(MCI(r))$, la quantité d'information attendue qui sert de référence. $\Delta MCI(r)$ ne dépend alors plus que des combinaisons d'espèces rencontrées autour des points.

$\Delta MCI(r)$ n'est pas liée directement à la variété des combinaisons d'espèces autour des points. Partons d'une distribution complètement aléatoire, pour laquelle $\Delta MCI(r)$ est égal à 0 (aux fluctuations stochastiques près) pour tout r. Pour augmenter la valeur de $MCI(r)$, il faut et il suffit de remplacer une combinaison autour d'un point par une combinaison plus rare. L'idée selon laquelle la valeur de $\Delta MCI(r)$ est liée à la variété des combinaisons est simplement due au fait que

la présence de nombreuses combinaisons rares nécessite qu'elles soient variées parce que les espèces rares ne peuvent pas être présentes autour de nombreux points.

$\Delta MCI(r)$ n'augmente pas non plus obligatoirement quand les nombre d'espèces augmente au voisinage des points. L'information apportée par une combinaison d'espèces F_k est $-\ln P(F_k, r)$. La probabilité de F_k est le produit des probabilités de présence de certaines espèces et d'absence des autres. Considérons $F_{k'}$ une combinaison obtenue en ajoutant une espèce nouvelle i à F_k. La probabilité de $F_{k'}$ est obtenue à partir de celle de F_k en remplaçant dans le produit la probabilité d'absence de la nouvelle espèce par sa probabilité de présence. $F_{k'}$ est moins probable que F_k si $p_i^1(r) < p_0^1(r)$, c'est-à-dire si $p_i^1(r) < 0{,}5$.

La répulsion est définie par les auteurs comme l'absence d'autres espèces autour des points considérés, l'agrégation par la présence de nombreuses espèces. En réalité, l'apparition d'une nouvelle espèce augmente la quantité d'information si et seulement si sa probabilité de présence est inférieure à 0,5. Pour les petites valeurs de r, on conçoit facilement que les probabilités de présence de toutes les espèces soient petites. Pour des grandes valeurs de r, la quantité d'information sera au contraire plutôt augmentée par l'absence d'espèces (à la limite, quand r atteint la taille du domaine, toutes les probabilités de présence tendent vers 1). On devra donc limiter l'application de $\Delta MCI(r)$ aux rayons pour lesquels les probabilités de présence des espèces sont inférieures à 0,5.

Enfin, par construction, la fonction $\Delta MCI(r)$ traite indifféremment la structure propre de chaque type de point et les structures dues à leurs interactions. Le premier exemple de Podani et Czárán (figure 2, p. 265) en est la meilleure illustration : plusieurs agrégats de points monospécifiques distribués indépendamment les uns des autres sont analysés. La fonction $\Delta MCI(r)$ détecte une répulsion significative à des distances de l'ordre de grandeur des agrégats. L'analyse par la fonction K intertype aurait conclu à l'indépendance entre les types de points, son hypothèse nulle prenant en compte la structure propre de chacune des espèces.

En pratique, la fonction paraît inutilisable parce que son écart à l'hypothèse nulle peut être due à des causes très différentes.

La méthode de reconnaissance des agrégats de Coomes *et al.* (1999)

Présentation

La méthode permet seulement de traiter des semis de points agrégés.

L'idée est de regrouper successivement les paires de points les plus proches jusqu'à ce que le semis de points obtenu ne soit plus distinguable d'un processus

de Poisson homogène. Alors, le nombre de points restant est le nombre d'agrégats, dont on connaît le nombre d'individus. Formellement, le procédé est le suivant :

- Vérification de l'existence d'agrégats : la probabilité qu'un point ait au moins un voisin à une distance r donnée, c'est-à-dire la fonction G de Diggle (1983) est calculée pour le semis de points et pour le processus de Poisson de référence ($G_0(w) = 1 - e^{-\pi\lambda w^2}$pour le processus de Poisson de densité λ,aux corrections des effets de bord près). On calcule la fonction G pour le semis de points réel, et on définit dw comme la plus grande différence entre les deux fonctions. Le test de Kolmogorov-Smirnov (Diggle, 1979) est ensuite appliqué : on calcule dw entre un grand nombre de simulations du processus de Poisson de référence et la fonction $G_0(w)$, on en tire une valeur limite de dw pour l'hypothèse nulle (méthode de Monte-Carlo), et on compare finalement la valeur de dw du semis de point réel au seuil. Si dw est inférieur à la valeur de référence au seuil de confiance choisi, on accepte l'hypothèse qu'il n'y a pas d'agrégats.
- Si la distribution est agrégée d'après le test précédent, on multiplie la distance entre chaque paire de points par le poids cumulé des deux points (au départ : 1), et on fusionne la paire de points ayant la plus petite distance pondérée. Le centre de gravité est affecté du poids total de la paire de points supprimée.
- On teste à nouveau l'agrégation et on continue les regroupements jusqu'à l'acceptation de l'hypothèse d'absence d'agrégats. On connaît alors le nombre d'agrégats et leur rayon moyen.

Cette méthode peut également servir à définir l'hypothèse nulle de l'indépendance des populations dans le cadre d'une analyse intertype (par exemple K_{12}). On a vu, page 71, que l'hypothèse nulle de l'indépendance de deux populations est difficile à tester parce qu'elle doit intégrer la structure propre de chacune des populations. La solution proposée ici consiste à distribuer aléatoirement les centres des agrégats, qui suivent par construction un processus de Poisson homogène, puis à régénérer autour de chacun de ces centres des agrégats correspondant aux caractéristiques du semis de points observé. Les auteurs ne précisent pas le processus de reconstruction, mais on peut supposer qu'il s'agit d'un processus de Neyman-Scott (on connaît le nombre d'agrégats, leur nombre de points et leur rayon moyens).

Discussion

Il s'agit d'une méthode complémentaire à la fonction de Ripley, qui ne donne pas les mêmes informations. Les auteurs la testent sur des semis de points particuliers qui permettent d'arriver aux conclusions suivantes :

- Dans certains cas, le regroupement des agrégats est plus puissant que la fonction de Ripley : il détecte l'agrégation alors que K ne rejette pas l'hypothèse nulle. Cette affirmation, étayée par un exemple (p. 563) dont le détail des calculs n'est pas donné pour la fonction K, est surprenante car le test de l'agrégation est ici fondé sur le seul plus proche voisin de chaque point, alors que la fonction de Ripley prend en compte tous les voisins. Le reste du test est similaire : on compare le semis de points réel à une distribution théorique par la méthode de Monte-Carlo.
- La méthode fonctionne parfaitement bien pour des distributions dont les agrégats sont distincts (Figure 2, page 32). Quand les agrégats s'interpénètrent (Figure 14, page 58), le nombre détecté dépend fortement du seuil de probabilité choisi pour l'acceptation de l'hypothèse nulle. Au seuil classique de 5%, on considère que l'hypothèse nulle n'est rejetée que si la probabilité que la distribution observée en soit une réalisation est inférieure à 5%, elle est donc assez facilement acceptée. Les auteurs utilisent des seuils bien moins restrictifs (par exemple 50%), qui conduisent à accepter l'hypothèse nulle beaucoup plus tard, donc à continuer le regroupement plus longtemps. En d'autres termes, augmenter le seuil conduit à définir moins d'agrégats, de plus grande taille. Les auteurs conseillent de tester diverses valeurs de seuil.
- Enfin, le processus ponctuel sous-jacent ne peut évidemment pas être caractérisé par une seule de ses réalisations. Les agrégats détectés sont les plus vraisemblables, ou du moins les plus apparents. Dans le cas d'un processus créant peu d'agrégats de grande taille, qui s'interpénètrent fortement, le nombre d'agrégats détecté est très supérieur au nombre réel. On peut donc se poser la question de la validité de l'utilisation de la méthode pour générer des semis de points pour tester l'hypothèse nulle d'indépendance de deux populations puisque le processus utilisé n'est pas forcément proche de celui qui a généré la population observée.
- Une dernière limite dans l'application de la méthode est due à la difficulté de sa mise en œuvre : la courbe de la fonction $G_0(w)$ pour le processus de Poisson de référence nécessite le calcul de la correction des effets de bord. La valeur seuil de la différence de probabilité entre la distribution observée et la courbe de référence nécessite ensuite la simulation de jeux de points pour l'application de la méthode de Monte-Carlo. L'utilisation de l'outil n'est donc pas envisageable sans une phase de développement informatique lourde.

Les fonctions intertypes à marques continues

La limite des fonctions intertypes vues au premier chapitre est qu'elles ne peuvent évaluer les interactions qu'entre deux types de points parmi un nombre fini et pas trop grand. Comment traiter alors les marques continues ?

Une question classique à traiter est la structure spatiale d'un peuplement forestier en fonction du diamètre. Une première réponse est de définir des classes de diamètre (par exemple petits, moyen et gros arbres) et de les traiter par une fonction intertype discrète. On peut émettre une réserve majeure sur cette méthode, outre le côté arbitraire de la classification et l'augmentation rapide du nombre de courbes à analyser avec le nombre de classes : rien ne garantit qu'une interaction détectée entre les gros arbres et les petits ne soit pas qu'un artefact résultant d'une attraction ou d'une répulsion réelle entre les petits arbres et les moyens d'une part, les moyens et les gros d'autre part. L'hypothèse nulle de chacune des fonctions intertypes prend bien en compte les structures des deux types de points mais pas les possibles interactions avec un troisième.

La solution est probablement dans les fonctions intertypes à marques continues, capables de traiter le diamètre en tant que variable continue et de donner des informations sur l'attraction ou la répulsion des arbres de même diamètre ou de diamètres très différents.

La fonction K_{cor} de Goreaud

Goreaud (2000) définit la fonction intertype à marques continues K_{cor} à partir de la définition du corrélogramme utilisé en géostatistiques.

On note m_i la valeur de la marque du point x_i. Le variogramme $\gamma(r)$ (Matheron, 1970) est défini par l'espérance du carré de la différence des marques des couples de points situés à la distance r l'un de l'autre :

$$\gamma(r) = \frac{1}{2}\mathbb{E}\left[\left(m_i - m_j\right)^2 | d(i,j)\right] = r \qquad \textbf{(134)}$$

Notons dès maintenant que le variogramme s'applique à un champ de points (c'est-à-dire un ensemble de variables aléatoires appliquées chacune à un point du plan), et non à des processus ponctuels (un ensemble de points discontinu). Cependant, le calcul du variogramme est réalisé à partir d'un échantillon de points et les calculs peuvent être formellement les mêmes. Une hypothèse préalable à l'utilisation du variogramme est la stationnarité de la variable traitée.

Dans le cadre des processus ponctuels, la distribution de la marque doit également être homogène.

Le corrélogramme est une expression différente de la même mesure :

$$\mathrm{Cor}(r) = \frac{\mathrm{Cov}\left(m_i, m_j | d(i,j) = r\right)}{\mathrm{Var}(m)} = \frac{\mathbb{E}\left[\left(\left(m_i - \mathbb{E}(m)\right)\left(m_j - \mathbb{E}(m)\right) | d(i,j) = r\right)\right]}{\mathrm{Var}(m)} \quad \textbf{(135)}$$

On montre aisément que $\gamma(r) = (1 - \mathrm{cor}(r))\,\mathrm{Var}(m)$.

Si les valeurs des marques sont indépendantes à une distance r donnée, le corrélogramme est nul. Si les valeurs sont corrélées positivement ou négativement, le corrélogramme se rapproche de 1 ou -1.

Goreaud définit la fonction K_{cor} comme la valeur du corrélogramme autour de chaque point dans un cercle de rayon r, par analogie avec la fonction K :

$$\boxed{K_{COR} = \frac{\mathrm{Cov}\left(m_i, m_j | d(i,j) < r\right)}{\mathrm{Var}(m)}} \quad \textbf{(136)}$$

(136) : Définition de la fonction K_{cor} de Goreaud

On note N_r le nombre de couples de points à distance inférieure ou égale à r et $\bar{m}$ la valeur moyenne de la marque. Un estimateur de la fonction est :

$$\widehat{K}_{COR}(r) = \frac{1}{\widehat{\mathrm{Var}}(m) N_r} \sum_{i=1}^{n(A)} \sum_{j=1, i \neq j}^{n(A)} (m_i - \bar{m})(m_j - \bar{m}) \mathbf{1}\left(\|x_i - x_j\| \leq r\right) c(i,j,r) \quad \textbf{(137)}$$

Le nombre de couples de points N_r est corrigé des effets de bord par Goreaud (2000). On peut également envisager, plus simplement, de ne réaliser aucune correction et d'utiliser la valeur du nombre de couples de points N_r réellement observée. Dans ce cas, les points proches des bords pèsent moins dans la valeur de K_{cor} mais aucun biais n'est introduit.

La valeur de K_{cor} se situe entre -1 et 1. L'hypothèse nulle est la distribution complètement aléatoire des marques sur le semis de points existant. On se référera à Goreaud (2000 p. 142-145), pour des exemples.

La fonction K_{mm} de Penttinen *et al.*

Penttinen *et al.* (1992) définissent une fonction intertype à marques continues en partant de la fonction K de Ripley.

Définition

On suppose que le processus ponctuel est homogène (sa densité est une tante λ) et isotrope (la fonction g ne dépend que de la distance r entre les deux surfaces élémentaires dx_1 et dx_2). On introduit la valeur des marques des points (par exemple le diamètre des arbres) dont la valeur moyenne est μ. On définit enfin une fonction $M(r)$ égale à la moyenne du produit des marques dans ces deux surfaces élémentaires :

$$M(r) = \lambda^2 g(r) k_{mm}(r) dx_1 dx_2 \quad \textbf{(138)}$$

$M(r)$ est égale à la probabilité de trouver deux points dans les deux surfaces élémentaires (102) multipliée par la fonction $k_{mm}(r)$. Dans le cas d'une distribution complètement aléatoire des marques, $k_{mm}(r) = \mu^2$. La fonction $k_{mm}(r)/\mu^2$ est analogue à la fonction g : les auteurs l'appellent *fonction de corrélation des marques*. Si les marques varient dans le même sens pour une distance r donnée (corrélation positive), $k_{mm}(r)$ sera supérieur à μ^2, et inversement $k_{mm}(r)$ sera inférieur à μ^2 en cas de corrélation négative des marques à la distance r.

On définit alors la fonction $K_{mm}(r)$:

$$\boxed{K_{mm}(r) = \int_{\rho=0}^{r} k_{mm}(\rho) g(\rho) 2\pi\rho d\rho} \quad \textbf{(139)}$$

(139) : Définition de la fonction K_{mm} de Penttinen *et al.*

La fonction K_{mm} vaut $\mu^2\pi r^2$ dans le cas d'une distribution complètement aléatoire des points et des marques. La fonction $L_{mm}(r) = \sqrt{\frac{K_{mm}(r)}{\pi\mu^2}}$ est plus facile à interpréter puisque sa valeur de référence est 0. On peut noter qu'il aurait été plus approprié de normaliser la fonction K_{mm} par μ^2 plutôt que la fonction L_{mm} pour conserver une analogie parfaite avec la fonction de Ripley.

Pratiquement, en reprenant les notations de (7), la fonction K_{mm} est estimée par :

$$\widehat{K}_{mm}(r) = \frac{1}{\hat{\lambda} n(A)} \sum_{i=1}^{n(A)} \sum_{j=1, i\neq j}^{n(A)} m_i m_j \mathbf{1}\left(\|x_i - x_j\| \leq r\right) c(i,j,r) \quad \textbf{(140)}$$

Discussion

Une hypothèse indispensable bien que jamais explicitée par les auteurs est la distribution homogène des marques. La valeur de $k_{mm}(r)$ n'est interprétable que dans ce cadre. Revenons à l'exemple du peuplement forestier : si on considère sans précautions deux sous-peuplements, l'un composé de petits arbres, l'autre de gros, il est clair que la valeur de $k_{mm}(r)$ sera supérieure à μ^2, sans qu'il n'y ait d'interaction entre les gros arbres et les petits. Cette grande valeur est seulement due à la non stationnarité du processus. Cette limite est exactement la même que pour la fonction g, qui n'a de sens que pour un processus stationnaire.

Cette hypothèse étant fixée, la fonction K_{mm} est assez intuitive : on peut imaginer de remplacer un point de poids n par n points superposés (si les poids ne sont pas entiers, on peut les multiplier par une puissance de 10). L'hypothèse de distribution homogène des marques implique que le processus ponctuel obtenu est lui-même homogène. Il peut être étudié par la fonction K de Ripley. En rapprochant (139) de (3), on voit immédiatement que la propriété de second ordre de ce nouveau processus est le produit des fonctions k_{mm} et g. La fonction de Penttinen *et al.* analyse donc un processus ponctuel de la même façon que la fonction de Ripley mais en décomposant sa propriété de second ordre :

- La première composante, g, est liée à la localisation des points
- La deuxième composante, k_{mm}, est liée à la corrélation entre les marques.

Application

Les auteurs donnent trois exemples d'application sur des peuplements forestiers. La démarche est toujours la même :

1. Calculer la fonction K classique au semis de points, et la fonction g correspondante.
2. À partir des courbes de K et de g, choisir un processus ponctuel connu permettant de décrire le semis de points. Dans le cas le plus simple, si la fonction K ne sort pas de l'intervalle de confiance de l'hypothèse nulle, ce sera un processus de Poisson homogène. Dans les deux premiers exemples de Penttinen *et al.* (1992), il s'agit d'un processus de Gibbs et d'un processus agrégatif de Matérn. Ce processus de référence est pris comme hypothèse nulle, il est validé si la courbe de K ne sort pas de son intervalle de confiance.
3. Calculer la fonction K_{mm} pour le semis de points. L'hypothèse nulle est la distribution aléatoire des marques ($k_{mm}(r) = \mu^2$) sur le processus ponctuel validé à l'étape précédente.

Il s'agit bien de détecter d'abord la structure du semis de points non marqué, et de détecter ensuite la structure de la distribution des marques, celle des points étant connue.

La difficulté se situe à la deuxième étape, où un processus connu convenable doit être trouvé. Dans le troisième exemple, les auteurs n'y parviennent d'ailleurs pas et donnent la structure de la distribution des marques sans hypothèse nulle de référence. Il est alors impossible de conclure rigoureusement.

Conclusion

La fonction K_{mm} est un outil remarquable pour la détection de la structure de la distribution de marques continues dans un semis de points stationnaire. Elle présente toutefois deux limites importantes :

- La distribution des marques doit être elle-même stationnaire, ce qui constitue une seconde hypothèse forte après la stationnarité de la distribution des points et limite la portée théorique de la fonction.
- La structure du semis de point non marqué doit pouvoir être modélisée pour servir d'hypothèse nulle dans l'étude de la distribution des marques. Il s'agit d'une difficulté pratique parfois insurmontable.

On remarquera enfin la grande similitude entre les fonctions K_{COR} et K_{mm}, rendue particulièrement visible par l'expression de leurs estimateurs (137) et (140). La fonction K_{COR} présente l'avantage d'être plus simple à utiliser puisqu'elle ne nécessite pas de modélisation du processus ponctuel non marqué. En revanche, elle ne permet pas de distinguer l'effet de la localisation des points de l'effet de la répartition des marques.

Annexe 3 : Méthodes alternatives en statistiques spatiales discrètes

De nombreux outils ont été développés principalement par les économistes pour détecter la concentration spatiale dans des jeux de données discrètes, en général des effectifs par zone administrative. Ceux qui ont une importance historique mais ne contribuent pas aux développements sur l'entropie sont passés en revue ici.

Une approche légèrement différente de la question est apportée par les méthodes mesurant l'autocorrélation spatiale. Bien qu'elles s'écartent du sujet, elles sont citées ici parce qu'elles peuvent être utilisées en complément des premières méthodes pour détecter un éventuel problème de MAUP (voir page 111). Elles permettent d'obtenir des informations complémentaires à celles des méthodes fondées sur les quadrats en ne mesurant pas la variabilité totale, mais les écarts entre des localisations voisines. Les indices d'autocorrélation spatiale (Moran et Geary sont les plus connues) et les corrélogrammes/variogrammes sont présentés.

Outils de détection de la concentration spatiale

La variance relative

La forme générale de l'indice a été introduite par Clapham (1936). Ses propriétés ont été étudiées par Hoel (1943) à qui on en attribue donc souvent la paternité.

On découpe la zone d'étude en quadrats de taille égale dans lesquels on compte le nombre de points. Selon le protocole, les quadrats peuvent être un simple échantillon du domaine. Le processus de Poisson homogène a une variance égale à son espérance. Le rapport de la variance à l'espérance donne une indication sur la dispersion : il est égal à 1 pour une distribution complètement aléatoire, une valeur supérieure indique l'agrégation, une valeur inférieure la dispersion. Plusieurs versions et améliorations successives existent dans la littérature (Upton et Fingleton, 1985 p. 29-32). On retiendra les plus courantes (n est le nombre de quadrats, $\bar{x}$ la moyenne du nombre de points par quadrat et s la variance) :

- L'indice de dispersion : $ID = (n-1)\frac{s^2}{\bar{x}}$ a une distribution de χ^2 à $n-1$ degrés de libertés pour l'hypothèse nulle de la distribution de Poisson (Hoel, 1943), ce qui permet de tester la significativité des valeurs observées.
- L'indice de la taille des agrégats (*Index of Cluster Size*) : $ICS = \frac{s^2}{\bar{x}} - 1$ dont la valeur positive peut être interprétée comme le nombre de points de chaque agrégat.

La difficulté principale réside dans le choix de la taille des quadrats : trop petits, ils contiennent trop peu de points, mais ils restituent d'autant moins d'information géographique qu'ils sont grands. La méthode de Greig-Smith tente d'y remédier en analysant plusieurs échelles successivement.

La méthode de Greig-Smith (1952)

La méthode des quadrats contigus, introduite par Greig-Smith (1952) est un grand classique, malgré ses limites théoriques et pratiques sérieuses. Une analyse détaillée est disponible dans Upton et Fingleton (1985 p. 34). Une simple présentation en est faite ici.

La méthode consiste à mesurer la variance du nombre d'individus à l'intérieur de surfaces doublées à chaque étape.

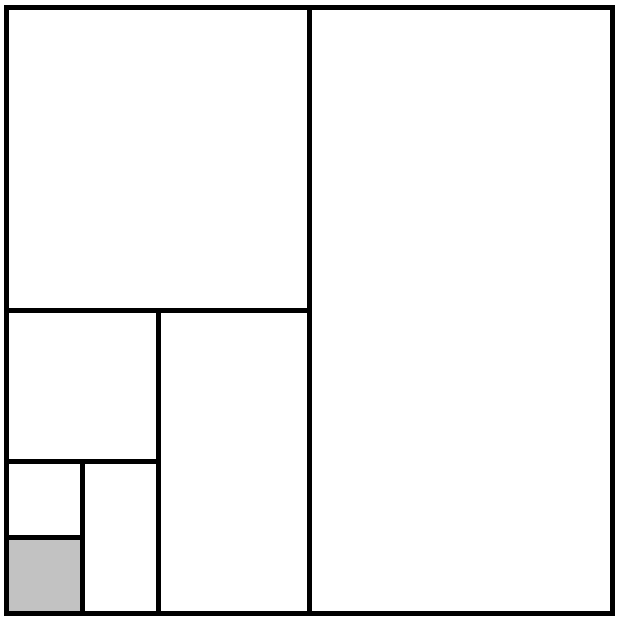


Figure 52 : Découpage successif des quadrats de Greig-Smith

La Figure 52 illustre le procédé : on calcule la somme des carrés des nombre d'individus en commençant par les 64 quadrats élémentaires, de la taille de celui représenté en bas à gauche. La somme est notée T_1. On double ensuite la surface des quadrats en les regroupant pour obtenir des rectangles (ici verticaux) répétant le motif des deux carrés en bas à gauche de la figure. La somme des carrés est notée T_2. On double la surface des quadrats jusqu'à couvrir toute la surface. Les dernières valeurs sont T_{32}, la somme du carré du nombre d'individus se trouvant dans la moitié gauche et dans la moitié droite du domaine, et T_{64}, le carré du nombre total d'individus. On calcule ensuite pour chaque niveau de regroupement r la valeur $G_r = 2T_r - T_{2r}$. Dans le cadre d'une répartition complètement aléatoire, G_r est

constant. Un pic de G_r correspond à une agrégation de la taille de r, alors que la dispersion est détectée par une brusque chute de G_r à la taille de la maille. La significativité des pics est vérifiée en calculant le rapport G_r/G_1 qui suit, au moins approximativement, une loi de Fisher à $N/2r$ et $N/2$ degrés de liberté.

La critique principale faite à la méthode est la progression géométrique de la taille des quadrats, qui ne permet pas de détecter des structures dont la taille est intermédiaire.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1					1											
2							1		1							
3			1			1		1								
4	1	1						1	1	1	2					
5						1		1						2		
6	1			2	2	1	1			1						
7		1	1		1	1	1		1	1	1					
8			1		2	1						1		1		
9		1		1												
10				1	2		2			1		1			1	1
11		1		2	1							1				
12			1								1					
13			1		1				1	2		1	1			
14						2			1							
15				1		1			1							
16						1	1		2			1		1		

Figure 53 : Distribution de Atripex hylmenelytra dans la Vallée de la Mort, Californie (in Upton et Fingleton, 1985 fig. 1.18 p. 35)

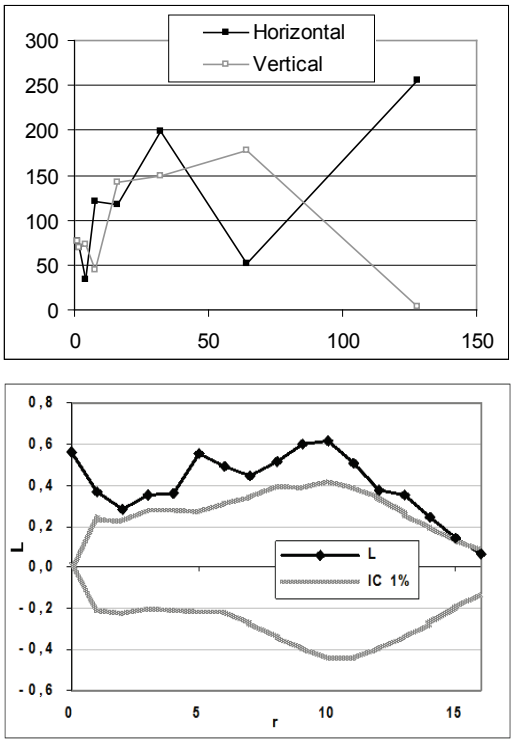


Figure 54: Courbes de G_r et $\hat{L}(r)$

L'exemple de la Figure 53 concerne la répartition spatiale d'un buisson pérenne dans une placette de 40m x 40m. Les coordonnées sont des carrés de 2,5 m de côté, les nombres les individus comptés dans chaque carré. La Figure 54 présente les résultats. G_r est tracé en regroupant les carrés pour obtenir des rectangles verticaux (comme précédemment) ou horizontaux (méthode alternative). La seule valeur de G_r significative (au seuil de 10%) est le pic de G_{32} obtenu par le regroupement horizontal. On peut donc conclure à l'existence d'agrégats de l'ordre de 4 à 8 carrés de dimension. On remarquera que les valeurs sont assez différentes selon le choix de regrouper verticalement ou horizontalement. La fonction de Ripley a été calculée sur les mêmes données en attribuant aux points la position du centre du carré (d'où la valeur non nulle de $\hat{L}(0)$ due à des points superposés). L'incertitude sur la position des points est de l'ordre de 1. L'intervalle de con-

fiance de l'hypothèse nulle d'une répartition complètement aléatoire est calculé au seuil de 1% à partir de 1 000 simulations. On observe une agrégation significative à toutes les distances jusqu'à 16 carrés (40 m), avec un pic à la distance de 6 détecté par la méthode de Greig-Smith, mais ce n'est pas le seul. La puissance de calcul limitée de l'époque justifiait l'utilisation de la méthode de Greig-Smith dans les années 1950 à 1970, mais on utilisera maintenant avantageusement les fonctions de Ripley.

L'indice de Gini

Considérons une espèce et une parcelle forestière divisé en N placettes (indicées par n). Dans un premier temps, nous calculons la part relative de chaque placette n dans la surface terrière de l'espèce considérée, notée s_n. De la même manière, nous calculons la part de chaque placette n pour la surface terrière de l'ensemble des espèces, que nous notons x_n. Puis, nous classons par ordre croissant les valeurs des ratios s_n/x_n. D'après ce dernier résultat, la courbe (dite de Lorenz) représentant les parts cumulées de la surface terrière pour l'espèce (en ordonnée) et celles pour la surface terrière totale (en abscisse) peut être construite. L'indice de Gini[3] pour l'espèce, noté G, correspond à l'aire (A) comprise entre la première bissectrice et la courbe de Lorenz (*cf.* Figure 55). De nombreuses formulations existent, nous proposons la suivante :

$$G = 0{,}5 - \sum_{n=1}^{N} \frac{s_n x_n}{2} - \sum_{n=2}^{N} x_n \sum_{j=2}^{n} s_{j-1} \qquad \textbf{(141)}$$

(141) : Indice de Gini

Le calcul de cet indice est illustré sur la Figure 55. L'indice de Gini (aire A, grisée) est obtenu en faisant la différence entre l'aire comprise sous la première bissectrice (0,5) et celle comprise sous la courbe de Lorenz (sommes des aires des triangles et des rectangles).

Le cas d'une distribution égalitaire est obtenu lorsque la distribution de l'espèce suit celle de l'ensemble du peuplement. Graphiquement, la courbe de Lorenz et la première bissectrice sont confondues : l'aire A est nulle et l'indice de Gini est égal à zéro. Inversement, la concentration de l'espèce est maximale lorsque tous les

[3] Les travaux originaux, cités par Hoover (1936), sont : Gini (1913) et Lorenz (1905)

arbres de cette espèce sont localisés dans une seule placette : l'indice de Gini tend alors vers 0,5.[4]

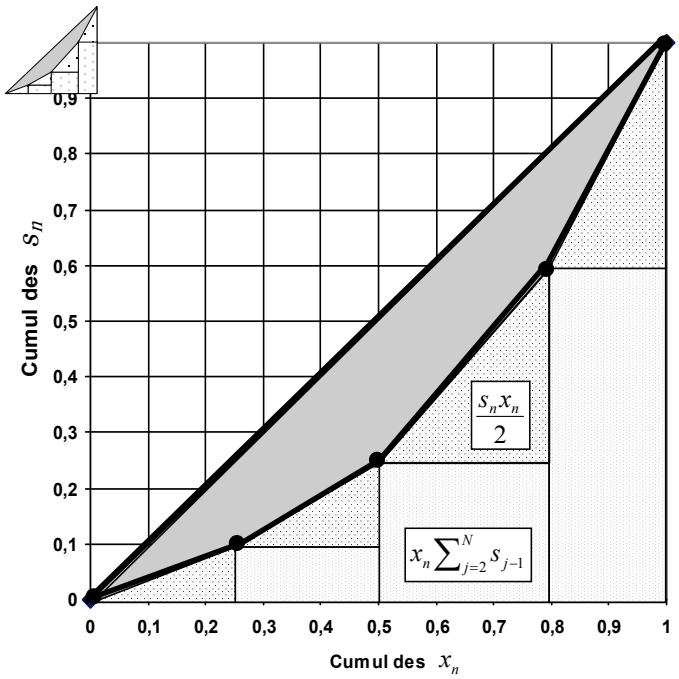


Figure 55 : Calcul de l'indice de Gini

L'indice est extrêmement commun en économie, mais rarement utilisé en écologie. En référence aux travaux de Hoover (1936), l'indice de Gini est également nommé « coefficient de localisation de Hoover » (cf. Kim, 1995). D'autres auteurs, préfèrent se référer aux travaux de Balassa (1965), le ratio est alors appelé « indice de Balassa » (Amiti, 1999).

L'indice d'Ellison et Glaeser (1997)

L'indice a été développé pour caractériser la localisation des établissements industriels dans différentes régions.

Le modèle sous-jacent est un choix de localisation en fonction :

- Des caractéristiques locales, regroupées sous le terme d'avantages naturels. En foresterie, on peut penser aux stations.
- Des externalités dues à la présence d'autres firmes. En forêt, les modèles rendant compte d'externalités positives entre les arbres sont rares (par

[4] L'indice de Gini peut-être également calculé comme le double de la surface comprise entre la première bissectrice et la courbe de Lorenz. Dans le cas d'une concentration maximale, la valeur de l'indice tendra non plus vers 0,5 mais vers 1 (Brülhart et Torstensson, 1996 ; Amiti, 1999 ; Holmes et Stevens, 2004).

exemple Goreaud *et al.*, 2002) alors que la concurrence, externalité négative tendant à la régularité du peuplement, est le mécanisme le plus évident. Le modèle d'Ellison et Glaeser ne prévoit pas la possibilité d'externalités négatives, c'est pourquoi il sera présenté pour plus de clarté ici dans son cadre original sans transposition en forêt.

Distribution complètement aléatoire

La distribution complètement aléatoire est figurée par un lancer de fléchettes sur une cible. Le nombre de fléchettes arrivant sur une zone de la cible dépend de la taille de la zone. La définition de la taille de la zone est donc cruciale. La tendance générale des activités humaines à se concentrer est considérée comme donnée, extérieure à la question, et doit donc être éliminée du calcul de l'indice. La taille de chaque zone de la cible, disons une région est définie comme une fraction de la taille totale de cible. On choisit généralement la part de la région dans l'emploi manufacturier total, mais ce n'est pas obligatoire. Cette taille est notée x_i. En écologie, la taille des établissements peut correspondre à la surface terrière des arbres, la taille d'une région à la surface terrière totale d'une placette.

S'il n'y a aucune force pour déterminer l'installation des firmes, l'espérance du nombre de firmes sera proportionnel à x_i.

Avantages naturels

Considérant une localisation, disons une région donnée, la probabilité qu'une firme d'une industrie donnée s'installe est définie par la réalisation d'une variable aléatoire p_i. L'espérance de p_i est nécessairement x_i pour que la distribution de l'emploi, toutes industries confondues, soit conforme à l'observation. Cependant, la région i présente certaines caractéristiques propres, utiles à certaines industries et défavorables à d'autre. La réalisation de p_i pour chaque industrie dépend de ces caractéristiques, regroupées sous le terme « avantages naturels ». Si les avantages naturels sont peu discriminants, toutes les industries auront la même valeur de p_i, c'est-à-dire x_i. La variance de p_i sera nulle. Dans le cas extrême inverse, les avantages naturels sont extrêmement discriminants et toutes les firmes de la région sont de la même industrie : $p_i = 1$ pour l'industrie en question, 0 pour les autres. La variance de p_i est alors maximale, égale à $x_i(1 - x_i)$. Entre ces deux cas extrêmes, Ellison et Glaeser définissent la variable γ^{na}, comprise entre 0 et 1, telle que la variance de p_i soit égale à $\gamma^{na} x_i(1 - x_i)$. La valeur de γ^{na} indique l'importance des avantages naturels dans l'implantation des firmes.

On peut remarquer que γ^{na} ne représente pas les avantages communs à toutes les industries, qui sont capturés par x_i, mais ceux qui font qu'une industrie s'implantera plutôt qu'une autre.

Externalités

Le choix d'implantation est ensuite influencé par les externalités. Les externalités sont ici des sources diverses de profit supplémentaire dues à l'existence de firmes de la même industrie dans la même région. Les externalités inter-régionales sont ignorées.

Formellement, les externalités interviennent sous la forme d'un multiplicateur de profit, proportionnel au nombre de firmes voisines, dans la même région de la même industrie et à un facteur γ^s qui est la probabilité qu'une firme voisine crée effectivement des externalités. Ce facteur γ^s indique l'importance des externalités : s'il est nul, le choix d'implantation ne sera pas lié aux autres firmes, mais seulement aux avantages naturels. S'il est égal à 1, l'importance des externalités est maximale.

Niveau de concentration géographique d'une industrie

Une façon de décrire la concentration géographique d'une industrie est de calculer l'écart entre sa distribution et celle de la totalité de l'emploi. Si on note s_i la part de la région i dans l'industrie, on peut définir l'indice de concentration géographique (parfois appelé indice de Hirschman-Herfindahl) par :

$$G = \sum_{i=1}^{M} (s_i - x_i)^2 \qquad \textbf{(142)}$$

Indice d'Ellison et Glaeser

Le résultat fondamental de Ellison et Glaeser est que l'espérance de G dépend de la structure géographique, de la concentration industrielle, et, de façon interchangeable, des avantages naturels et des externalités.

La structure géographique est décrite par la variabilité de la taille des régions. Elle est définie par :

$$D = 1 - \sum_{i=1}^{M} x_i^2 \qquad \textbf{(143)}$$

D varie entre 0 si une seule région contient tout l'emploi, et $1 - 1/M$ si les M régions sont de taille identique.

La concentration industrielle est définie par la variabilité de la taille des firmes, c'est-à-dire l'indice d'Herfindahl :

$$H = \sum_{k=1}^{N} z_k^2 \quad \textbf{(144)}$$

z_k est la part de la firme k dans la taille totale de l'industrie, son nombre d'employés divisé par le nombre d'employés de toute l'industrie. H varie de 1, si une seule firme de l'industrie emploie toute la main d'œuvre, à $1/N$, si toutes les N firmes sont de taille égale.

Il reste à définir γ, indicateur composite de l'intensité des avantages naturels et des externalités :

$$\gamma = \gamma^{na} + \gamma^{s} - \gamma^{s}\gamma^{na} \quad \textbf{(145)}$$

Alors, l'espérance de la concentration géographique est :

$$E(G) = D[\gamma + (1-\gamma)H] \quad \textbf{(146)}$$

G peut être estimé à partir des données pour obtenir un estimateur de γ :

$$\boxed{\hat{\gamma} = \frac{\hat{G} - DH}{D(1-H)}} \quad \textbf{(147)}$$

(147) : Indice d'Ellison et Glaeser

Dans le cadre d'une distribution des firmes sans avantages naturels ni externalités, $\gamma = 0$. Cette valeur est prise comme référence. Une valeur positive de $\hat{\gamma}$ dénote une concentration, une valeur négative une dispersion (bien que ce cas de figure ne soit pas prévu par le modèle théorique).

La variance de G est calculée par les auteurs pour l'hypothèse nulle d'absence de concentration, $\gamma = 0$:

$$V(G) = 2H^2\left[\sum_{i=1}^{M} x_i^2 - 2\sum_{i=1}^{M} x_i^3 + \left(\sum_{i=1}^{M} x_i^2\right)^2\right] - \sum_{j=1}^{N} z_j^4\left[\sum_{i=1}^{M} x_i^2 - 4\sum_{i=1}^{M} x_i^3 + 3\left(\sum_{i=1}^{M} x_i^2\right)^2\right] \quad \textbf{(148)}$$

L'intervalle de confiance de l'hypothèse nulle peut donc être calculé (en admettant la normalité de *G)*, et la significativité des valeurs de $\hat{\gamma}$ observées peut être donnée.

Dans la définition de l'indice d'Ellison et Glaeser, l'indice de concentration géographique G est la valeur calculée à partir des données. Dans le cadre d'une distribution complètement aléatoire des firmes, son espérance est DH. On pourrait utiliser cet indice tel quel, mais sa valeur n'est pas simple à interpréter. Si la localisation des firmes respecte le modèle sous-jacent, alors l'écart entre la valeur de G observée et DH est due à l'existence d'avantages naturels et d'externalités, dont la valeur synthétique γ peut être calculée.

L'avantage de γ est d'être indépendante du découpage en zones, de la taille de l'industrie et de sa concentration industrielle (la distribution de l'emploi entre ses firmes, considérée comme exogène). Son inconvénient est double : l'indice n'a de sens que si le modèle de choix est valide, ce qui est invérifiable, et sa valeur absolue n'a pas de signification intuitive, même si les auteurs donnent un certain nombre d'exemples permettant de fixer les idées du lecteur.

Le passage de G à γ répond donc à un besoin de lisibilité de la valeur de la concentration obtenue. Le simple calcul de G et sa comparaison à un intervalle de confiance de l'hypothèse nulle d'une distribution complètement aléatoire des firmes de taille prédéfinie dans des régions de taille elle-même connue par la méthode de Monte-Carlo permettrait de détecter une éventuelle concentration ou dispersion significative, mais pas la comparaison entre des industries ou des lieux différents.

Notons pour conclure que dans l'estimation de γ, seule s_i, la part de la région i dans l'industrie, dépend de la localisation des firmes : toutes les autres valeurs sont considérées comme exogènes et étrangères à la question traitée. Estimer γ revient à estimer l'écart quadratique moyen entre s_i et x_i, puis à normaliser le résultat.

L'indice est utilisé très fréquemment dans la littérature économique, mais apparemment jamais dans d'autres domaines. Il a été remis en cause par Kim *et al.* (2000) qui affirment qu'il est parfois biaisé, mais à tort (Marcon, 2003). Des méthodes concurrentes sont toujours en cours de développement en économie, dont un exemple significatif est présenté dans le paragraphe suivant.

Le test multinomial d'agglomération et de dispersion (MTAD) de Rysman et Greenstein

Présentation

La méthode de Rysman et Greenstein (2003) s'applique au même contexte que l'indice d'Ellison et Glaeser, mais utilise une approche différente, consistant à mesurer la différence de vraisemblance entre la distribution observée et une distribution complètement aléatoire.

On considère K firmes, indicées par k, appartenant à C secteurs d'activité, indicés par c, localisées dans M régions indicées par m. Le nombre de firmes dans la région m est n_m, et le nombre de firmes du secteur c dans la région m, x_m^c. Le vecteur des x_m^c dans une région m est noté x_m. La probabilité qu'une firme appartienne au secteur c est notée p_c ; le vecteur des probabilités de tous les secteurs est noté p. La vraisemblance du jeu de données réel est rapportée à celle d'une distribution complètement aléatoire. On note $\binom{n_m}{x_m^1, \dots, x_m^c}$ le nombre de façons de choisir $x_m^1, \dots, x_m^c$ firmes de chaque secteur parmi toutes les firmes, sous la contrainte que leur nombre total soit égal à n_m. La vraisemblance des données observée est calculée pour chaque région :

$$L(x_m, n_m, p) = \binom{n_m}{x_m^1, \dots, x_m^c} \prod_c (p_c)^{x_m^c} \quad \textbf{(149)}$$

La moyenne du logarithme de vraisemblance est calculée pour les données observées :

$$L = \frac{1}{M} \sum_m \ln L(x_m, n_m, p) \quad \textbf{(150)}$$

On considère maintenant la vraisemblance d'un ensemble de firmes se localisant complètement aléatoirement. Le nombre de firmes n_m dans chaque région est supposé aléatoire, suivant une distribution (densité de probabilité discrète pour n fini) notée $f(n_m)$. Alors, l'espérance de la vraisemblance, que nous noterons E, est calculée à partir de toutes les valeurs possibles de n_m, en fonction de leur probabilité de réalisation, et pour chaque valeur de n_m pour toutes les valeurs possibles de x_m :

$$E = \sum_{x_m} \sum_{n_m} \ln L(x_m, n_m, p)\, f(n_m) \quad \textbf{(151)}$$

Le test consiste à calculer la différence des logarithmes de vraisemblance :

$$t = L - E \quad \textbf{(152)}$$

Les auteurs montrent que t a une distribution asymptotiquement normale centrée quand le nombre de région devient grand.

Pratiquement, la valeur du test est calculée à partir des données : les probabilités sont estimées à partir de la proportion de firme de chaque secteur, et la distribution $f(n_m)$ à partir des valeurs observées de n_m. La valeur de référence est 0, et un intervalle de confiance pour l'hypothèse nulle peut être calculé par la méthode de Monte Carlo.

Discussion

Le test MTAD est introduit comme une alternative à l'indice d'Ellison et Glaeser. Un jeu de données fictif présenté par les auteurs montre une grande similarité dans les résultats, certains cas particuliers montrant une plus grande puissance de l'un ou l'autre.

Une différence de taille existe toutefois : le MTAD ne prévoit pas de pondération des firmes. La limite peut facilement être contournée en exécutant le test sur les employés au lieu des firmes, mais alors la concentration industrielle serait totalement négligée. D'autre part, la normalité du test est asymptotique : si le nombre de région est trop petit (et il n'y a aucun moyen d'évaluer un seuil critique), rien ne garantit que l'espérance de la valeur du test pour l'hypothèse nulle soit 0. Enfin, il s'agit d'un test, dont la valeur n'est ni interprétable ni comparable pour deux jeux de données différents.

Les méthodes mesurant l'autocorrélation spatiale

La question de l'autocorrélation spatiale est marginale dans le cadre traité ici. Il ne s'agit pas de décrire comment des objets sont localisés dans un domaine d'étude, mais, leur localisation étant donnée, de savoir si leurs valeurs (par exemple la taille) sont liées. Les objectifs sont donc différents :

- Les analyses de structure spatiale intertypes, comme la fonction K_{12} de Ripley, répondent généralement à la question « les points du type 1 attirent-ils ou repoussent-ils les points du type 2 ? »

- Les analyses d'autocorrélation spatiale répondent à la question « les valeurs des points proches sont-elles semblables ? ». Cette question est proche de la précédente, mais en permet un traitement plus large puisque la limite de deux types de points n'existe plus.

L'approche du problème est finalement l'inverse de celle des analyses de structures spatiales : on cherche à savoir si des points proches sont semblables, au lieu de rechercher si des points semblables ont tendance à se rapprocher. Il n'est pas question de traiter ici la question en profondeur mais seulement de présenter les principes des méthodes utilisées, notamment celles assez répandues comme l'indice I de Moran.

La présentation suivante est largement issue de Upton et Fingleton (1985).

Approche unifiée

De nombreux indices d'autocorrélation spatiale ont été développés, mais leur unification a été réalisée par Hubert *et al.* (1981). On considère un domaine d'étude divisé en un nombre quelconque de zones. On définit une proximité spatiale W_{ij}, définissant la proximité entre toutes les paires de zones i et j, ainsi qu'une « distance » (au sens commun du terme, mais pas nécessairement au sens mathématique) de la variable étudiée (par exemple la taille firmes) notée Y_{ij}. On retient souvent comme valeur de la proximité $W_{ij} = 1$ si les zones sont contiguës, 0 sinon. Si on note x_i, la valeur de la variable étudiée, on choisit souvent $Y_{ij} = \left(x_i - x_j\right)^2$.

Hubert *et al.* ont défini la valeur r :

$$r = \sum_i \sum_j W_{ij} Y_{ij} \qquad \textbf{(153)}$$

Cette statistique résume tous les indices précédents, qui n'en sont que des variantes selon la forme de W_{ij} et Y_{ij}. La forme est clairement analogue à une corrélation entre la distance spatiale et la distance qualitative.

La valeur de r est calculée d'après les données puis comparée à la valeur de l'hypothèse nulle de la répartition aléatoire des zones. La façon la plus simple de traiter l'hypothèse nulle consiste à générer de nombreuses permutations des zones pour obtenir un intervalle de confiance par la méthode de Monte-Carlo. Dans certains cas, l'espérance et la variance de r sont connus sous l'hypothèse nulle, ce qui permet de mettre en place des tests statistiques simples.

Les cartes noires et blanches

Une application très classique prend en compte des zones dont la valeur est binaire (noire ou blanche). La question consiste à savoir si les zones noires sont agrégées. On définit les valeurs suivantes :

- $W_{ij} = 1$ si les zones sont contiguës, 0 sinon
- $x_i = 1$ si le carré est noir, 0 s'il est blanc. $Y_{ij} = (x_i - x_j)^2$. En d'autres termes, Y_{ij} vaut 1 si les deux zones sont de couleurs différentes, 0 sinon. On aurait pu définir une autre distance, pour compter les frontières entre carrés noirs par exemple.

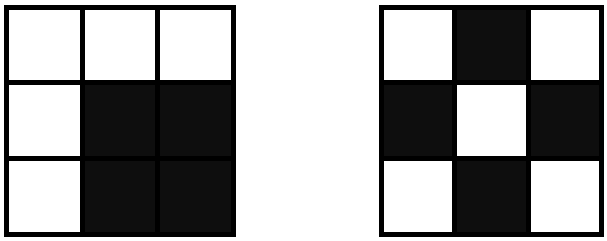


Figure 56 : Cartes de zones noires et blanches

Dans ce cadre, le calcul de r est particulièrement simple : r est le double du nombre de frontières entre deux carrés de couleurs différentes. L'espérance et la variance de r sont connues analytiquement à condition que le nombre de zones de chaque couleur soit suffisamment grand (Upton et Fingleton, 1985 p. 159) pour accepter la normalité de la distribution de r.

La Figure 56 est un exemple. À gauche, on compte 4 frontières entre carrés noirs, la valeur de r est 8. La normalité de la distribution ne peut pas être avancée étant donné le faible nombre de carrés ; la permutation aléatoire des zones a été réalisée par Upton et Fingleton (p. 156). La valeur 8 est la plus faible obtenue (8 fois) sur 99 simulations, ce qui permet de conclure à une agrégation statistiquement significative au seuil de confiance de 10%. La valeur de r pour la figure de droite est 24, atteinte seulement 2 fois et jamais dépassée par les simulations, ce qui permet de conclure à une dispersion à un seuil inférieur à 5%.

Selon la définition de Y_{ij}, qui peut être une distance ou une proximité (on choisit la valeur 1 si les carrés sont de même couleur), les valeurs correspondant à l'agrégation peuvent être les plus grandes ou les plus petites.

L'indice *I* de Moran

L'indice de Moran (1950) est défini de la façon suivante (n est le nombre total de zones) :

$$I = \frac{n}{\sum_i \sum_{j,i\neq j} W_{ij}} \frac{\sum_i \sum_{j,i\neq j} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad \textbf{(154)}$$

(154) : Indice *I* de Moran

Hubert *et al.* (1981) ont montré qu'il s'agissait d'un cas particulier de r, défini par $Y_{ij} = (x_i - \bar{x})(x_j - \bar{x})$, à quelques ajustements formels près. En absence d'autocorrélation, I vaut $-1/(n-1)$. Une autocorrélation très élevée fait tendre I vers 1. Sous réserve de normalité (donc d'une valeur de n suffisante), la variance de I sous l'hypothèse nulle est connue analytiquement, ce qui permet des tests statistiques simples. Sinon, des simulations et la méthode de Monte-Carlo sont nécessaires.

L'indice peut être calculé en utilisant des définitions diverses de la distance W_{ij}. Plusieurs variantes ou indices dérivés peuvent être rencontrés dans la littérature, par exemple dans Feser et Sweeney (2002), les plus fréquents étant :

- L'indice c de Geary (1954) : $c = \frac{n-1}{2\sum_i \sum_{j,i\neq j} W_{ij}} \frac{\sum_i \sum_{j,i\neq j} W_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2}$, très proche de l'indice de Moran. La différence réside dans la définition de $Y_{ij} = (x_i - x_j)^2$. Upton et Fingleton (1985) indiquent que les tests utilisant I sont plus puissants que ceux utilisant c.
- L'indice LISA (local index of spatial autocorrélation, (Anselin, 1995) est la version individuelle de l'indice de Moran : $I_i = (x_i - \bar{x}) \sum_j W_{ij}(x_j - \bar{x})$.

Les corrélogrammes

Les corrélogrammes sont des représentations de l'autocorrélation (calculée par exemple par *I*) en fonction de la distance entre les zones i et j. L'avantage du corrélogramme est de permettre de détecter l'autocorrélation entre des zones qui ne sont pas nécessairement contiguës. Une formulation simple (n'utilisant pas *I*) en est donnée par Goreaud (2000) :

$$\mathrm{COR}(r) = \frac{\mathrm{cov}(x_i, x_j | W_{ij} = r)}{\mathrm{var}(x)} \quad \textbf{(155)}$$

La courbe typique d'un corrélogramme ressemble à une exponentielle négative, montrant une certaine corrélation à faible distance, tendant vers 0 à grande distance.

Les variogrammes sont l'équivalent des corrélogrammes, mais montrent l'écart des valeurs à une distance donnée plutôt que leur similarité :

$$\gamma(r) = \mathbb{E}\left(\left(x_i - x_j\right)^2 | W_{ij} = r\right) = \mathrm{var}(x)\left(1 - \mathrm{COR}(r)\right) \quad \textbf{(156)}$$

On se reportera à Matheron (1970) pour une présentation complète. Ces fonctions sont adaptées à l'étude de variables régionalisées, dans le but de prévoir leur valeur en tous points à partir de quelques points de sondage (krigeage), mais assez peu à celle de semis de points, dont de petits écarts de localisation entraînent de fortes irrégularités. Goreaud (2000) propose une généralisation de la fonction *K* intertype inspirée du corrélogramme, qui permet de lisser les résultats (voir page 179).

Conclusion

L'analyse de l'autocorrélation spatiale de densités de points entre zones contiguës permet de détecter l'agrégation, ce qui les rapproche beaucoup des méthodes d'analyse fondées sur les quadrats, comme la variance relative. Les limites en sont les mêmes, notamment le problème de la définition des zones. Barrios *et al.* (2003) utilisent l'indice de Moran pour détecter un éventuel problème de zonage (MAUP, voir page 111) : les auteurs traitent des données économiques (des établissements manufacturiers) disponibles à un niveau de détail administratif fin (équivalent de la commune), et mesurent la concentration spatiale par l'indice d'Ellison et Glaeser à un niveau plus agrégé (équivalent du département) qu'ils supposent pertinent. L'absence d'autocorrélation spatiale entre les zones voisines assure que le niveau de regroupement choisi est suffisant : s'il est trop détaillé, la concentration mesurée est sous-estimée et les zones voisines sont autocorrélées.

Un usage encore plus intéressant est l'analyse de la corrélation d'une variable quantitative (par exemple la taille moyenne des objets mesurés). On peut y voir une extension des fonctions d'analyse intertype (*K* ou *M*) libérées de la nécessité de regrouper les valeurs en classes permettant l'analyse simultanée de toutes les valeurs (alors que *K* et *M* ne peuvent mesurer les interactions qu'entre 2 classes). On doit cependant garder à l'esprit les hypothèses de départ différentes : l'autocorrélation spatiale suppose que l'emplacement des objets est donné. Enfin, l'autocorrélation ne s'applique pas à des points mais à des zones, supposées implicitement homogènes comme les carrés de la Figure 56. L'application de ces outils à des points ne pose pas de problème de calcul mais l'existence de fortes différences entre des points proches rend les corrélogrammes peu lisibles.

ANNEXE 4 : CODE INFORMATIQUE

Ktest

Code utilisé pour le test d'un jeu de points contre l'hypothèse nulle d'un processus de Poisson homogène (page 44).

Bibliothèque de fonctions

```
# Ktest  : Test d'un semis de points contre un processus de Poisson homogène
# ppData : semis de points à tester, au format ppp de spatstat
# n      : taille du côté du domaine carré
# r      : vecteur des distances de calcul de la fonction K (exemple : c(1,2) pour
calculer K(1) et K(2)
# Retourne un seuil de risque (p-value)
Ktest<-function(ppData, n, r)
{
  # espKi : espérance de K, rho inconnu, calculée en fonction du nombre de points
  espKi_<-espKi(r,n,ppData$n)
  # Variance calculée. Dépend du nombre de points observés
  sigmaKi_<-sigmaKi(r,ppData$n,n)

  # Construction de la matrice de distances entre les points.
  pairdist_<-pairdist.ppp(ppData)  # consomme beaucoup de mémoire.Limité à 8000
points avec R 32bits.

  # Estimation de K
  Kest<-mat.or.vec(1,length(r))
  # Pour chaque distance r
  for (i in 1:length(r))
  {
    # NbPaires : nombre de paires de points à distance <=r (a>0 élimine les
distances d'un point à lui-même)
    # *1.0 pour forcer le passage en réel et éviter les valeurs infinies dans b*n*n
    NbPaires<-sum(pairdist_>0 & pairdist_<r[i])*1.0
    # Kest reçoit l'estimateur de K, centré sur la valeur attendue.
    Kest[i]<-NbPaires*n*n/(ppData$n*(ppData$n-1))-espKi_[i]
  }

  VecteurT<-invsqrtmat(sigmaKi_) %*% t(Kest) * sqrt(sqrt(ppData$n*(ppData$n-1)))
  Ktest<-1-pchisq(sum(VecteurT*VecteurT), length(r))
}


##############################################################################
############################
#-----------------------------------------------------------------------------
# ern : Probabilité qu'un point se trouve au voisinage d'un autre (pi * r² / n²
corrigé des effets de bord)
ern<-function(r,n)
{
  rap<-r/n
  ern<-rap*rap*(pi+rap*(-8/3+rap/2))
```

```
}
#-------------------------------------------------------------------------------
# espKc : espérance de K, densité connue
espKc<-function(r,n)
{
  espKc<-n*n*ern(r,n)
}
#-------------------------------------------------------------------------------
# espKi : espérance de K, densité inconnue. rho n² est estimé par le nombre de
points (lambda).
#        la différence avec espc devient nulle dès 20 points.
espKi<-function(r,n,lambda)
{
  espKi<-espKc(r,n)*(1-(1+lambda)*exp(-lambda))
}
#-------------------------------------------------------------------------------
# eh : espérance de h1²(U,r)
eh<-function(r,n)
{
  rap<-r/n
  eh<-rap^5*(8*pi/3-256/45) +rap^6*(11*pi/48-56/9) +rap^7*8/3 -rap^8/4
}
# ------------------------------------------------------------------------------
# sigmaKi : Matrice de variance de l'estimateur de K (rho inconnu), normalisée en
multipliant par n²rho
# Retourne trois valeurs différentes, suivant l'approximation faite sur covh1
# vec : vecteur des distances de calcul de la fonction K (exemple : c(1,2) pour
calculer K(1) et K(2)
# lambda : nombre de points
# n   : taille du côté du domaine carré
sigmaKi<-function(vec,lambda,n)
{
  # calculs intermédiaires
  d<-length(vec)
  ern_<-ern(vec,n)
  # Estimation de rho à partir du nombre de points
  rho<-sqrt(lambda*(lambda-1))/n^2
  c1<-2*n^6*rho/(lambda*(lambda-1))
  c2<-n^4*exp(-lambda)*(1+lambda)*(1-exp(-lambda)-lambda*exp(-lambda))
  # Préparation d'une matrice carrée
  sigmaKi_<-matrix(nrow = d, ncol = d)
  for (i in 1:d )
  {
    for (j in 1:d )
    {
      covh1_<-covh1(vec[i],vec[j],n)
      sigmaKi_[i,j]<-c1*ern_[min(i,j)]+(c2-
c1)*ern_[i]*ern_[j]+4*n^6*rho/lambda*covh1_
    }
    sigmaKi_[i,i]<-c1*ern_[i]+(c2-c1)*ern_[i]*ern_[i]+4*n^6*rho/lambda*eh(vec[i],n)
  }

  # Retour
  sigmaKi<-sigmaKi_
}
#-------------------------------------------------------------------------------
# invsqrtmat : Transforme une matrice en la racine de son inverse
# telle que invsqrtmat %*% mat %*% t(invsqrtmat) = Id
invsqrtmat<-function(mat)
{
  if (length(mat) > 1)
  {
    e<-eigen(mat)
    # Vecteurs propres
    p<-e$vectors
    # Racines des valeurs propres
    d<-sqrt(e$values)
```

```
    # Construction de la matrice diagonale contenant les racines des valeurs
propres
    rd<-diag(d)
    # Résolution
    invsqrtmat<-solve(p%*%rd)
  }
  else
  {
    invsqrtmat<-1/sqrt(mat)
  }
}
#-------------------------------------------------------------------------------
# covh1 : covariance de h1
# r1, r2 : distances
# n : côté du carré
covh1<-function(r1,r2,n)
{
  # remise en ordre des rayons
  ra1<-min(r1,r2)
  ra2<-max(r1,r2)
  # calculs des rapports des parametres de taille
  rap1<-ra1/n
  rap2<-ra2/n
  rapr<-ra1/ra2
  r12<-rap1*rap1
  r22<-rap2*rap2

  # calculs des biais renormalisés par n^2/r^2
  b1<-brn(ra1,n)
  b2<-brn(ra2,n)
  #taille du carré central A^{1,1}
  cote<-1-2*rap2
  #taille du carré médian A^{1,2}+
  cote1<-1-2*rap1

  # calcul approché de l'intégrale elliptique
  int2<-integrate(integrand3, lower=0, upper=1, r1=ra1, r2=ra2)
  intcorner<-adapt(ndim=2, lower=c(0,0), upper=c(1,1), minpts=100, maxpts=NULL,
eps=0.001, corner, r1=ra1, r2=ra2, n=n)

  #ligne 1
  covh1_<-cote*cote*b1*b2
  #ligne 2
  covh1_<-covh1_+4*b1*cote *rap2*(b2-foncG(1))
  #ligne 3
  covh1_<-covh1_+4*rap1*cote*(int2$value-b2*foncG(1))+4*r22*intcorner$value

  # multiplication par le facteur commun
  covh1_<-covh1_*r12*r22

  # Retour
  covh1<-covh1_
}

foncg<-function(x){if (x<=1) {foncg<-acos(x)-x*sqrt(1-x*x)} else {foncg<-0}}
foncG<-function(x){x*acos(x)-sqrt(1-x*x)*(2+x*x)/3+2/3}

# valeur de h1 sur les différentes zones sans la normalisation en r^2/n^2
brn<-function(r,n){8*r/(3*n)-r*r/(2*n*n)}
indic<-function(a){as.numeric(a)}
foncA1<-function(x,r,n){brn(r,n)*indic(x[1]>=1)*indic(x[2]>=1)}
foncA2<-function(x,r,n){(brn(r,n)-
foncg(x[2]))*indic(x[1]>=1)*indic(x[2]<1)+(brn(r,n)-
foncg(x[1]))*indic(x[2]>=1)*indic(x[1]<1)}
foncA3<-function(x,r,n){(brn(r,n)-foncg(x[1])-
foncg(x[2]))*indic(x[1]<1)*indic(x[2]<1)*indic(x[1]^2+x[2]^2>1)}
```

```
foncA4<-function(x,r,n){(brn(r,n)+x[1]*x[2]-(foncg(x[1])+foncg(x[2]))/2-
pi/4)*indic(x[1]<1)*indic(x[2]<1)*indic(x[1]^2+x[2]^2<=1)}
# valeur du produit dans le coin en coordonnees x'=(n-xi)/r2 sans la normalisation
en r'^2r^2/n^4
corner<-
function(x,r1,r2,n){(foncA1(r2*x/r1,r1,n)+foncA2(r2*x/r1,r1,n)+foncA3(r2*x/r1,r1,n)
+foncA4(r2*x/r1,r1,n))*(foncA3(x,r2,n)+foncA4(x,r2,n))}

integrand3<-function(x,r1,r2){foncg(r1*x/r2)*foncg(x)}
```

Code pour le test

Le semis de points doit être transformé en un objet ppp. Ici, il est simulé. La fonction `Ktest (ppData, n, r, alpha)` est ensuite appelée. Elle retourne la probabilité de rejeter l'hypothèse nulle par erreur.

```
# Environnement
rm(list=ls(all=TRUE))
require ("spatstat")
require ("adapt")
source ("Ktest.r")

# Mettre un objet ppp dans ppData, par simulation ou par lecture d'un fichier csv
puis conversion as.ppp
ppData<-  rpoispp(3, win=owin(c(0,10),c(0,10)))
(pValue<-Ktest(ppData, n=10, r=c(1,2,3,4,5), alpha=5))
```

Les paramètres sont :

- `ppData` : le semis de points à tester, au format `ppp` de Spatstat.
- `n` : le côté du carré.
- `r` : le vecteur des distances auxquelles K doit être calculé. La valeur maximum de r doit être inférieure ou égale à la moitié de n.
- `alpha` : le seuil de risque (5 pour 5%).

Indice β de Shannon

Code utilisé pour le calcul des indices de biodiversité de Shannon et le test de significativité.

```
###################################
# DECOMPOSITION OF SHANNON'S INDEX OF DIVERSITY
# B. Hérault, bruno.herault@ecofog.gf
# Last updated: April 19, 2010
###################################
# Initialization
rm(list=ls(all=TRUE))

###################################
###### DATA IMPORTATION ##########
# Load a .csv file with species in rows and plots in columns
```

```
# The name of the first column must be 'Species'
# Check that data file has no empty cells and no supplementary rows and/or columns
# If so, remove them
###################################
# Choose the name of the file
dataD<-read.csv2(file="Data.csv")
Nspecies<-length(dataD[,1])
Nplots<-length(dataD[1,])-1
Nrecords<-sum(dataD[1:Nspecies,2:(Nplots+1)])
dataM<-as.matrix(dataD[,2:(Nplots+1)])
dimnames(dataM)[[1]]<-dataD$Species
Nrecords.plots<-apply(dataM, 2, sum)
Nrecords.species<-apply(dataM, 1, sum)


###################################
## Diversity Decomposition########
###################################
#####ALPHA DIVERSITY###############
dataAlpha<--log(dataM%*%diag(1/Nrecords.plots))*dataM%*%diag(1/Nrecords.plots)
dataAlpha[dataM==0]<-0
Halpha<-sum(apply(dataAlpha, 2, sum)*Nrecords.plots/Nrecords)
Halpha
#####GAMMA DIVERSITY###############
dataGamma<-Nrecords.species/Nrecords*log(Nrecords.species/Nrecords)
dataGamma[Nrecords.species==0]<-0
Hgamma<--sum(dataGamma)
Hgamma
#####BETA DIVERSITY################
dataBeta<-
dataM%*%diag(1/Nrecords.plots)*log(diag(1/Nrecords.species)%*%dataM%*%diag(1/Nrecor
ds.plots)*Nrecords)
dataBeta[dataM==0]<-0
Hbeta<-sum(apply(dataBeta, 2, sum)*Nrecords.plots/Nrecords)
Hbeta

###################################
## Test of Significance ##########
##    BETA DIVERSITY    ##########
###################################
HbetaSIM<-numeric()
# Choose the number of simulations
SIM<-1:1000
a<-t(matrix(rep(Nrecords.plots,Nspecies),nrow=Nplots,ncol=Nspecies))
b<-matrix(rep(Nrecords.species/Nrecords,Nplots),nrow=Nspecies,ncol=Nplots)
for (i in SIM)
    {
    dataSIM<-matrix(rbinom(Nplots*Nspecies,a,b),nrow=Nspecies, ncol=Nplots)
    NrecordsSIM<-sum(dataSIM)
    Nrecords.plotsSIM<-apply(dataSIM, 2, sum)
    Nrecords.speciesSIM<-apply(dataSIM, 1, sum)
    dataBetaSIM<-
dataSIM%*%diag(1/Nrecords.plotsSIM)*log(diag(1/Nrecords.speciesSIM)%*%dataSIM%*%dia
g(1/Nrecords.plotsSIM)*NrecordsSIM)
    dataBetaSIM[dataSIM==0]<-0
    HbetaSIM[i]<-sum(apply(dataBetaSIM, 2, sum)*Nrecords.plotsSIM/NrecordsSIM)
    }
plot(density(HbetaSIM), col="red", lwd=3, main="Red: Expected ; Green: 95% CI ;
Blue: Observed", xlab="Beta Diversity" )
# Choose the risk level of the test
abline(v=quantile(HbetaSIM, c(0.025,0.975)), col="green", lwd=3)
abline(v=Hbeta, col="blue", lwd=3)
```

ANNEXE 5 : PUBLICATIONS

Cette thèse a donné lieu aux publications suivantes :

1. Marcon, E., Hérault, B., Baraloto, C. et Lang, G. (en préparation). Decomposition and test of Shannon diversity.

2. Marcon, E., Traissac, S. et Lang, G. (en préparation). A Global Test for Ripley's K Function Poisson Null Hypothesis Rejection.

3. Lang, G. and Marcon, E. (2010). Testing randomness of spatial point patterns with the Ripley statistic. *ArXiv e-prints* 1006.1567

4. Marcon, E. and Puech, F. (2010). Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods. *Journal of Economic Geography* **10**(5): 745-762

5. Marcon, E. et Puech, F. (2009). Generalizing Ripley's K function to inhomogeneous populations. *HAL* halshs-00372631.

Les références bibliographiques des deux premiers articles sont incluses dans la bibliographie générale.

THE DECOMPOSITION OF SHANNON'S ENTROPY AND A TEST FOR BETA DIVERSITY

ERIC MARCON[1], BRUNO HÉRAULT[2], CHRISTOPHER BARALOTO[3] AND GABRIEL LANG[4]

[1] AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana. Corresponding author, e-mail: Eric.Marcon@ecofog.gf

[2] Université de s Antilles et de la Guyane, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana.

[3] INRA, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana.

[4] AgroParisTech, UMR 518 Math. Info. Appli., F-75005 Paris, France INRA, UMR 518 Math. Info. Appli., F-75005 Paris, France.

Abstract

We present the explicit formula of Shannon's β diversity and the mathematical framework to justify it. We also show the need for a significance test to avoid misinterpretations and to provide direct statistical tests of H_β. Finally, we show how to decompose Shannon diversity into several nested levels. We provide an example from four tropical forest plots in French Guiana, from which we generate hierarchical Shannon entropies at the plot, forest and regional level that can be tested for significance and readily interpreted using the Hill numbers.

Running head: Decomposition of Shannon diversity

Keywords: Shannon diversity, Kullback-Leibler divergence, Entropy, Biodiversity, Decomposition.

Introduction

Alpha, beta and gamma diversities are among the most employed theoretical concepts in ecology and biodiversity conservation. For most ecologists, alpha diversity traditionally reflects the within-habitat diversity (MacArthur, 1965) while beta diversity is the component of 'total diversity' that is produced by differences in species composition among the sampling units, i.e. "the extent of change of community composition" (Whittaker, 1960). The need to partition diversity within and among habitats has both theoretical (e.g. gradient analyses) and applied (e.g. drawing management plans) consequences such that they are employed in both ecology (e.g. Crist *et al.*, 2003) and conservation biology (e.g. Steinitz *et al.*, 2005). From these studies emerges the idea that a lot of natural and/or anthropogenic factors (among others dispersal limitation, habitat fragmentation and environmental heterogeneity) affect the values of alpha, beta and gamma diversity in a given set of local communities (e.g. Crist *et al.*, 2003).

The partitioning of diversity began with niche ecological studies (Allan, 1975), but recent interest has focused (i) in partitioning biodiversity measures into independent components (Jost, 2007 ; Pélissier et Couteron, 2007 ; Jost *et al.*, 2009) and (ii) in analyzing patterns of diversity sampled from hierarchically scaled studies (Lande, 1996 ; Loreau, 2000). This recent interest largely builds on Lande's (1996) explanations for additive partitions of total diversity (gamma) into components within-samples (alpha) and among-samples (beta), following thus the original concepts of alpha, beta and gamma diversity (Whittaker, 1972). Up to now, diversity partitioning have been used over a wide range of ecosystems, including tropical (e.g. Condit *et al.*, 2002) as well as temperate landscapes (e.g. Qian *et al.*, 2005). Although diversity partitioning has deep conceptual meanings for ecologists, its usefulness to date has been impeded by (i) the lack of theoretical basis for interpreting the results (Jurasinski *et al.*, 2009) and (ii) the lack of statistical methods for testing null hypotheses (Crist *et al.*, 2003).

As a good starting point, it has been acknowledged that most usual measures of diversity are particular cases of generalized entropy measures (Tsallis, 1988). In this framework, the species richness is an entropy-based diversity measure of order 0, the Shannon diversity (Shannon, 1948 ; Shannon et Weaver, 1963) of order 1 and the Simpson (1949) of order 2 (Jost, 2007). Increasing the order of the diversity estimates is conceptually equivalent to enhancing the weight of rare species in the final diversity estimates (Keylock, 2005). The Shannon estimates (order 1) are ecologically meaningful because all species are strictly weighted by their frequency. Furthermore, Jost (2007) shows that Shannon's estimates are the only entropy measures that can be decomposed additively, so that the gamma diversity H_γ is the weighted sum of α diversity of partitions, H_α, and a between-

partition diversity H_β, even when community weights are unequal. And transforming Shannon entropies into Hill numbers makes possible the derivation of the so-called 'true diversity' indices (Jost, 2006 ; Tuomisto, 2010), i.e. the number of equally-likely elements needed to produce a given value of Shannon entropy. And finally, Shannon measures have several intuitively expected properties of a diversity measure (Jost, 2007): (i) alpha and beta components are orthogonal, meaning that a high value of alpha does not force the beta component to be high and vice versa, (ii) gamma is completely determined by alpha and beta, (iii) alpha is never greater than gamma. These unique properties of Shannon measures give them a privileged place as estimates of diversity. But, as far as we know, still lacking is an explicit mathematical formulation of H_β, whose value is always obtained by the difference $H_\gamma - H_\alpha$. Recently, Tuomisto (2010) provided an extremely detailed review of literature defining "beta diversity as a function of alpha and gamma diversity" and, from this review, it emerges that a self-contained definition of H_β, with no reference to H_α and H_γ, has never been proposed.

The aim of this paper is to give the explicit mathematical formulation of a dataset measure of biodiversity (γ) into within- (α) and between-(β) partition diversities following Whittaker's concepts (1972).

The paper is organized as follows. First, we derive the decomposition of Shannon's index of diversity H using Kullback-Leibler divergence, yielding the analytic formulation of H_β. Then, we propose a test of significance for H_β against the null hypothesis that all plots are samples of the same community. We use simulated datasets to show that testing the significance of H_β is necessary for its interpretation. We use a real dataset to show how Shannon's index of diversity H can be decomposed, including hierarchical nested levels of decomposition.

Methods

Derivation of the decomposition

Kullback-Leibler divergence

A Kullback-Leibler (1973) divergence measures how different two distributions of probabilities are. Given a model distribution p, actual frequencies q of a finite number of observations i, the Kullback-Leibler divergence T between p and q is:

$$T = \sum_i q_i \ln \frac{q_i}{p_i} \quad (1)$$

Economists know this measure T as Theil's (1967) dissimilarity index. Note that T is a measure of divergence s.s., neither a dissimilarity nor a distance measure, because p and q do not play a symmetric role. Mathematically, the observed values in a given system are only estimates of the actual frequencies q. Unlike Mori et al. (2005), we abuse notations for simplicity, confusing q and $\hat{q}$, without consequences for our purpose in the following.

Consider an ecological community partitioned into plots. The number of individuals y of each species s in each plot i is denoted y_{si}. The number of individuals in plot i is $y_{+i} = \sum_s y_{si}$, the number of individuals of species s is $y_{s+} = \sum_i y_{si}$. The total number of individuals is y_{++}. The corresponding actual frequencies are $q_{si} = {}^{y_{si}}/_{y_{++}}$, with $\sum_i \sum_s q_{si} = 1$. The expected distribution will be $p_{si} = {}^{y_{+i}}/_{Sy_{++}}$ where S is the number of species. In other words, we expect that all species have the same frequency, and the number of trees is proportional to the size of the plot.

Grouping rule

In this section, we derive a general equality (5) for grouping plots or species.

Data are organized in a table where lines are species, indexed by s and columns are plots indexed by i. Consider any group of cells G: the contribution of the group to the whole entropy is the sum of each cell's entropy. We denote it T_G^α:

$$T_G^\alpha = \sum_{g \in G} q_g \ln \frac{q_g}{p_g} \quad (2)$$

After grouping, a single cell remains. We denote its entropy T_G^γ:

$$T_G^\gamma = q_G \ln \frac{q_G}{p_G} = \left(\sum_{g \in G} q_g \right) \ln \frac{\sum_{g \in G} q_g}{\sum_{g \in G} p_g} \quad (3)$$

Proof: the probability for an individual to belong to the group is the sum of the probabilities that it belongs to any cell of the group.

The between-cell entropy is:

$$T_G^\beta = \sum_{g\in G} \frac{q_g}{\sum_{g\in G} q_g} \ln \frac{\frac{q_g}{\sum_{g\in G} q_g}}{\frac{p_g}{\sum_{g\in G} p_g}} = \left(\sum_{g\in G} q_g\right)^{-1} \left[\sum_{g\in G} q_g \ln \frac{q_g}{p_g} - \left(\sum_{g\in G} q_g\right) \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g}\right] \quad (4)$$

Proof: within the group, the sum of probabilities is 1. Within-group probabilities are normalized consequently.

Finally, the entropy of the group equals its gamma entropy plus its between-cell entropy:

$$T_G^\alpha = T_G^\gamma + \left(\sum_{g\in G} q_g\right) T_G^\beta \quad (5)$$

At this step, alpha, beta and gamma are purely conventional notations. They will be justified later.

Application to Shannon's index

We apply the previous result to Shannon's index of diversity. The expected probability for species s in plot i is $1/S$ (all species are expected to have the same frequency) multiplied by y_{+i}/y, the weight of plot i. The observed frequency is $q_{si} = y_{si}/y_{++}$. We group all the cells of species s. The entropy of the forest for species s is:

$$T_s^\alpha = \sum_i q_{si} \ln \frac{q_{si}}{p_{si}} = \sum_i \frac{y_{si}}{y_{++}} \left(\ln \frac{y_{si}}{y_{+i}} + \ln S\right) \quad (6)$$

The gamma entropy of species s is:

$$T_s^\gamma = q_{s+} \ln \frac{q_{s+}}{p_{s+}} = \frac{y_{s+}}{y_{++}} \left(\ln \frac{y_{s+}}{y_{++}} + \ln S\right) \quad (7)$$

The between-plot entropy of species s is:

$$\left(\sum_i q_{si}\right) T_s^\beta = \left(\sum_i q_{si}\right) \sum_i \frac{q_{si}}{q_{s+}} \ln \frac{\frac{q_{si}}{q_{s+}}}{\frac{p_{si}}{p_{s+}}} = \frac{y_{s+}}{y_{++}} \sum_i \frac{y_{si}}{y_{s+}} \ln \frac{\frac{y_{si}}{y_{s+}}}{\frac{y_{+i}}{y_{++}}} = \sum_i \frac{y_{si}}{y_{++}} \ln \frac{\frac{y_{si}}{y_{s+}}}{\frac{y_{+i}}{y_{++}}} \quad (8)$$

We know (5) that $T_s^\alpha = T_s^\gamma + (\sum_i q_{si}) T_s^\beta$. This equality will be summed over all species to introduce diversity measures:

$$T_\alpha = \sum_s T_s^\alpha = \ln S + \sum_i \frac{y_{+i}}{y_{++}} \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{y_{si}}{y_{+i}} = \ln S - \sum_i \frac{y_{+i}}{y_{++}} H_i^\alpha = \ln S - H_\alpha \quad \textbf{(9)}$$

H_i^α is the alpha diversity of plot i. It is computed according to local frequencies y_{si}/y_{+i}. H_α is the weighted sum of H_i^α. T_α is the Kullback-Leibler divergence between p and q for all plots and all species.

The gamma entropy sums to give the Kullback-Leibler divergence for the forest:

$$T_\gamma = \sum_s T_s^\gamma = \ln S + \sum_s \frac{y_{s+}}{y_{++}} \ln \frac{y_{s+}}{y_{++}} = \ln S - H_\gamma \quad \textbf{(10)}$$

Finally, we sum between-plot entropy:

$$T_\beta = \sum_s \left(\sum_i q_{si} \right) T_s^\beta = \sum_i \sum_s \frac{y_{si}}{y_{++}} \ln \frac{\frac{y_{si}}{y_{s+}}}{\frac{y_{+i}}{y_{++}}} = \sum_i \frac{y_{+i}}{y_{++}} \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{\frac{y_{si}}{y_{+i}}}{\frac{y_{s+}}{y_{++}}} \quad \textbf{(11)}$$

Combining equations (9), (10) and (11) and assuming $H_\gamma = H_\alpha + H_\beta$, we identify β diversity:

$$H_\beta = \sum_i \frac{y_{+i}}{y_{++}} H_i^\beta = \sum_i \frac{y_{+i}}{y_{++}} \sum_s \frac{y_{si}}{y_{+i}} \ln \frac{\frac{y_{si}}{y_{+i}}}{\frac{y_{s+}}{y_{++}}} \quad \textbf{(12)}$$

H_β is the weighted sum of contributions of plots i, H_i^β. These contributions are Kullback-Leibler divergences. The expected probabilities are $p_{si} = y_{s+}/y_{++}$. The probability to find an individual of species s in plot i is proportional to the frequency of the species in the forest. All plots are expected to be identical. Observed frequencies are $q_{si} = y_{si}/y_{+i}$: actual frequencies differ from plot to plot. In agreement with intuition, β diversity is the divergence between identical plots and real plots.

Hill (1973) numbers are the numbers of equiprobable species yielding the same measure of diversity as the actual data, also called the effective number of species. They allow one to transform non-intuitive values of Shannon diversity into easy-to-understand numbers. The Hill number for β diversity is the number of equally-weighted, completely distinct plots giving the same value of H_β, that is to say the effective number of plots.

Test of significance

We may want to test H_β against this null hypothesis: plots are samples of the same forest, i.e. the same community. This does not mean that species composition does not differ between plots but that, whatever the plot and the species, the observed values q_{si} are realizations of p_{si}, meaning that H_β is not null due to stochasticity.

The test can be done by Monte-Carlo simulations as follows:

- Draw each value of y_{si} in a binomial law $\mathcal{B}(y_{++}, {}^{y_{s+}}/_{y_{++}})$ and compute H_β
- Repeat the simulation a great number of times, say 10,000, and eliminate extreme values according to the chosen risk level α. For $\alpha = 5\%$, the confidence interval of the null hypothesis is between the 251st and the 9750th simulated values of H_β.

Rejection of the null hypothesis will occur if species frequencies are different. It may happen that H_β be smaller than expected if frequencies vary less than a binomial law, in an artificial ecosystem for example.

The R (R Development Core Team, 2010) code we wrote and used in this paper can be found as a supplementary material for use in further studies.

Examples

We first use a simulated dataset to illustrate that the numeric value of H_β depends on both the number of species in the community and the sampling effort. We simulate two frequency distributions of respectively 20 and 40 tree species in forest plots. Frequencies follow the same uniform law. Then we draw a pair of plots 10,000 times from these communities, with an expectancy of 500 trees or 5,000 trees. H_β is computed for each pair of plots and results are reported in a frequency histogram of H_β, smoothed as a density function by the `density` function of R.

We also provide a real example to show how the test can be applied to actual data. We measure Shannon diversity in four 1-ha plots of tropical rain forest at the Nouragues and Paracou field stations in French Guiana. Both sites are seasonal lowland forests receiving about 3m of annual precipitation, with tree composition dominated by Fabaceae, Chrysobalanaceae, Lecythidaceae, Sapotaceae, and Burseraceae (Bongers *et al.*, 2001 ; Gourlet-Fleury *et al.*, 2005). The two plots within each site were chosen to represent the most common contrasting environments found for hilltop terra firme forest at each site. At Paracou, the two example plots occur on migmatites associated with the Bonidoro formation, with one plot exhib-

iting blocked vertical drainage (P06) and the other with strong vertical drainage and incipient podzolization (P18). At Nouragues, the two example plots occur on weathered granite with sandy soils (NH20) and metavolcanic rock of the Paramaca formation with clay-rich laterite soils. In all four plots all trees were sampled in 2008 by professional climbers to obtain herbarium vouchers, each of which was identified to distinct morphospecies at the Cayenne regional herbarium (Baraloto *et al.*, 2010). We assume for simplicity here that an acceptable sample of each forest is obtained when its two plots are united.

Results

The simulations exemplified how the significance of a given H_β value depends on the number of individuals sampled. H_β does not change if all numbers of individuals are multiplied by 10, while maintaining actual frequencies the same. But the null hypothesis of the test is that plots are a random draw of the same community: when more individuals are drawn, species frequencies from the binomial law converge to their probability due to the law of large numbers so H_β converges to 0. In a 20-species community, an observed value $H_\beta = 0.005$ shows a significant difference between plots if it is obtained from two 5,000-tree plots (Figure 1, solid curve on the left). If the plots contain only 500 trees (Figure 1, solid curve on the right), the same value is no longer significant. H_β also tends to be higher when the number of species increases.

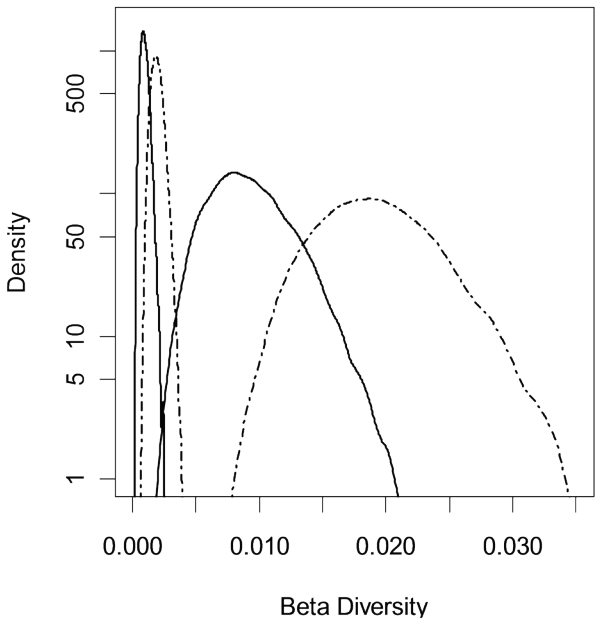


Figure 1 - Probability densities of H_β obtained from 10,000 simulations of the model described in details in the text. Two forest plots are drawn from the same community. H_β is not zero because of stochastic differences between the plots. The first two curves on the left concern plots around 5,000 trees, the right ones plots around 500 trees. Dotted lines are for 40-species plots, solid lines for 20 species. Everything else equal, expected H_βs decrease with the number of trees and increase with the number of species.

Plots	*NH20*	*NL11*	*P006*	*P018*
Number of trees	558	515	643	481
Number of species	203	182	147	149
H_{plots}	4.74	4.63	4.19	4.42
Hill Number	114	103	66	83
Weighted H_{plots}	2.46	2.22	2.40	1.89
H_{β}^{plots}	0.42 [0.11; 0.18]		0.45 [0.08; 0.14]	
Hill Number plots	1.52 [1.12; 1.20]		1.57 [1.08; 1.15]	
H_{forest}	5.11		4.74	
Weighted H_{forest}	2.49		2.42	
$H_{\beta}^{forests}$	0.38 [0.08; 0.13]			
Hill Number forests	1.46 [1.08; 1.14]			
H_{total}	5.29			
Hill Number total	199			

Table 1 – An example of hierarchical decomposition of Shannon diversity (H) for tropical tree communities. Trees with Diameter at breast Height > 10cm were inventoried in four 1ha tropical rain forest plots in French Guiana. The first two plots (NH20, NL11) are from the Nouragues forest station, the last two from the Paracou (P006, P018) forest station. Within forests, the sum of weighted alpha (Weighted H_{plots}) and beta diversities (H_{β}^{plots}) equals the within forest gamma diversity (H_{forest}). This within forest gamma diversity can be considered as the alpha diversity at the between-forest level. In this way, the Weighted H_{forest} added to the beta diversity between forests ($H_{\beta}^{forests}$) gives the total diversity. Hill numbers are the numbers of equiprobable species or completely different plots or forests yielding the same measure of diversity as the actual data. Beta diversities are given with the 99.99% confidence interval of the null hypothesis of a single community, between square brackets.

The second example illustrates how Shannon diversity can be hierarchically partitioned (Table 1). The first result is that plots at the Nouragues site are more diverse than those at Paracou. Hill numbers offer an intuitive representation of the level of diversity. For example, the Nouragues NH20 plot is as diverse as one of the same size with 114 equally frequent species, almost twice the value ob-

tained at Paracou P006. Next, plots can be grouped into forests. The value of H_{forest}, that is to say the γ diversity of the forest, is the sum of the weighted H_α of plots plus the H_β between plots. In turn, H_{forest} can be treated as an α diversity and one can follow the same procedure. Successive values of H are given in Table 1. H_β can be tested against the null hypothesis of identical probabilities in all plots. For example, for the Nouragues plots, the distribution of H_β under the null hypothesis averaged 0.144, corresponding to a Hill number equal to 1.16 plots. Note that theoretical Hill values for this example of two plots are between 1 (perfect equality of distribution, $H_\beta = 0$) and 2 (equal number of trees with no species in common, $H_\beta = \ln 2 \approx 0.7$). H_β values below 0.1 (N=1.10) or over 0.2 (N=1.22) have a probability so close to zero that they can be considered as impossible results of a random draw of two plots following the probability distribution of species in the whole forest. The observed H_β for the Nouragues plots is 0.42 (Table 1), far over the upper confidence limit. All values of H_β in Table 1 are highly significant (over 99.99%). We can see that diversity within forests is roughly the same as that between forests (all values of Hill Numbers are around 1.5). We could have chosen to group the plots directly. In this case, H_β between all plots would be 0.81 (significance over 99.99%). The corresponding Hill number would be 2.25 meaning that the four plots are as different as 2.25 completely different ones.

Discussion

In this paper, we propose a self-contained definition of H_β, with no reference to H_α and H_γ. Bourguignon (1979) showed that Theil's entropy was the only inequality measure that could be decomposed the way we do it here. The form of H_β we provide has already been derived by Ricotta and Avena (2003), but they did not relate it with H_α and H_γ. Also, Ludovisi et Taticchi (2006) decomposed a Kullback-Leibler divergence with a different approach, in order to develop new measures of β diversity.

Kullback-Leibler divergences provide the necessary framework to decompose Shannon diversity. Shannon's α diversity is the difference between the logarithm of the number of species and Theil's relative entropy, that is to say the Kullback-Leibler divergence between a distribution where all species have the same frequency and actual data. Shannon's γ diversity has the same definition after grouping plots. Shannon's β diversity is the Kullback-Leibler divergence between actual plots and identical ones. Jurasinski et al. (2009) distinguish two general definitions for β diversity: proportional diversity as the variation of inventory di-

versity across scales and differentiation diversity as the compositional dissimilarity between scales. As the difference between γ and α indices, Shannon's β diversity measures proportional diversity. We show that it also addresses differentiation diversity. The particular properties of Shannon index allow a reconciliation of these two approaches.

In the proposed diversity partitioning framework, H_β is very different from H_α and H_γ because it is a measure of divergence, not of diversity s.s. The more similar the species relative abundances, the higher H_α and H_γ are, but H_β increases when plots are less similar. This is in agreement with the original definition of diversity expressed by Whittaker (1960). Converting Shannon's entropy to Hill numbers allows a unified definition of diversity as a number of effective objects (species or plots), or "true diversity" measures (see Tuomisto, 2010, page 8, for a thorough discussion).

The maximum theoretical value of H_β is lnS when all plots have an α diversity equal to zero (i.e. they contain a single species different among plots); and γ diversity also has its maximum value, equal to lnS. This is possible only if the number of samples equals the number of species and the number of individuals in every sample is the same. A more realistic case is H_β equal to the logarithm of the number of samples. In this case, samples contain the same number of individuals (equal weight) but have no species in common. The minimum value is 0 when all samples are completely identical in species relative abundances.

Note that $T_\alpha = T_\beta + T_\gamma$ because the divergence is greater for disaggregated data. This equality is similar to decomposition of variance: the total variance equals the sum of within-group and between-group variances. In this way, economists decompose Theil's indices into within- and between-group components (see Brülhart et Traeger, 2005 for instance). Because $T_\alpha = \ln S - H_\alpha$ and $T_\gamma = \ln S - H_\gamma$, this leads to $H_\gamma = H_\alpha + H_\beta$.

Conclusion

In this paper, we provided the explicit formula of Shannon's β diversity and the mathematical framework to justify it. We also showed the need for a significance test to avoid misinterpretations and to provide direct statistical tests of H_β. Finally, we showed how to decompose Shannon diversity into several nested levels. This diversity partitioning is flexible enough to analyze any a priori determinant of species diversity. We believe that the use of explicit diversity partitioning may help ecologists to a better understanding of the factors shaping the spatial and

temporal distribution of biodiversity and, in fine, help nature practitioners to design effective strategies for protecting it (Veech *et al.*, 2002).

A GLOBAL TEST FOR RIPLEY'S K FUNCTION POISSON NULL HYPOTHESIS REJECTION

ERIC MARCON[1*], STÉPHANE TRAISSAC[1,2] AND GABRIEL LANG[3,4]

[1] AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana;

[2] E-mail Stephane.Traissac@ecofog.gf;

[3] AgroParisTech, UMR 518 Math. Info. Appli., F-75005 Paris, France;
INRA, UMR 518 Math. Info. Appli., F-75005 Paris, France;

[4] E-mail Gabriel.Lang@agroparistech.fr;

*Corresponding author; fax: +594594324302; E-mail: Eric.Marcon@ecofog.gf

Abstract

Question: How to test Ripley's K function against complete spatial randomness rigorously and faster than classical Monte-Carlo simulations?

Methods: We follow Lang and Marcon (2010) to provide a statistical test. We test it on simulated data and show how to use it on real data.

Results: The test returns the p-value to reject the null hypothesis of complete spatial randomness erroneously. We show that the test works as expected.

Conclusion: Our test should be used to characterize homogenous point patterns with the K function. It complements the usual graphical presentation of K by a global, rigorous statistic of test.

Running head: K function global test

Keywords: Ripley, K function, statistical test, point process, point pattern.

Introduction

The commonest tool used to characterize the spatial structure of a point set is Ripley's K statistic (Ripley, 1976; 1977). It has been widely used in ecology and other scientific fields and is well referenced in handbooks (Ripley, 1981 ; Diggle, 1983 ; Stoyan *et al.*, 1987 ; Cressie, 1993 ; Illian *et al.*, 2008). Classically, an observed set of points is tested against a homogeneous Poisson point process taken as a null model. Since little is known about the distribution of the K function, the test of rejection of the null hypothesis relies on Monte Carlo simulations. Large point patterns are difficult to deal with because computation time is roughly proportional to the square of the number of points (to calculate the distances between all pairs of points) multiplied by the number of simulations. A theoretical problem is less frequently mentioned (Duranton et Overman, 2005 is an exception): the test is valid for one value of r, and no rigorous global test is available (although Duranton and Overman provide a heuristic global test).

Lang and Marcon (2010) developed a global, asymptotical test able to return a classical p-value, that is to say the probability to erroneously reject the null model. They showed that the test can be used with a very little number of points. Their test relies on exact values of the bias and variances of the statistics. For simplicity, it was computed only on square domains. We extend it in this paper so that it can be used in a rectangular window, as most applications require.

We first provide the mathematical framework supporting the test. Then we show how to use it and we verify that the test actually works. Last, we apply it to a dataset and discuss the results. We compare the results of the test to the classical Monte Carlo technique and we give its graphical interpretation to help the reader understand the mathematics.

Materials and Methods

Mathematical framework

We consider a point pattern in a rectangular window denoted A_{l_1,l_2}. l_1 and l_2 are the sides of the window (width and length). ρ is the intensity of the point process the point pattern is a realization of, estimated by the number of observed points N divided by the area of the window. r and r' are two distances the function is estimated at, r' is larger than r. Only when necessary, distances are denoted $r_1, \ldots, r_i, \ldots, r_d$ but r and r' are preferred for readability of equations. Points are

denoted ξ_i and $I\{d(\xi_i,\xi_j) \leq r\}$ is an indicator function equal to 1 when the distance between two point is less or equal than r, 0 else. Details of the calculation are in supplementary materials: we follow exactly Lang and Marcon (2010).

Ripley's K function is estimated from the data for each distance r, without correction for edge effects as in Lang and Marcon (2010):

$$\widehat{K}(r) = \frac{l_1 l_2}{N(N-1)} \sum_{\xi_i \neq \xi_j} I\{d(\xi_i,\xi_j) \leq r\} \quad \textbf{(1)}$$

Assumptions are those of Ripley's K function: we test the independence of locations of an observed point pattern, assumed to be a realization of a homogenous point process. Homogeneity means both stationarity (the process in unchanged by translation) and isotropy (the process is unchanged by rotation). Thus the null hypothesis of complete spatial randomness (CSR) is that the point process is a homogenous Poisson process.

Edge effects introduce a bias, computed for the null model:

$$B(r) = -\frac{4r^3(l_1+l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2} \quad \textbf{(2)}$$

Estimated $\widehat{K}(r)$ can be corrected for the bias of the null model to test them against it. We get a vector of results of length d:

$$\boldsymbol{K} = \left(\widehat{K}(r_1) - B(r_1), \widehat{K}(r_2) - B(r_2), \ldots, \widehat{K}(r_d) - B(r_d)\right) \quad \textbf{(3)}$$

When the spatial structure of trees is investigated in a forest plot, r values may be, say, each meter from 1 to 50. The classical, local test consists in comparing each observed $\widehat{K}(r)$ to the confidence interval of $\widehat{K}(r)$ obtained by Monte-Carlo simulations of the null model (a homogenous Poisson point process). Rejection of the null model at the chosen significance level α, say 5%, is retained when the observed $\widehat{K}(r)$ is out of the corresponding confidence interval.

In our example, we have 50 values of $\widehat{K}(r)$. If we draw a point pattern in a Poisson process we can expect 5% of them, *i.e.* 2 or 3 of them, to be out of the confidence interval. To address the global question "can we reject the null hypothesis that the observed point pattern is a realization of a Poisson process?", a Bonferroni correction should be applied. As a consequence, the local significance level of the test should be decreased dramatically to have a global significance level of 5%. Actually, $\widehat{K}(r)$ are highly correlated because K is a cumulative function:

roughly speaking, most of $\widehat{K}(r)$ value comes from that of the previous one. This reduces the need for a correction but does not eliminate it completely (Marcon et Puech, 2003). Since no quantification of the correction is available, the local test is used, keeping in mind that the global significance level of the test is somehow higher than announced.

To address this issue, solutions have been proposed (see Lang et Marcon, 2010 for a review). Ward and Ferrandino (1999) opened the way for a global test but failed to compute it correctly (Lang et Marcon, 2010). Duranton and Overman (2005) proposed a heuristic test consisting in eliminating simulated $\boldsymbol{K}$ vectors globally when one of their values is an extreme one. A rigorous test is still missing.

Lang and Marcon (2010) showed that for a homogenous point process the vector $\boldsymbol{K}$ is asymptotically normal, with explicit variance matrix Σ.

$$\Sigma = \begin{pmatrix} \mathrm{Var}\left(\widehat{K}(r_1)\right) & \cdots & \mathrm{cov}\left(\widehat{K}(r_1), \widehat{K}(r_d)\right) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}\left(\widehat{K}(r_1), \widehat{K}(r_d)\right) & \cdots & \mathrm{Var}\left(\widehat{K}(r_d)\right) \end{pmatrix} \quad \textbf{(4)}$$

Consequently $\left\|\Sigma^{-\frac{1}{2}}\boldsymbol{K}\right\|$, follows a χ^2 law. Asymptotic value of the variance is reached with dozens of thousands points, so it is of little use, but normality is acceptable with very few points so the test can be used with real data, as the exact value of Σ can be calculated. We give here the value of Σ in a rectangular window.

Variances are:

$$\begin{aligned} \mathrm{Var}\left(\widehat{K}(r)\right) = {} & 2l_1^2 l_2^2 \mathbb{E}\left(\frac{I(N>1)}{N(N-1)}\right)\left(e_{r,l_1,l_2} - e_{r,l_1,l_2}^2\right) \\ & + 4l_1^2 l_2^2 \mathbb{E}\left(\frac{I(N>1)(N-2)}{N(N-1)}\right) V(r, l_1, l_2) \\ & + l_1^2 l_2^2 e^{-\rho l_1 l_2}(1+\rho l_1 l_2)(1-e^{-\rho l_1 l_2} - \rho l_1 l_2 e^{-\rho l_1 l_2}) e_{r,l_1,l_2}^2 \end{aligned} \quad \textbf{(5)}$$

Where:

$$e_{r,l_1,l_2} = \frac{\pi r^2}{l_1 l_2} - \frac{4r^3(l_1+l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2} \quad \textbf{(6)}$$

And:

$$V(r,n_1,n_2)=\frac{r^5(l_1+l_2)}{l_1^3l_2^3}\left(\frac{8}{3}\pi-\frac{256}{45}\right)+\frac{r^6}{l_1^3l_2^3}\left(\frac{11}{48}\pi-\frac{8}{9}-\frac{16(l_1+l_2)^2}{l_1l_2}\right)+\frac{4r^7(l_1+l_2)}{3l_1^4l_2^4}-\frac{r^8}{4l_1^4l_2^4} \quad (7)$$

l_1,l_2e_{r,l_1,l_2} is the expectation of $K(r)$. The main term is πr^2 and the other terms correspond to the bias due to edge effects.

ρ in equation (5) is unknown, so it is estimated by $N/(l_1l_2)$.

Covariances are:

$$\begin{aligned}\operatorname{cov}\left(\widehat{K}(r),\widehat{K}(r')\right)&=2l_1^2l_2^2\mathbb{E}\left(\frac{I(N>1)}{N(N-1)}\right)\left(e_{r,l_1,l_2}-e_{r,l_1,l_2}e_{r\prime,l_1,l_2}\right)\\&+4l_1^2l_2^2\mathbb{E}\left(\frac{I(N>1)(N-2)}{N(N-1)}\right)\mathrm{C}(r,r',l_1,l_2)\\&+l_1^2l_2^2e^{-\rho l_1l_2}(1+\rho l_1l_2)(1-e^{-\rho l_1l_2}-\rho l_1l_2e^{-\rho l_1l_2})e_{r,l_1,l_2}e_{r\prime,l_1,l_2}\end{aligned} \quad (8)$$

Where:

$$\begin{aligned}\mathrm{C}(r,r',l_1,l_2)&=(l_1-2r')(l_2-2r')\frac{r^2{r'}^2}{l_1^3l_2^3}b_{r,l_1,l_2}b_{r',l_1,l_2}\\&+2(l_1+l_2-4r')\frac{r^2{r'}^3}{l_1^3l_2^3}b_{r,l_1,l_2}\int_{r/r'}^1\left(b_{r',l_1,l_2}-g(x'_1)\right)dx'_1\\&+2(l_1+l_2\\&-4r')\frac{r^3{r'}^2}{l_1^3l_2^3}b_{r,l_1,l_2}\int_0^1\left(b_{r',l_1,l_2}-g\left(\frac{rx_1}{r'}\right)\right)\left(b_{r,l_1,l_2}-g(x_1)\right)dx_1\\&+4\frac{r^2{r'}^4}{l_1^3l_2^3}\int_0^1\int_0^1\left[h_{A1}\left(\frac{r'x'}{r},r\right)+h_{A2}\left(\frac{r'x'}{r},r\right)+h_{A3}\left(\frac{r'x'}{r},r\right)\right.\\&\left.+h_{A4}\left(\frac{r'x'}{r},r\right)\left(h_{A3}(x',r')+h_{A4}(x',r)\right)\right]dx'_1dx'_2\end{aligned} \quad (9)$$

And:

$$b_{r,l_1,l_2}=\pi-\frac{l_1l_2}{r^2}e_{r,l_1,l_2}=-\frac{4r(l_1+l_2)}{3l_1l_2}+\frac{r^2}{2l_1l_2} \quad (10)$$

And:

$$g(x) = I(x < 1)\left(\arccos x + x\sqrt{1 - x^2}\right) \quad \textbf{(11)}$$

And:

$$\begin{aligned}
h_{A1}(x,r) &= b_{r,l_1,l_2} I(x_1 \geq 1) I(x_2 \geq 1) \\
h_{A2}(x,r) &= \left(b_{r,l_1,l_2} - g(x_2)\right) I(x_1 \geq 1) I(x_2 < 1) \\
&\quad + \left(b_{r,l_1,l_2} - g(x_1)\right) I(x_2 \geq 1) I(x_1 < 1) \\
h_{A3}(x,r) &= \left(b_{r,l_1,l_2} - g(x_1) - g(x_2)\right) I(x_1 < 1) I(x_2 < 1) I(x_1^2 + x_2^2 \geq 1) \\
h_{A4}(x,r) &= \left(b_{r,l_1,l_2} - \frac{\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2}\right) I(x_1^2 + x_2^2 \leq 1)
\end{aligned} \quad \textbf{(12)}$$

$\mathbb{E}\left(\frac{I(N>1)}{N(N-1)}\right)$ and $\mathbb{E}\left(\frac{I(N>1)(N-2)}{N(N-1)}\right)$ are estimated by $\frac{1}{N(N-1)}$ and $\frac{(N-2)}{N(N-1)}$ as N follows a Poisson law (Lang et Marcon, 2010).

Application

The test is applied as follows:

1. Compute $\boldsymbol{K}$ from the observed point pattern, following equation (3).
2. Compute Σ according to the window's size and the number of observed points.
3. Finally compare $\left\|\Sigma^{-\frac{1}{2}}\boldsymbol{K}\right\|$ to a χ^2 distribution with d degrees of freedom and return the p-value.

We provide the code to run the code with R (R Development Core Team, 2010) as a supplementary material.

The maximum distance $\widehat{K}(r)$ is computed at must be less or equal than half the width of the rectangle. This is a classical geometrical limitation, already faced by local edge-effect corrections (Goreaud et Pélissier, 1999).

The test provides a p-value (the risk of error if H0 is rejected). It is more rigorous than the usual Monte-Carlo test and much faster: $\boldsymbol{K}$ is computed only once, with no local, point-by-point edge-effect correction. It is very efficient for large data sets.

Examples

We consider the distribution of two tree species in Paracou field station, French Guiana (Gourlet-Fleury *et al.*, 2004). All trees over 10 cm DBH have been plotted. We use data from a 400.6 by 522.3 meters rectangle included in the four plots of the southern block of the experimental device. A map of trees is in figure 1.

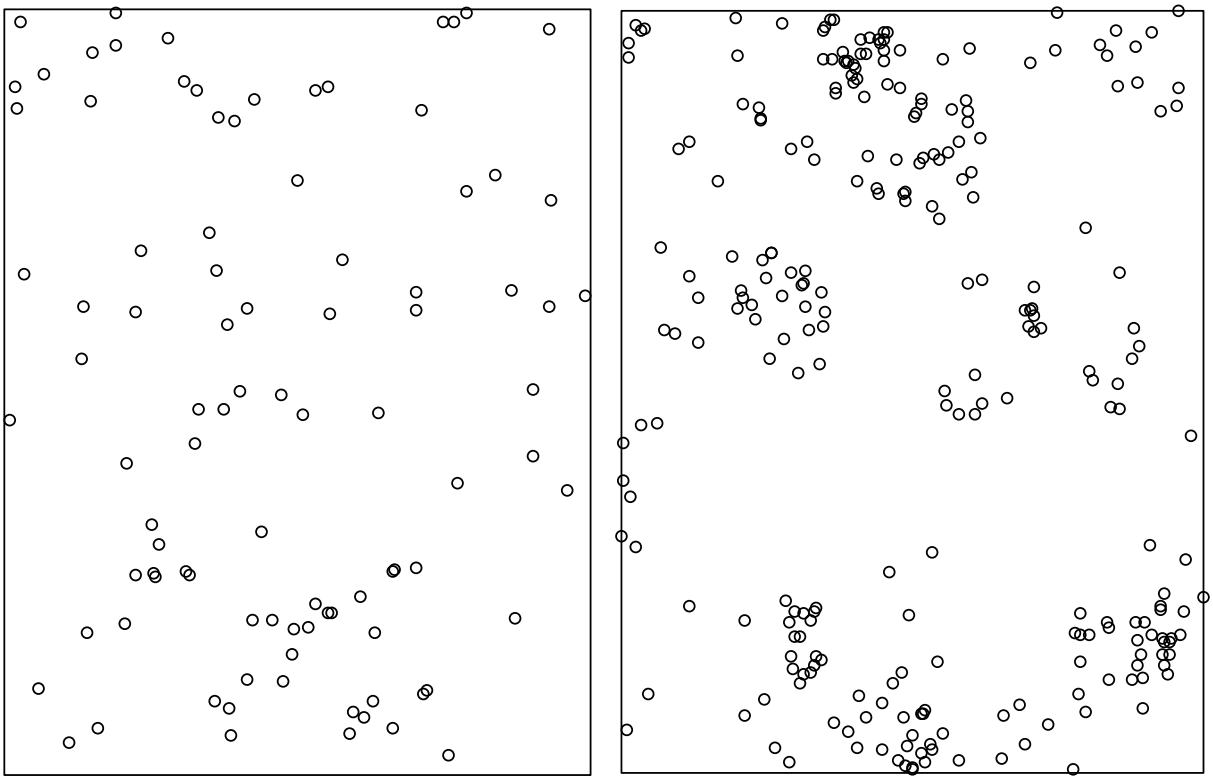


Figure 1: Map of *Tachigali melinonii* (94 trees, left) and *Dicorynia guianensis* (254 trees, right).

Dicorynia guianensis is a widely studied species in French Guiana, its spatial structure has been characterized for a long time (Goreaud *et al.*, 1997): as a visual inspection of the map allows to guess, *Dicorynia guianensis* is an aggregated species.

Much less is known about *Tachigali melinonii*. The species has been studied for its special biomechanical behavior (Jaouen *et al.*) or leaf trait plasticity (Coste *et al.*, 2009). The spatial structure of its saplings has been reported by Flores *et al.(2006)* but the structure of adult trees is not clear.

We apply our test to these two point sets. We also provide a classical plot of $L(r) = \sqrt{\frac{K(r)}{\pi}} - r$ (Besag, 1977) against r, computed every 5 meters up to 250m. 1,000 simulations of a binomial process with the same number of points as the real data are run. At each distance r, the 25 lower and greater values are eliminated to build the local 5% confidence interval. The global confidence interval is

built iteratively (Duranton et Overman, 2005 ; Marcon et Puech, 2010): simulations corresponding to extreme values (maximum or minimum) at any distance are eliminated. This process is repeated until 5% of the simulations are concerned. The extreme remaining values are plotted. Interpolation is used if the last iteration eliminates more simulations than required.

Results

Test of the test

The test can be tested by simulation. We simulate 10,000 realizations of a Poisson point process and test them. At the 5% significance level, we expect 500 of them to be rejected. As rejection of a pattern follows a Bernoulli law with probability 5% and tests are independent, The number of rejected patterns follows a binomial law $\mathcal{B}(10{,}000, 5\%)$. It can be approximated by a normal law $\mathcal{N}\left(500, \sqrt{475}\right)$ so we have a 95% probability that the observed number of rejections will be between 457 and 543.

The test has been run 10,000 times in a 10x15 rectangle window. A Poisson process was drawn. $\boldsymbol{K}$ was calculated at distances 1, 2, 3, 4 and 5. The hypothesis of a Poisson process was rejected around 500 times for a wide range of intensities (Table 1). The number of rejections is a little too high when points are less numerous, but always under 6%. This little bias is acceptable for practical purposes: it appears that the test can be applied for any actual point set.

Table 1: Number of rejections of the null hypothesis (the point process is Poisson) out of 10,000 simulations of a homogenous point process in a rectangular 10 by 15 window. The significance level is 5%, so 500 simulations are expected to be rejected. The intensity varies from 0.3 to 64 so that the expected number of points varies from 45 to 9600, covering the range of usually-studied point patterns.

Expected Number of points	**5**	**0**	**80**	**00**	**50**	**00**	**200**	**400**	**800**	**600**
Number of rejections	75	36	58	09	21	11	10	74	06	00

Structure of *Dicorynia guianensis*

Aggregation is obvious on Figure 2. Our test applied with the vector of distances (10,20,30,…,150) meters returns a p-value equal to zero, that is that the quantile of the χ^2 distribution with 15 degrees of freedom for $\left\|\Sigma^{-\frac{1}{2}}\boldsymbol{K}\right\|$ is so low that R returns 0.

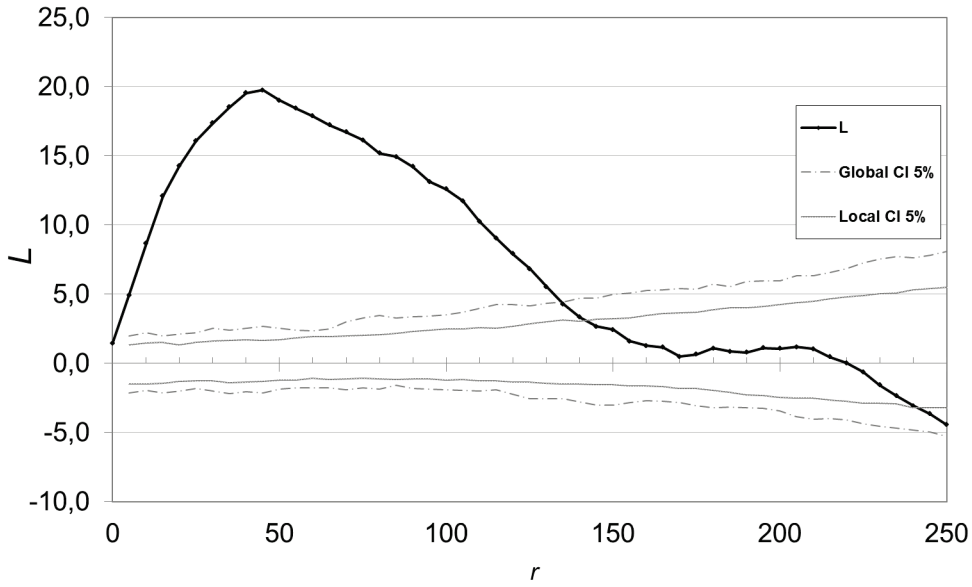


Figure 2: *L* values for *Dicorynia guianensis*. Distances are in meters. Confidence intervals are computed for the null hypothesis of complete spatial randomness at the 5% significance level. The local and global confidence intervals are calculated by Monte-Carlo simulations as explained in the text.

Structure of *Tachigali melinonii*

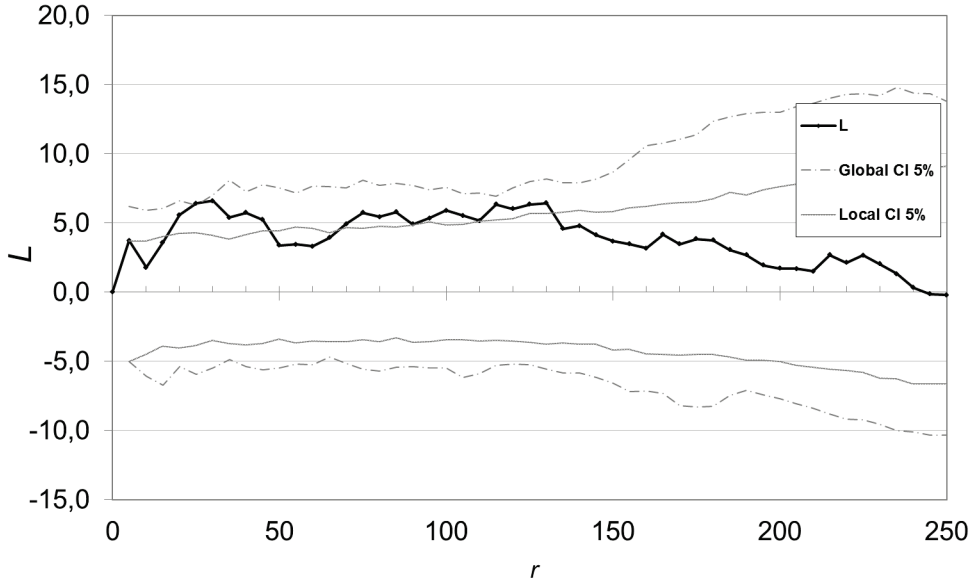


Figure 3: L values for *Tachigali melinonii*. See Figure 2 for details.

Figure 3 shows a less clear structure than that of *Dicorynia guianensis*. The curve leaves the local confidence interval many times, but not the global one. Our test applied with the same distance vector (10 to 150m) returns a p-value equal to 2.5%: aggregation is significant.

Discussion

Application of the test

The test does not give information on what values of r are responsible for its significance. Practically, in order not to lose usual references, a graphical representation of $\boldsymbol{K}$ or, better, its transformed function L, should be provided with local confidence intervals. If the number of points is great, the number of Monte-Carlo simulations can be reduced since these intervals are not used for the test.

Local confidence interval values are correct but testing a curve made of 50 points against local confidence intervals is not. If $L(r)$ were independent from each other, we would expect 2 or 3 of them to leave the local confidence interval under the null hypothesis at the 5% significance level: the Bonferroni correction should be applied. Actually, values are highly correlated (K and L are cumulative functions) so this issue is minimized, but it cannot be quantified. Global confidence intervals plotted in the figures are heuristic: they do not rely on any mathematical proof. They appear to be too conservative for *Tachigali melinonii*.

Our test allows to reject the null hypothesis more rigorously. Yet it relies on the arbitrary choice of r values (the vector of distance). K values are classically computed at many distances (50 steps of 5 meters in our examples). Because we have to invert a matrix of variances whose dimension is that of the vector of distance, too many values will cause numerical issues: all plotted values of $\boldsymbol{K}$ should not be used for the test. We used a vector of length 15, every 10 meters from 10 to 150m. Other choices were possible, but some rules should be followed: choosing the distances up the expected range of interactions, with uniform steps, allows an "objective" analyze of the data, better than selecting values from the plots.

Graphical interpretation

Figure 4 shows the correlation of values of $\widehat{K}(r) - B(r)$ for two different values of r. Each point represents a simulation of a Poisson process. The plot should be imagined in a number of dimensions d equal to the number of r values.

As K is a cumulative function, its values are highly autocorrelated. Figure 4 presents the results of simulations of a Poisson point process. Some are slightly aggregated (positive values), other are dispersed (negative values) due to stochasticity. Multiplying by $\Sigma^{-\frac{1}{2}}$ yields values of $T = \Sigma^{-\frac{1}{2}}\boldsymbol{K}$ indepent, centered, of variance 1 (Figure 5).

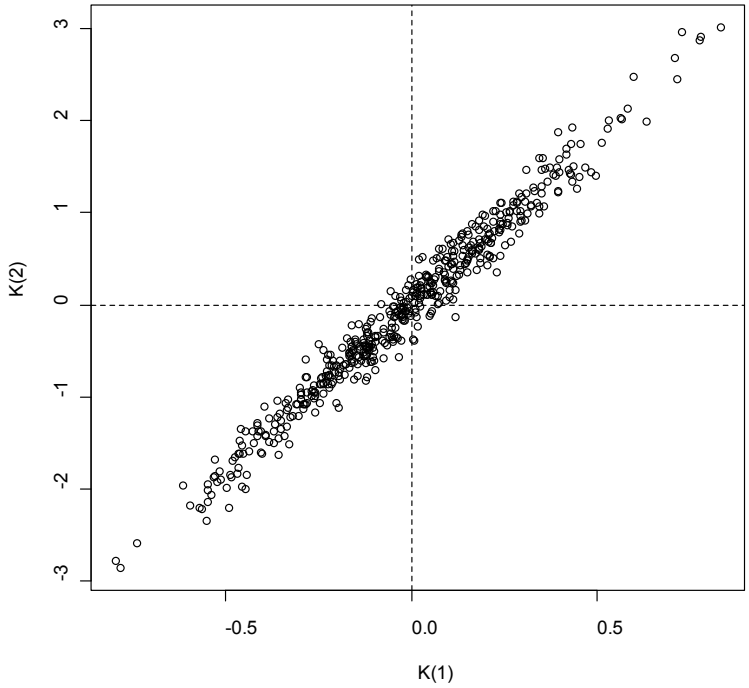


Figure 4: Plot of K in two dimensions ($r = 2$ and $r = 5$) for 500 simulations of a Poisson process of intensity $\rho = 5$ drawn in a square window of size 10.

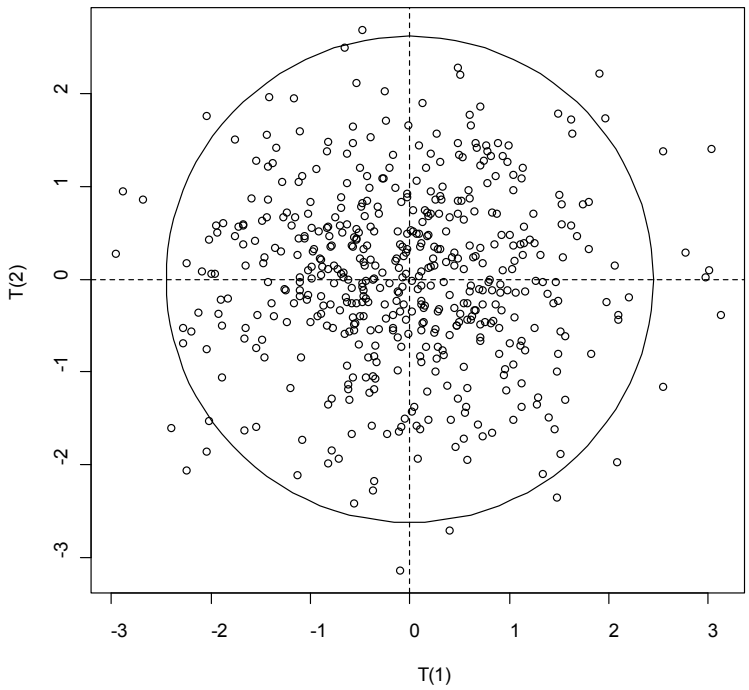


Figure 5: Comparison of values of $T(2)$ and $T(5)$, after transformation of Figure 4. 22 simulations of the homogenous point process out of 500 lie out of the critical circle corresponding to $\|T\| > \chi^2_{5\%}(2)$ so they are rejected by the test.

The circle's radius is the square root of the 5% critical value of a χ^2 distribution with 2 degrees of freedom. Point patterns corresponding to plots outside the circle are rejected. Thus, the test detects significant regularity of points (example not shown) as well as aggregation.

Conclusion

We provide a rigorous statistical test to reject the null hypothesis that K values of an observed point pattern in a rectangle window are that of a realization of a homogenous Poisson point process. Homogeneity of the point process is assumed. This test replaces advantageously the classical Monte-Carlo one. It will rather complete it in practical applications since Monte-Carlo simulations provide useful local information on the point process.

The test is ready to use with the provided R code.

Testing randomness of spatial point patterns with the Ripley statistic

Gabriel Lang*

UMR 518 Mathématique et Informatique appliquées,
AgroParisTech,
19 avenue du Maine,
75732 PARIS CEDEX 15, France.
e-mail: gabriel.lang@agroparistech.fr

Eric Marcon

UMR 745 Ecologie des Forêts de Guyane,
AgroParisTech,
Campus agronomique BP 316,
97379 KOUROU CEDEX, France.
e-mail: eric.marcon@ecofog.gf

Abstract: Aggregation patterns are often visually detected in sets of location data. These clusters may be the result of interesting dynamics or the effect of pure randomness. We build an asymptotically Gaussian test for the hypothesis of randomness corresponding to a Poisson point process. We first compute the exact first and second moment of the Ripley K-statistic under the homogeneous Poisson point process model. Then we prove the asymptotic normality of a vector of such statistics for different scales and compute its covariance matrix. From these results, we derive a test statistic that is chi-square distributed. By a Monte-Carlo study, we check that the test is numerically tractable even for large data sets and also correct when only a hundred of points are observed.

AMS 2000 subject classifications: Primary 60G55, 60F05; secondary 62F03.
Keywords and phrases: Central limit theorem, Gaussian test, Höffding decomposition, K-function, point pattern, Poisson process, U-statistic.

1. Introduction

Analysis of point patterns is relevant in many sciences: cell biology, ecology or spatial economics. The observation of clusters in point locations is considered as a hint for non observable dynamics. For example the clustering of tree locations in a forest may come from better soil conditions or from spreading of seeds of a same mature individual; but clusters are also observed in random distribution as a Poisson point process sample. It is therefore essential to distinguish between clusters resulting from relevant interactions or from complete randomness. Ripley (1976, 1977)— is a widely used tool to quantify the structure of point patterns, especially in ecology, and is well referenced in handbooks (Ripley,

1981; Diggle, 1983; Stoyan et al., 1987; Cressie, 1993; Møller & Waagepetersen, 2004; Ilian et al., 2008). Up to a renormalization by the intensity of the process, this statistic denoted here $\hat{K}(r)$ estimates the expectation $K(r)$ of the number of neighbors at distance less than r of a point in the sample. The observed $\hat{K}(r)$ is compared to the value of $K(r)$ for a homogeneous Poisson point process with the same intensity as the data, chosen as a null hypothesis: the Poisson point process is characterized by an independence of point locations, modelling an absence of interactions between individuals in ecosystems. In this case $K(r)$ is simply the mean number of points in a ball of radius r divided by the intensity, that is πr^2. If $\hat{K}(r)$ is significantly larger than πr^2 (respectively smaller), the process is considered as aggregated (respectively over-dispersed) at distance r.
To decide if the difference is statistically significant, we build a test of the Poisson process hypothesis; we need to know the distribution of $\hat{K}(r)$ for this process. But even the variance is not known and statistical methods generally rely on Monte-Carlo simulations. Ripley (1979) used them to get confidence intervals. Starting from previous results (Saunders & Funk, 1977), he also gave critical values for the L function, a normalized version of K introduced by Besag (1977). These critical values are valid asymptotically, for a large number of points but low intensity, so that both edge effects and point-pair dependence can be neglected. Further computations of confidence interval bands based on simulation have been proposed in Koen (1991) and corrected in Chiu (2007). But the simulation is a practical issue for large point patterns, because computation time is roughly proportional to the square of the number of points (one has to calculate the distances between all pairs of points) multiplied by the number of simulations.
We propose here to compute the exact variance of the Ripley statistic. Ward & Ferrandino (1999) studied this variance. But they ignored that point pairs are not independent even though points are (eq. A8, p. 235), thus their derivation of the variance of $\hat{K}(r)$ was erroneous. The right way to compute the covariance is to consider that it is a U-statistic as remarked in Ripley (1979), then to use the Höffding decomposition. As the variance is not enough to build a test, we study the distribution of the statistic. We prove its asymptotic normality as the size of the observation window grows. It is then easy to build an asymptotically Gaussian test.
Another concern is to test simultaneously the aggregation/dispersion at different scales. This is rarely correctly achieved in practical computations with Monte-Carlo simulations. The confidence bands or test rejection zone are often determined without taking the dependence between the numbers of neighbors at different scales into account. As an exception Duranton & Overman (2005) provide a heuristic multiscale test. In our main theorem, we consider a set of scales $(r_1, \dots, r_d)$, compute the covariance matrix of the $K(r_i)$ and prove the asymptotic normality for the vector $(K(r_1), \dots, K(r_d))$. From this we propose the first rigorous multiscale test of randomness for point patterns.
The paper is built as follows: Section 2 introduces the precise definition of $K(r)$ and the current definition of $\hat{K}(r)$. In Section 3, after the definition of our statistics (no edge-effects correction, known or unknown intensity), we list

the main results of the paper: exact bias due to the edge effects and exact variance of $\hat{K}(r)$ for a homogeneous Poisson process with known or unknown intensity; covariance between $\hat{K}(r)$ and $\hat{K}(r')$ for two different distances r and r'. The main theorem contains the convergence of the vector $(K(r_1), \dots, K(r_d))$ to a Gaussian distribution with explicit covariance in the following asymptotic framework: data from the same process are collected on growing squares of observation. These results allow a simple, multiscale and efficient test procedure of the Poisson process hypothesis. Section 4 provides a Monte-Carlo study of the test and Section 5 gives our conclusions. The last section contains the proofs. Technical integration lemmas are postponed in the appendix.

2. Definition of the Ripley K-function

We recall the characterizations of the dependence of the locations for a general point process X over $\mathbb{R}^2$. We refer to the presentation of Møller & Waagepetersen (2004).

2.1. Definitions

For a point process X, define the point process $X^{(2)}$ on $\mathbb{R}^2 \times \mathbb{R}^2$ of all the couples of two different points of the original process. The intensity of this new process gives information on the simultaneous presence of points in the original process. Denote $\rho^{(2)}(x, y)$ its density (called the second-order product density). The Poisson process of density $\rho(x)$ is such that $\rho^{(2)}(x, y) = \rho(x)\rho(y)$.
The Ripley statistic is a way to estimate the density $\rho^{(2)}(x, y)$. Precisely it is an estimate of the integral on test sets of the ratio $g(x, y) = \rho^{(2)}(x, y)/\rho(x)\rho(y)$. The function $g(x, y)$ characterizes the fact that the points x and y appear simultaneously in the samples of X. If $g(x, y) = 1$, the points appear independently. If $g(x, y) < 1$, they tend to exclude each other; if $g(x, y) > 1$, they appear more frequently together.
We assume the translation invariance of the point process: $g(x, y) = g(x - y)$. In order to estimate the function g, we define its integral as the set function $\mathcal{K}$. Let A be a Borel set:

$$\mathcal{K}(A) = \int_A g(x)dx.$$

If we also assume that the point process is isotropic, we define the Ripley K-function as

$$K(r) = \mathcal{K}(B(x, r)),$$

where $B(x, r)$ is the closed ball with center x and radius r. The translation invariance implies that $\mathcal{K}(B(x, r))$ does not depend on x. For example, if the process is a Poisson process then $g(x) = 1$ and $K(r) = \pi r^2$. We define the Ripley statistic that estimates the K-function. Let A be a bounded Borel set of the plane $\mathbb{R}^2$, m the Lebesgue measure and $\widehat{\rho}$ an estimator of the local intensity

of the process; for a realization S of the point process X, $S = \{X_1, \ldots, X_N\}$, the Ripley statistic is defined by

$$\widehat{K}_A(r) = \frac{1}{m(A)} \sum_{X_i \neq X_j \in S} \frac{\mathrm{I}\{d(X_i, X_j) \leq r\}}{\widehat{\rho}\,(X_i)\ \widehat{\rho}\,(X_j)}.$$

3. Main results

This section presents the theoretical results on the Ripley statistic and the resulting test.

3.1. Definitions

Throughout the paper, we refer to the indicator function I, the expectation $e_{r,n}$, the centred indicator function h and its conditional expectation h_1. We gather here these definitions.
Let n be an integer; A_n denotes the square $[0, n]^2$; U is a random location in A_n with an uniform random distribution; its density is $1/n^2$ with respect to the Lebesgue measure $d\xi_1 d\xi_2$ over A_n. V is a random location with the same distribution as U and independent of U. We denote $d(x, y)$ the Euclidean distance between x and y in the plane, and $\mathrm{I}\{A\}$ the indicator function of set A. We define $e_{r,n} = \mathbb{E}(\mathrm{I}\{d(U, V) \leq r\})$, $h(x, y, r) = \mathrm{I}\{d(x, y) \leq r\} - e_{r,n}$ and $h_1(x, r) = \mathbb{E}(h(U, V, r)|\ V = x)$.

3.2. Assumptions

We assume that X is a homogeneous Poisson process on $\mathbb{R}^2$ with intensity ρ. We consider that the data are available on the square A_n. $S = \{X_1, \ldots, X_N\}$ is the sample of observed points. We consider two cases:

1. If the intensity ρ is known, the Ripley statistic is expressed as

$$\widehat{K}_{1,n}(r) = \frac{1}{n^2\rho^2} \sum_{X_i \neq X_j \in S} \mathrm{I}\{d(X_i, X_j) \leq r\}.$$

2. If the intensity ρ is unknown, we choose to estimate ρ^2 by the unbiased estimator $\widehat{\rho^2} = N(N-1)/n^4$ (Stoyan & Stoyan, 2000) and define

$$\widehat{K}_{2,n}(r) = \frac{n^2}{N(N-1)} \sum_{X_i \neq X_j \in S} \mathrm{I}\{d(X_i, X_j) \leq r\}.$$

3.3. Bias

It is known that a large number of neighbors of the points located near the edges of A_n may lie outside A_n causing a bias in the estimation. We compute the bias due to this edge effect.

Proposition 1. *Assume that $r/n < 1/2$.*

$$\begin{aligned}
\mathbb{E}\widehat{K}_{1,n}(r) - K(r) &= r^2\left(-\frac{8r}{3n} + \frac{r^2}{2n^2}\right).\\
\mathbb{E}\widehat{K}_{2,n}(r) - K(r) &= r^2\left(-\frac{8r}{3n} + \frac{r^2}{2n^2}\right)\\
&\quad -r^2e^{-\rho n^2}\left(\pi - \frac{8r}{3n} + \frac{r^2}{2n^2}\right)\left(1+\rho n^2 e^{-\rho n^2}\right).
\end{aligned}$$

Notes:

- The assumption that r/n is less than $1/2$ means that at least some balls of radius r are included in the square A_n.
- The additional term for $K_{2,n}$ corresponds to the probability to draw a sample with zero or one point in the square. This probability is so low that the term gives a zero contribution as soon as the mean number of points ρn^2 is larger than 20.
- The proof may be adapted for a convex polygon of perimeter Ln to compute the first order term of the bias; for $u = 1$ or 2:

$$\mathbb{E}\widehat{K}_{u,n}(r) - K(r) = -\frac{2Lr^2}{3}\frac{r}{n} + O\left(\frac{r^2}{n^2}\right).$$

3.4. Variance

We compute the covariance matrix of $\widehat{K}_{u,n}(r)$ for $u = 1$ or 2. We get an exact computation for the variance, that can be used for any value of n.

Proposition 2. *For $0 < r < r'$,*

$$\begin{aligned}
var(\widehat{K}_{1,n}(r)) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2e_{r,n}^2}{\rho} + \frac{4n^2}{\rho}\mathbb{E}h_1^2(U,r),\\
cov(\widehat{K}_{1,n}(r),\widehat{K}_{1,n}(r')) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2e_{r',n}e_{r,n}}{\rho} + \frac{4n^2}{\rho}cov(h_1(U,r'),h_1(U,r)),
\end{aligned}$$

$$\begin{aligned}
var(\widehat{K}_{2,n}(r)) &= 2n^4\mathbb{E}\left(\frac{I\{N>1\}}{N(N-1)}\right)\left(e_{r,n} - e_{r,n}^2\right)\\
&+ 4n^4\mathbb{E}\left(\frac{I\{N>1\}(N-2)}{N(N-1)}\right)\mathbb{E}h_1^2(U,r)\\
&+ n^4e^{-\rho n^2}\left(1+\rho n^2\right)\left(1-e^{-\rho n^2}-\rho n^2e^{-\rho n^2}\right)e_{r,n}^2,\\
cov(\widehat{K}_{2,n}(r),\widehat{K}_{2,n}(r')) &= 2n^4\mathbb{E}\left(\frac{I\{N>1\}}{N(N-1)}\right)\left(e_{r,n} - e_{r',n}e_{r,n}\right)\\
&+ 4n^4\mathbb{E}\left(\frac{I\{N>1\}(N-2)}{N(N-1)}\right)cov(h_1(U,r'),h_1(U,r))\\
&+ n^4e^{-\rho n^2}\left(1+\rho n^2\right)\left(1-e^{-\rho n^2}-\rho n^2e^{-\rho n^2}\right)e_{r',n}e_{r,n},
\end{aligned}$$

where

$$\begin{aligned} e_{r,n} &= \frac{\pi r^2}{n^2} - \frac{8r^3}{3n^3} + \frac{r^4}{2n^4}, \\ \mathbb{E}h_1^2(U,r) &= \frac{r^5}{n^5}\left(\frac{8}{3}\pi - \frac{256}{45}\right) + \frac{r^6}{n^6}\left(\frac{11}{48}\pi - \frac{56}{9}\right) + \frac{8}{3}\frac{r^7}{n^7} - \frac{1}{4}\frac{r^8}{n^8}. \end{aligned}$$

Notes:

- The variances of both estimators are exact and can be computed at any precision, as inverse moments of the Poisson variable correspond to fast converging series. But these series may be difficult to evaluate with mathematical softwares, because of the large value of the Poisson parameter.
- The covariances are not explicit because the terms $\operatorname{cov}(h_1^2(U,r'), h_1^2(U,r))$ involve terms that have to be numerically integrated.
- The leading terms of the variances of $K_{1,n}(r)$ and $K_{2,n}(r)$ as n tends to infinity are $2\pi r^2/n^2\rho^2 + 4\pi r^4/n^2\rho$ and $2\pi r^2/n^2\rho^2$.

3.5. Central Limit Theorem

We show that a normalized vector of Ripley statistics for different r converges in distribution to a normal vector. Let $\mathcal{N}(0,\Sigma)$ denote the Gaussian multivariate centred distribution with covariance matrix Σ.

Theorem 1. *Let d be an integer, $0 < r_1 < \ldots < r_d$ a set of reals and define $\mathcal{K}_{u,n} = (\widehat{K}_{u,n}(r_1), \ldots, \widehat{K}_{u,n}(r_d))$. Then $n\sqrt{\rho}(\mathcal{K}_{u,n} - \pi(r_1^2, \ldots r_d^2))$ converges in distribution to $\mathcal{N}(0,\Sigma)$ as n tends to infinity, where for s and t in $\{1,\ldots,d\}$*

- *if $u = 1$, $\Sigma_{s,t} = \dfrac{2\pi(r_s^2 \wedge r_t^2)}{\rho} + 4\pi^2 r_s^2 r_t^2$.*
- *if $u = 2$, $\Sigma_{s,t} = \dfrac{2\pi(r_s^2 \wedge r_t^2)}{\rho}$.*

Note: The first term of the variance corresponds to a situation where the couples of points are independent from each others; this was used as an approximation without proof in Ward & Ferrandino (1999); our work proves that the actual variance and limit process are different in the first case and that the approximation holds only in the second case.

3.6. Applications to test statistics

From Theorem 1, we deduce that $T_u = \Sigma^{-1/2}\mathcal{K}_{u,n}$ is asymptotically $\mathcal{N}(0, I_d)$ distributed. For the hypothesis

$$H_0\text{: } X \text{ is a homogeneous Poisson process of intensity } \rho$$

we use $T^2 = \|T_u\|_2^2$ as a test statistic with rejection zone for the level α:

$$T^2 > \chi_\alpha^2(d).$$

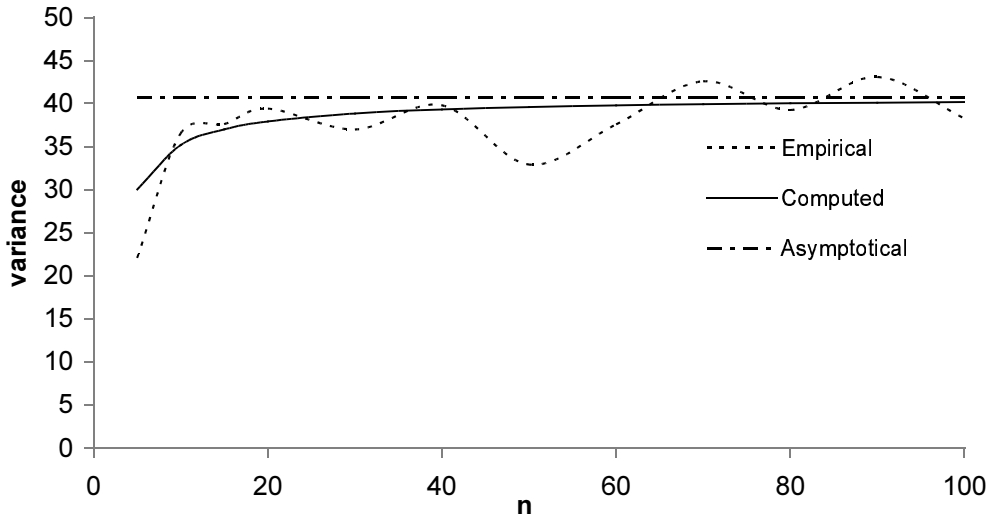


FIGURE 1. *Comparison of normalized variances for* $K_1(1)$, $\rho = 5$

where $\chi^2_\alpha(d)$ is the $(1-\alpha)$-quantile of the $\chi^2(d)$ distribution.
Note: the covariance matrix Σ depends on the intensity parameter ρ, so that in the case of the unknown parameter we have to use an estimate of ρ in the formula defining Σ.

4. Simulations

We study the empirical variance of the proposed statistics by a Monte-Carlo simulation. Then we apply the test procedure to simulated data sets, observe the number of rejections and compare it to the level of the test.

4.1. Variance

We simulate a sample of 1000 repetitions with $\rho = 5$ and compare (after renormalization by $n\sqrt{\rho}$) the empirical variance and the exact computed variance with the limit variance for different value of n (figure 1). With 1000 repetitions, the oscillations of the empirical variance are still large; we will use a larger number of repetitions in the following study of the test.
The convergence of the computed variance to the limit value is not so fast and for applications with hundreds of points (corresponding in figure 1 to $n < 15$) the distance between the variances is still large. A preliminary study, not presented here, showed that the test procedure is perturbed by an small error in the covariance matrix, as we tried simplified versions of the covariance by bounding

TABLE 1
Percentile of rejection over 10000 *repetitions of the test with level* $\alpha = 0.05$.

Poisson			T_1^*	T_1'	T_1	T_2^*	T_2
$n = 30$	$\rho = 1$	$r = (1, 2, 5)$	5.40	5.04	5.20	5.01	5.10
$n = 10$	$\rho = 5$	$r = (1, 2, 5)$	5.61*	5.40	5.19	5.38	5.37
$n = 10$	$\rho = 5$	$r = (1, 2, \ldots, 10)$	5.13	5.32	5.76*	6.67*	6.01*
$n = 10$	$\rho = 1$	$r = (1, 2, 5)$	5.67*	5.86*	5.81*	5.30	5.25
$n = 10$	$\rho = .5$	$r = (1, 2, 5)$	5.52*	5.73*	5.52*	5.60*	4.91
$n = 10$	$\rho = .2$	$r = (1, 2, 5)$	6.40*	6.84*	6.59*	6.59*	5.22

or ignoring the corner contribution $C(A_n^{3,3})$ (see in the proof section). It is crucial to use an accurate computation of the covariance matrix to have a correct approximation of the square root inverse matrix $\Sigma^{-1/2}$. Therefore we will use the exact formula instead of the asymptotic formula in the test procedure.

4.2. Test

In the known parameter case, the computation of the test statistic T_1 is straightforward; we also build a statistic T_1^* using the empirical covariance matrix of the sample. The advantage of T_1^* is that it is orthogonal by construction and should lead to better results. But the covariance matrix is not observable when we dispose of one sample, so that the test procedure based on T_1^* is unfeasible. It is an idealized version, used to compare the corresponding number of rejections. To avoid the statistical dependence between the sample and the estimator of the covariance matrix, we also build a statistic T_1' where we generate a additional independent sample of the Poisson process with intensity ρ to compute the empirical covariance matrix.
In the unknown parameter case, the computation of the test statistic T_2 is similar. In the variance formula the unknown parameter ρ is replaced by the estimator N/n^2. We also choose to replace the expectation $\mathbb{E}\left(\mathrm{I}\{N > 1\}/(N(N-1))\right)$ by the observed value $1/(N(N-1))$ and $\mathbb{E}\left(\mathrm{I}\{N > 1\}(N-2)/(N(N-1))\right)$ by $(N-2)/(N(N-1))$, because the dispersion of a Poisson variable is low with respect to the expectation when its intensity is large. The construction of T_2^* is the same as for T_1^*. The case of T_2' is not studied because, as ρ is unknown, one would have to generate an additional sample for each estimated value of ρ.
The test output is a Bernoulli random variable with parameter α. With a sufficient index of repetition m, the mean number of rejection is close to a normal variable with expectation α and variance $\alpha(1-\alpha)/m$. We consider that the test works when the observed frequency of rejection is in the 95% Gaussian confidence interval $[\alpha - 1.96\sqrt{\alpha(1-\alpha)/m}, \alpha + 1.96\sqrt{\alpha(1-\alpha)/m}]$. With $m = 10000$ and $\alpha = 0.05$, the interval is $[0.0457; 0.0543]$ so that the percentile of rejection in table 1 should lie in $[4.57; 5.43]$. Stars indicate the values outside the confidence interval.
The performances in the case of a known parameter (T_1, T_1^* and T_1') are good except when the number of points is small. The unfeasible tests T_1^* and T_1' based

TABLE 2
Percentile of rejection over 10000 *repetitions of the test with level* $\alpha = 0.05$.

Thomas			T_2
$n = 10$	$(\kappa, \mu, \sigma) = (1, 5, 3)$	$r = (1, 2, 5)$	71.6
$n = 10$	$(\kappa, \mu, \sigma) = (0.5, 10, 0.5)$	$r = (1, 2, 5)$	100

on the empirical covariance have no better performance than the test T_1. The error of the empirical covariance is probably still to large. The only exception is the third line where a large number of values of r are considered simultaneously. The test T_2 performs better than T_1 for small data sets. The only exception is the case of a large number of scales. The poor performance of T_1 and T_2 in this case may result from numerical instabilities in the covariance matrix inversion as its dimension is larger. The departure from normality may also be larger in this case (some classes of inter-point distances being weakly represented in the sample). With this exception, the test based on T_2 works perfectly.
In table 2, we investigate the power of the test T_2 by simulating two Thomas cluster processes (Thomas, 1949). A Thomas process is a Neyman-Scott process; the germs of the clusters are drawn as a sample of a homogeneous Poisson process of intensity κ . For each germ, an inhomogeneous Poisson process is drawn with intensity measure μf, where f is the density of the Gaussian two-dimensional vector centered on the germ and with independent coordinates of variance σ. The Thomas process results from the superposition of these Poisson processes. The germs are not conserved. The parameters of the two processes are such that clusters are not visually detectable in the first process and evident in the second one. The test rejects 71% of the first sample and systematically the second one. The test is more powerful than a visual observation of the data, detecting invisible clusters. A rigorous analysis of the distribution of the statistic for dependent point process models should allow to conclude on the power of our test but such a study is beyond the scope of this paper.

5. Conclusion

We provide an efficient test of the null hypothesis of a homogeneous Poisson process for point patterns in a square domain. This is a theoretical and practical improvement on preexisting methods: Monte-Carlo simulations are untractable when the number of points increases. With a personal computer, calculating K for 10,000 simulations of a 10,000-point set is not feasible (or it will take months). Marcon & Puech (2003) applied K to a 36,000-point data set (the largest ever published as far as we know), but had to limit the number of simulations to 20. We suggest to change the treatment of edge effects. Instead of correcting edge effect on each sample to reduce the bias, we compute the exact bias. The use of sample correction (for each point of the data) has not been questioned since Ripley's original paper, except by Ward & Ferrandino (1999).
We also point out that the test can be used on samples with a few dozens of points as encountered in actual data sets. It works correctly with such small

data sets, even if it is based on asymptotic normality. This is due to the fact that the bias and variance are known exactly and not asymptotically; the non-normality of the statistics for small data sets seems to have lesser effects than approximating the variance.
Our work should be extended in two directions: to other domain shapes that are of interest for the practitioners and to 3-dimensional data for high resolution medical imagery. A further study of the asymptotics of the distribution of $\hat{K}(r)$ for dependent point process models such as Markov or Cox processes should also be achieved to inform on the power of our test.

6. Proofs

6.1. Proof of proposition 1

Recall that U and V are two independent uniform variables on A_n. The expectations of the Ripley statistics are

$$\begin{aligned}
\mathbb{E}\widehat{K}_{1,n}(r) &= \frac{1}{n^2\rho^2}\mathbb{E}\left(\sum_{X_i\neq X_j\in S} \mathrm{I}\{d(X_i,X_j)\leq r\}\right)\\
&= \frac{\mathbb{E}\left(N(N-1)\right)}{n^2\rho^2}\mathbb{E}(\ \mathrm{I}\{d(U,V)\leq r\})\\
&= n^2 e_{r,n}.\\
\mathbb{E}\widehat{K}_{2,n}(r) &= n^2\mathbb{E}\left(\frac{1}{N(N-1)}\sum_{X_i\neq X_j\in S} \mathrm{I}\{d(X_i,X_j)\leq r\}\right)\\
&= n^2\mathbb{P}\left(N>1\right)\mathbb{E}(\ \mathrm{I}\{d(U,V)\leq r\})\\
&= n^2\left(1-e^{-\rho n^2}-\rho n^2 e^{-\rho n^2}\right)e_{r,n}.
\end{aligned}$$

The following lemma allows to conclude:

Lemma 1.

$$e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8r^3}{3n^3} + \frac{r^4}{2n^4}.$$

Proof: We split A_n into four parts to compute $e_{r,n}$:

$$\begin{aligned}
e_{r,n} &= \int_{\xi\in A_n^1}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\leq r\}\frac{1}{n^4}d\xi d\eta &(1)\\
&+ \int_{\xi\in A_n^2}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\leq r\}\frac{1}{n^4}d\xi d\eta &(2)\\
&+ \int_{\xi\in A_n^3}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\leq r\}\frac{1}{n^4}d\xi d\eta &(3)\\
&+ \int_{\xi\in A_n^4}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\leq r\}\frac{1}{n^4}d\xi d\eta &(4)
\end{aligned}$$

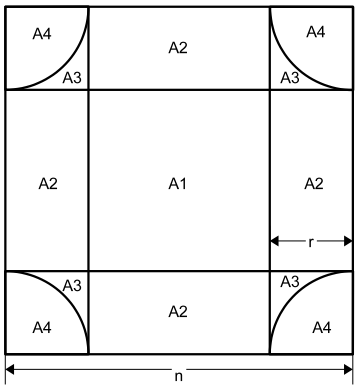


FIGURE 2. *Zones in the square*

where (see figure 2)

- (interior) $A_n^1 = \{\xi,\ \xi$ is at distance larger than r from the boundary$\}$
- (edge) $A_n^2 = \{\xi,\ \xi$ is at distance less than r from an edge, larger than r from the others$\}$
- (two edges) $A_n^3 = \{\xi,\ \xi$ is at distance less than r from two edges and larger than r from the corner$\}$
- (corner) $A_n^4 = \{\xi,\ \xi$ is at distance less than r from the corner$\}$

Note that A_n^2, A_n^3 and A_n^4 are composed of four parts that contribute identically. We establish formulas only for one of these parts.

Lemma 2. *Define function* $g(x) = \arccos(x) - x\sqrt{1-x^2}$.
If $\xi \in A_n^1$,

$$\int_{\eta \in A_n} I\{d(\xi,\eta) \le r\} d\eta = \pi r^2.$$

If $\xi \in A_n^2$, *with* $n - r < \xi_1 < n$, $x_1 = \frac{1}{r}(n - \xi_1)$,

$$\int_{\eta \in A_n} I\{d(\xi,\eta) \le r\} d\eta = r^2(\pi - g(x_1))$$

If $\xi \in A_n^3$, *with* $n - r < \xi_1 < n$, $n - r < \xi_2 < n$ *and* $(x_1, x_2) = \frac{1}{r}(n - \xi_1, n - \xi_2)$,

$$\int_{\eta \in A_n} I\{d(\xi,\eta) \le r\} d\eta = r^2(\pi - g(x_1) - g(x_2)).$$

If $\xi \in A_n^4$, *with* $n - r < \xi_1 < n$, $n - r < \xi_2 < n$ *and* $(x_1, x_2) = \frac{1}{r}(n - \xi_1, n - \xi_2)$,

$$\int_{\eta \in A_n} I\{d(\xi,\eta) \le r\} d\eta = r^2 \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right).$$

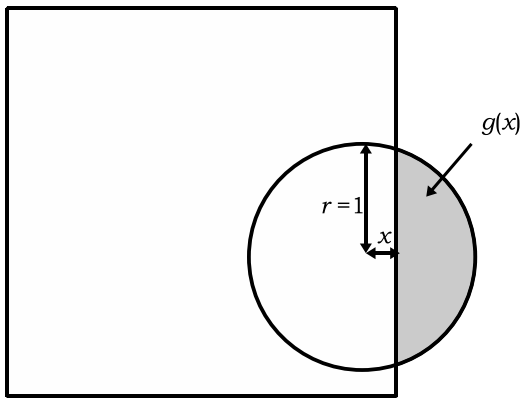


FIGURE 3. *Geometrical interpretation of g*

Note: Function $g(x)$ is the area of the part of a ball of radius 1 that lies outside the square when the ball intersects one of its edges (see figure 3).
Proof. For the interior points $\xi \in A_n^1$, $B(\xi, r) \subset A_n$.
Let $\xi \in A_n^2$. We compute the area of $B(\xi, r) \cap A_n$.

$$\begin{aligned}\int_{\eta \in A_n} \mathrm{I}\{d(\xi,\eta) \leq r\} d\eta &= \frac{\pi r^2}{2} + 2r^2 \int_0^{x_1} \sqrt{1-t^2}dt \\ &= r^2 \left(\pi - \arccos(x_1) + x_1\sqrt{1-x_1^2}\right) \\ &= r^2 \left(\pi - g(x_1)\right).\end{aligned}$$

Note that $r^2 g(x)$ is the part of the ball that lies out of the square A_n if the center is at distance xr from the edge of the square.
Let $\xi \in A_n^3$. Here the ball intersects two edges of the square and the area of $B(\xi, r) \cap A_n$ is

$$\int_{\eta \in A_n} \mathrm{I}\{d(\xi,\eta) \leq r\} d\eta = r^2 \left(\pi - g(x_1) - g(x_2)\right).$$

Let $\xi \in A_n^4$. Divide the ball into four quarters along axes parallel to the coordinate axes. One of the quarter is inside the square, two intersect the edges, leaving outside an area equal to $(g(x_1) + g(x_2))/2$. The area of the intersection of the last quarter with the square is $x_1 x_2$ so that the area of $B(\xi, r) \cap A_n$ is

$$\int_{\eta \in A_n} \mathrm{I}\{d(\xi,\eta) \leq r\} d\eta = r^2 \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2}\right). \quad \square$$

Proof of lemma 1(continued). The left-hand side of (1) is $m(A_n^1)\pi r^2 = \pi(n - 2r)^2 r^2$. Recall that A_n^2 is composed of four parts that contribute identically. We integrate function g.

Lemma 3.

$$G(x) = \int_0^x g(u)du = x\arccos(x) - \sqrt{1-x^2} + \frac{1}{3}(1-x^2)^{3/2} + \frac{2}{3}.$$

Proof. Changing variables and integrating by parts

$$\begin{aligned}\int_0^x \arccos(u)du &= -\int_{\pi/2}^{\arccos(x)} t\sin(t)dt \\ &= [t\cos(t)]_{\pi/2}^{\arccos(x)} + \int_{\pi/2}^{\arccos(x)} \cos(t)dt \\ &= x\arccos(x) - \sqrt{1-x^2} + 1.\end{aligned}$$

Changing the variable $v = \sqrt{1-u^2}$, we get

$$-\int_0^x u\sqrt{1-u^2}du = \int_1^{\sqrt{1-x^2}} v^2 dv = \frac{1}{3}\left((1-x^2)^{3/2} - 1\right). \quad \square$$

Then the contribution (2) is equal to

$$4r\int_r^{n-r} d\xi_2 \int_0^1 r^2(\pi - g(x))dx = 4r^3(n-2r)(\pi - G(1)) = \left(4\pi - \frac{8}{3}\right) r^3(n-2r).$$

We consider A_n^3; the domain of integration is symmetric in (x_1, x_2) so that the contribution (3) is equal to

$$4r^4\int_0^1 dx_1 \int_{\sqrt{1-x_1^2}}^1 (\pi - 2g(x_1))dx_2 = 4r^4\left(\pi\left(1-\frac{\pi}{4}\right) - 2\int_0^1 g(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 dx_2\right).$$

From Lemma 6,

$$\int_0^1 g(x_1)dx_1 \int_{\sqrt{1-x_1^2}}^1 dx_2 = G(1) - \int_0^1 g(x_1)\sqrt{1-x_1^2}dx_1 = \frac{2}{3} - \frac{\pi^2}{16}.$$

so that contribution (3) is equal to $r^4\left(4\pi - \frac{\pi^2}{2} - \frac{16}{3}\right)$.

We consider A_n^4; the contribution (4) is equal to

$$\begin{aligned}4r^4\int_0^1 dx_1 \int_0^{\sqrt{1-x_1^2}} \left(\frac{3\pi}{4} + x_1x_2 - g(x_1)\right)dx_2 &= r^4\left(\frac{3\pi^2}{4} + \frac{1}{2} - 4\int_0^1 g(x_1)\sqrt{1-x_1^2}dx_1\right) \\ &= r^4\left(\frac{\pi^2}{2} + \frac{1}{2}\right).\end{aligned}$$

Gathering the four contributions, we get

$$\begin{aligned}e_{r,n} &= \frac{r^2}{n^2}\left(\pi\left(1-\frac{2r}{n}\right)^2 + \left(4\pi - \frac{8}{3}\right)\frac{r}{n}\left(1-\frac{2r}{n}\right) + \left(4\pi - \frac{29}{6}\right)\frac{r^2}{n^2}\right) \\ &= \frac{r^2}{n^2}\left(\pi - \frac{8}{3}\frac{r}{n} + \frac{1}{2}\frac{r^2}{n^2}\right). \quad \square\end{aligned}$$

6.2. Proof of proposition 2

We decompose the variance of $K_{s,A_n}(r)$ by conditioning the variable with respect to the number N of points in the sample. Conditionally to N, $K_{s,A_n}(r)$ has the form of a U-statistic. Then we apply the Höffding decomposition to this U-statistic.
For $s = 1, 2$, we use the relation

$$\operatorname{var}(\widehat{K}_{s,A_n}(r)) = \operatorname{var}\mathbb{E}(\widehat{K}_{s,A_n}(r)|N) + \mathbb{E}\operatorname{var}(\widehat{K}_{s,A_n}(r)|N).$$

We first consider the conditional expectation of $\widehat{K}_{s,A_n}(r)$.

$$\begin{aligned}
\mathbb{E}(\widehat{K}_{1,n}(r)|N) &= \frac{1}{n^2\rho^2}\left(\sum_{i\neq j=1}^{N} \mathbb{E}\ \mathrm{I}\{d(X_i, X_j) \leq r\}\right) = \frac{N(N-1)e_{r,n}}{n^2\rho^2},\\
\mathbb{E}(\widehat{K}_{2,n}(r)|N) &= \frac{n^2}{N(N-1)}\sum_{i\neq j=1}^{N} \mathbb{E}\ \mathrm{I}\{d(U_i, U_j) \leq r\} = n^2 e_{r,n}\ \mathrm{I}\{N > 1\}.
\end{aligned}$$

Because N is a Poisson variable with intensity ρn^2

$$\begin{aligned}
\mathbb{E}N^2(N-1)^2 &= \mathbb{E}N(N-1)(N-2)(N-3)\\
&+ 4\mathbb{E}N(N-1)(N-2) + 2\mathbb{E}N(N-1)\\
&= \rho^4 n^8 + 4\rho^3 n^6 + 2\rho^2 n^4.\\
\operatorname{var} N(N-1) &= 4\rho^3 n^6 + 2\rho^2 n^4. \qquad (5)
\end{aligned}$$

Then

$$\begin{aligned}
\operatorname{var}\mathbb{E}(\widehat{K}_{1,n}(r)|N) &= \frac{(4\rho n^2 + 2)e_{r,n}^2}{\rho^2}. \qquad (6)\\
\operatorname{var}\mathbb{E}(\widehat{K}_{2,n}(r)|N) &= n^4\mathbb{P}\{N > 1\}(1 - \mathbb{P}\{N > 1\})e_{r,n}^2\\
&= n^4 e^{-\rho n^2}\left(1 + \rho n^2\right)\left(1 - e^{-\rho n^2}\left(1 + \rho n^2\right)\right)e_{r,n}^2. \qquad (7)
\end{aligned}$$

We compute the conditional variances.

$$\begin{aligned}
\operatorname{var}(\widehat{K}_{1,n}(r)|N) &= \frac{1}{n^4\rho^4}\operatorname{var}\left(\sum_{i\neq j=1}^{N} h(X_i, X_j, r)\right),\\
\operatorname{var}(\widehat{K}_{2,n}(r)|N) &= \frac{n^4}{N^2(N-1)^2}\operatorname{var}\left(\sum_{i\neq j=1}^{N} h(X_i, X_j, r)\right).
\end{aligned}$$

Conditionally to N, the locations of the points are independent and uniformly distributed variables U_i over A_n. We introduce the Höffding decomposition of the U-statistic kernel h:

$$h(x, y, r) = h_1(x, r) + h_1(y, r) + h_2(x, y, r),$$

where $h_1(x) = \mathbb{E}(h(U,V,r)|V = x)$, (U,V) being two independent uniform random variables on A_n.
Then $\mathbb{E}h_1(U,r) = 0$ and $\mathbb{E}(h_2(U,V,r)|U) = \mathbb{E}(h_2(U,V,r)|V) = 0$, so that

$$\begin{aligned}\operatorname{var} h(U,V,r) &= \operatorname{var} h_1(U,r) + \operatorname{var} h_1(V,r) + \operatorname{var} h_2(U,V,r)\\ &= 2\mathbb{E}h_1^2(U,r) + \operatorname{var} h_2(U,V,r).\end{aligned}$$

From

$$\sum_{i\neq j=1}^{N} h(U_i,U_j,r) = 2(N-1)\sum_{i=1}^{N} h_1(U_i,r) + \sum_{i\neq j=1}^{N} h_2(U_i,U_j,r).$$

we get

$$\begin{aligned}\operatorname{var}(\widehat{K}_{1,n}(r)|N) &= \frac{4(N-1)^2}{n^4\rho^4}\operatorname{var}\left(\sum_{i=1}^{N} h_1(U_i,r)\right) + \frac{1}{n^4\rho^4}\operatorname{var}\left(\sum_{i\neq j=1}^{N} h_2(U_i,U_j,r)\right)\\ &= \frac{4N(N-1)^2}{n^4\rho^4}\mathbb{E}h_1^2(U,r) + \frac{2}{n^4\rho^4}\sum_{i\neq j=1}^{N}\operatorname{var} h_2(U_i,U_j,r)\\ &= \frac{4N(N-1)^2}{n^4\rho^4}\mathbb{E}h_1^2(U,r) + \frac{2N(N-1)}{n^4\rho^4}(\operatorname{var} h(U,V,r) - 2\mathbb{E}h_1^2(U,r))\\ &= \frac{4N(N-1)(N-2)}{n^4\rho^4}\mathbb{E}h_1^2(U,r) + \frac{2N(N-1)}{n^4\rho^4}\operatorname{var} h(U,V,r),\end{aligned}$$

Now $\operatorname{var} h(U,V,r) = e_{r,n} - e_{r,n}^2$ and using factorial moments of the Poisson distribution

$$\mathbb{E}\operatorname{var}(\widehat{K}_{1,n}(r)|N) = \frac{4n^2}{\rho}\mathbb{E}h_1^2(U,r) + \frac{2}{\rho^2}\left(e_{r,n} - e_{r,n}^2\right). \tag{8}$$

Lemma 4 gives the exact value of $\mathbb{E}h_1^2(U,r)$. With relations (6) and (8), we get

$$\begin{aligned}\operatorname{var}(\widehat{K}_{1,n}(r)) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2e_{r,n}^2}{\rho} + \frac{4n^2}{\rho}\mathbb{E}h_1^2(U_j,r)\\ &= \frac{1}{n^2}\left(\frac{2\pi r^2}{\rho^2} + \frac{4\pi^2r^4}{\rho}\right)\\ &- \frac{1}{n^3}\left(\frac{16}{3}\frac{r^3}{\rho^2} + \left(\frac{32\pi}{3} + \frac{1024}{45}\right)\frac{r^5}{\rho}\right)\\ &+ \frac{1}{n^4}\left(\frac{r^4}{\rho^2} + \left(\frac{59\pi}{12} + \frac{32}{9}\right)\frac{r^6}{\rho}\right).\end{aligned}$$

Similarly

$$\begin{aligned}
\operatorname{var}(\widehat{K}_{2,n}(r)|N) &= \frac{\mathrm{I}\{N>1\}(N-2)}{N(N-1)}\mathbb{E}h_1^2(U,r) + \frac{2n^4\ \mathrm{I}\{N>1\}}{N(N-1)}\operatorname{var} h(U,V,r),\\
\mathbb{E}\operatorname{var}(\widehat{K}_{2,n}(r)|N) &= \mathbb{E}\left(\frac{\mathrm{I}\{N>1\}(N-2)}{N(N-1)}\right)\mathbb{E}h_1^2(U,r)\\
&+ 2n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}}{N(N-1)}\right)\left(e_{r,n}-e_{r,n}^2\right).
\end{aligned}$$

From this and relation (7), we get

$$\begin{aligned}
\operatorname{var}(\widehat{K}_{2,n}(r)) &= 2n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}}{N(N-1)}\right)\left(e_{r,n}-e_{r,n}^2\right)\\
&+ 4n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}(N-2)}{N(N-1)}\right)\mathbb{E}h_1^2(U_j,r)\\
&+ n^4e^{-\rho n^2}\left(1+\rho n^2\right)\left(1-e^{-\rho n^2}-\rho n^2e^{-\rho n^2}\right)e_{r,n}^2.
\end{aligned}$$

We now apply the same decomposition to $\operatorname{cov}(\widehat{K}_{1,n}(r),\widehat{K}_{1,n}(r'))$,

$$\operatorname{cov}(\mathbb{E}(\widehat{K}_{1,n}(r')|N),\mathbb{E}(\widehat{K}_{1,n}(r)|N)) = \frac{(4\rho n^2+2)e_{r',n}e_{r,n}}{\rho^2}. \tag{9}$$

$$\begin{aligned}
\operatorname{cov}(\widehat{K}_{1,n}(r'),\widehat{K}_{1,n}(r)|N) &= \frac{4(N-1)^2}{n^4\rho^4}\operatorname{cov}\left(\sum_{i=1}^N h_1(U_i,r'),\sum_{i=1}^N h_1(U_i,r)\right)\\
&+ \frac{1}{n^4\rho^4}\operatorname{cov}\left(\sum_{i\neq j=1}^N h_2(U_i,U_j,r'),\sum_{i\neq j=1}^N h_2(U_i,U_j,r)\right)\\
&= \frac{4N(N-1)(N-2)}{n^4\rho^4}\operatorname{cov}(h_1(U,r'),h_1(U,r))\\
&+ \frac{2N(N-1)}{n^4\rho^4}\operatorname{cov}(h(U,V,r'),h(U,V,r)).
\end{aligned}$$

$$\mathbb{E}\operatorname{cov}(\widehat{K}_{1,n}(r'),\widehat{K}_{1,n}(r)|N) = \frac{4n^2}{\rho}\operatorname{cov}(h_1(U,r'),h_1(U,r)) + \frac{2}{\rho^2}(e_{r,n}-e_{r',n}e_{r,n}).$$

To compute $\operatorname{cov}(h_1(U,r'),h_1(U,r))$, the square A_n should now be split into 16 different zones according to the 4 zones of the preceding section with respect to r and the 4 zones with respect to r'. Because of inclusions, the actual number of zones to consider is reduced to 9. The corresponding computation is easy in the center zone, but can not be achieved in a close form in the edge bands and in the corner. We consider the following zones:

- (interior) $A_n^{1,1}$ =$\{\xi,\ \xi$ is at distance larger than r' from the boundary$\}$,

- (interior-edge) $A_n^{1,2}$ =$\{\xi,\ \xi$ is at distance between r and r' from an edge, larger than r' from the others$\}$,
- (edge) $A_n^{2,2}$ =$\{\xi,\ \xi$ is at distance less than r from an edge, larger than r' from the others$\}$,
- (corner) $A_n^{3,3}$ =$\{\xi,\ \xi$ is at distance less than r' from two edges$\}$.

Denoting $x_1 = \frac{1}{r}(n-\xi_1)$ and $x_1' = \frac{1}{r'}(n-\xi_1)$ we get

$$\begin{aligned} h_1(X_j,r')h_1(X_j,r) =& \left(\frac{\pi r'^2}{n^2} - e_{r',n}\right)\left(\frac{\pi r^2}{n^2} - e_{r,n}\right) \text{ on } A_n^{1,1},\\ =& \left(\frac{\pi r'^2}{n^2} - e_{r',n} - \frac{r'^2}{n^2}g(x_1')\right)\left(\frac{\pi r^2}{n^2} - e_{r,n}\right) \text{ on } A_n^{1,2},\\ =& \left(\frac{\pi r'^2}{n^2} - e_{r',n} - \frac{r'^2}{n^2}g(x_1')\right)\left(\frac{\pi r^2}{n^2} - e_{r,n} - \frac{r^2}{n^2}g(x_1)\right) \text{ on } A_n^{2,2}. \end{aligned}$$

Denote $b_{r,n} = \left(\pi - \frac{n^2}{r^2}e_{r,n}\right) = \frac{8r}{2n} - \frac{r^2}{2n^2}$.

$$\operatorname{cov}\left(h_1(X_j,r'),h_1(X_j,r)\right) = C(A_n^{1,1}) + C(A_n^{1,2}) + C(A_n^{2,2}) + C(A_n^{3,3})$$

$$\begin{aligned} C(A_n^{1,1}) &= \frac{r'^2r^2}{n^4}\left(1-\frac{2r'}{n}\right)^2 b_{r',n}b_{r,n}\\ C(A_n^{1,2}) &= 4\left(1-\frac{2r'}{n}\right)\frac{r'^3r^2}{n^5}b_{r,n}\int_{r/r'}^1 (b_{r',n}-g(x_1'))dx_1'\\ C(A_n^{2,2}) &= 4\left(1-\frac{2r'}{n}\right)\frac{r^3r'^2}{n^5}\int_0^1 (b_{r',n}-g(rx_1/r'))(b_{r,n}-g(x_1))dx_1. \end{aligned}$$

The first integral may be expressed in terms of function G, the second integral is elliptic and has to be numerically evaluated; as the integrand is bounded and very smooth this can be achieved without difficulties. To compute the term $C(A_n^{3,3})$, we rewrite the different values of function h_1 with the help of indicator functions:

$$\begin{aligned} h_{A1}(x,r) &= b_{r,n}\ \mathrm{I}\{x_1\geq 1; x_2\geq 1\}\\ h_{A2}(x,r) &= (b_{r,n}-g(x_2))\ \mathrm{I}\{x_1\geq 1;\ x_2<1\} + (b_{r,n}-g(x_1))\ \mathrm{I}\{x_2\geq 1;\ x_1<1\}\\ h_{A3}(x,r) &= (b_{r,n}-g(x_1)-g(x_2))\ \mathrm{I}\{x_1<1;\ x_2<1;\ x_1^2+x_2^2\geq 1\}\\ h_{A4}(x,r) &= (b_{r,n}-\pi/4+x_1x_2-(g(x_1)+g(x_2))/2)\ \mathrm{I}\{x_1^2+x_2^2<1\} \end{aligned}$$

For $x' = \frac{1}{r'}(n-\xi_1, n-\xi_2)$

$$C(A_n^{3,3}) = 4\frac{r^2r'^4}{n^6}\int_0^1\int_0^1 \sum_{i=1}^4 h_{Ai}(r'x'/r,r)\times\sum_{i=3}^4 h_{Ai}(x',r')dx_1'dx_2'$$

and this integral also can be numerically evaluated.
Note: the whole computation of this term of the covariance could be numerically

achieved, but it is preferable to use an exact computation whenever it is possible. The case of the covariance of $K_{2,n}(r)$ is analogous:

$$\operatorname{cov}\left(\mathbb{E}(\widehat{K}_{2,n}(r')|N), \mathbb{E}(\widehat{K}_{2,n}(r)|N)\right) = n^4 e^{-\rho n^2}\left(1+\rho n^2\right)\left(1-e^{-\rho n^2}(1+\rho n^2)\right)e_{r',n}e_{r,n}.$$

$$\begin{aligned}\mathbb{E}\operatorname{cov}\left(\widehat{K}_{2,n}(r'), \widehat{K}_{2,n}(r)|N\right) &= 4n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}(N-2)}{N(N-1)}\right)\operatorname{cov}\left(h_1(U,r'), h_1(U,r)\right)\\ &\quad + 2n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}}{N(N-1)}\right)\left(e_{r,n} - e_{r',n}e_{r,n}\right). \quad \square\end{aligned}$$

6.3. Proof of Theorem 1.

We show that any linear combination of the $K_{1,n}(r_t)$ is asymptotically normal. Let $\Lambda = (\lambda_1, \dots \lambda_d)$ be a vector of real coefficients. Define $Z_1 = \sum_{t=1}^{d} \lambda_t K_{1,n}(r_t)$. We use the Bernstein blocks technique (Bernstein, 1939): we divide the square A_n into squares of side p with $p = o(n)$. These squares are separated by gaps of width $2r_d$ so that the sums over couples of points in each square are independent. The couples of points with at least one point in the gaps give a negligible contribution, so that the statistic Z_1 is equivalent to a sum of independent variables and asymptotically normal.
Set $p = n^{1/4}$. Assume that the Euclidean division of n by $(p + 2r_d)$ gives a quotient a and a remainder q. For $l = 0, \dots, a$, we define the segment $I_l = [(p + 2r_d)l, (p + 2r_d)l + p - 1]$. We order the set $\{0, \dots, a\}^2$ by the lexicographic order. To any integer i such that $1 \le i \le k = (a + 1)^2$, corresponds an element (j_1, j_2) of this set; we define the block $P_{i,n} = I_{j_l} \times I_{j_2}$ and $Q = A_n \backslash \cup_i P_{i,n}$ the set of points that are in none of the $P_{i,n}$'s. For each block $P_{i,n}$ and Q, we define the partial sums:

$$\begin{aligned} u_{i,n} &= \frac{1}{n\rho^{3/2}} \sum_{X_l \neq X_m \in P_{i,n}} \sum_{t=1}^{d} \lambda_t \ \mathrm{I}\{d(X_l, X_m) \le r_t\},\\ v_{i,n} &= \frac{1}{n\rho^{3/2}} \sum_{X_l \in P_{i,n}, X_m \in Q} \sum_{t=1}^{d} \lambda_t \ \mathrm{I}\{d(X_l, X_m) \le r_t\}\\ w_n &= \frac{1}{n\rho^{3/2}} \sum_{X_l \neq X_m \in Q} \sum_{t=1}^{d} \lambda_t \ \mathrm{I}\{d(X_l, X_m) \le r_t\}. \end{aligned}$$

then

$$n\sqrt{\rho}(Z_1 - \mathbb{E}Z_1) = \sum_{i=1}^{k}(u_{i,n} - \mathbb{E}u_{i,n}) + \sum_{i=1}^{k}(v_{i,n} - \mathbb{E}v_{i,n}) + w_n - \mathbb{E}w_n,$$

We show that the sum of the $u_{i,n}$ converges in distribution to a Gaussian variable and that the other term are negligible in [2]. We check the conditions of the following CLT adapted from Bardet et al. (2008).

Theorem 2. *Let $(z_{i,n})_{0\leq i\leq k(n)}$ be an array of random variables satisfying*

1. *There exists $\delta > 0$ such that $\sum_{i=0}^{k(n)} \mathbb{E}|z_{i,n}|^{2+\delta}$ tends to 0 as n tends to infinity,*
2. *$\sum_{i=0}^{k(n)} var\, z_{i,n}$ tends to σ^2 as n tends to infinity,*

then $\sum_{i=0}^{k(n)} z_{i,n}$ tends in distribution to $\mathcal{N}(0,\sigma^2)$ as n tends to infinity.

To check Condition 1, we compute the fourth order moment of $u_{i,n} - \mathbb{E}u_{i,n}$. Let N_i be the number of points of S that fall in $P_{i,n}$. Define

$$f(x,y) = \sum_{t=1}^{d} \lambda_t \ (\ \mathrm{I}\{d(x,y) \leq r_t\} - e_{r,p}) = \sum_{t=1}^{d} \lambda_t \ h(x,y,r_t)$$

$$\mathbb{E}((u_{i,n} - \mathbb{E}u_{i,n})^4 | N_i) = \frac{1}{n^4\rho^6}\mathbb{E}\left(\sum_{l\neq m=1}^{N_i} f(U_l, U_m)\right)^4$$

Denote f_1 and f_2 the decomposing functions of f:
$\mathbb{E}(f_1(U_l)) = 0$, $\mathbb{E}(f_1(U_l)f_2(U_l,U_m)) = \mathbb{E}(f_1(U_m)f_2(U_l,U_m)) = 0$, for U_l and U_m two independent uniform variables on $P_{i,n}$.

$$\sum_{l\neq m=1}^{N_i} f(U_l,U_m) = 2(N_i-1)\sum_{l=1}^{N_i} f_1(U_l) + \sum_{l\neq m=1}^{N_i} f_2(U_l,U_m).$$

Note that $|h_1(x,r)| \leq \pi r^2 p^{-2}$ so that f_1 is bounded by Cp^{-2}.
Define $M_1 = \mathbb{E}\left(\sum_{l=1}^{N_i} f_1(U_l)\right)^4$. Then $M_1 = N_i E(f_1^4(U)) + 6N_i(N_i-1)E(f_1^2(U))^2$ and

$$\mathbb{E}(N_i-1)^4 M_1 = O(1).$$

Define $M_2 = \mathbb{E}\left(\sum_{l\neq m=1}^{N_i} f_2(U_l,U_m)\right)^4$. Because f_2 is zero mean with respect to one coordinate, only the products where variables appear at least two times contribute.

$$\begin{aligned}
M_2 \ &= \ 8 \sum_{l\neq m=1}^{N_i} \mathbb{E}f_2^4(U_l,U_m) + 16 \sum_{l\neq m\neq u=1}^{N_i} \mathbb{E}f_2^2(U_l,U_u)f_2^2(U_m,U_u) \\
&+ \ 32 \sum_{l\neq m\neq u=1}^{N_i} \mathbb{E}f_2^2(U_l,U_m)f_2(U_m,U_u)f_2(U_l,U_u) \\
&+ \ 4 \sum_{l\neq m\neq u\neq v=1}^{N_i} \mathbb{E}f_2^2(U_l,U_m)f_2^2(U_u,U_v) \\
&+ \ 16 \sum_{l\neq m\neq u\neq v=1}^{N_i} \mathbb{E}f_2(U_l,U_m)f_2(U_m,U_u)f_2(U_u,U_v)f_2(U_v,U_l).
\end{aligned}$$

Because f_2 is bounded, $\mathbb{E}M_2 = O(\mathbb{E}N_i(N_i-1)(N_i-2)(N_i-3)) = O(p^8)$, so that

$$\sum_{i=0}^{k} \mathbb{E}(u_{i,n} - \mathbb{E}u_{i,n})^4 = O(p^6 n^{-2}).$$

As $p = n^{1/4}$, we get condition 1.
To check condition 2, note that the vector $(K_{1,P_i}(r_1), \dots, K_{1,P_i}(r_d))$ has a covariance matrix Σ_p defined by Proposition 2 by substituting p to n in the expressions. The $u_{i,n} = \frac{p^2\sqrt{\rho}}{n}\sum_{t=1}^{d} \lambda_t(K_{1,P_i}(r_t) - \mathbb{E}K_{1,P_i}(r_t))$ are i.i.d variables with variance equal to $\frac{p^4\rho}{n^2}\Lambda^t\Sigma_p\Lambda$. But $p^2\rho\Sigma_p$ tends to Σ as p tends to infinity and

$$\sum_{i=0}^{k} \operatorname{var} u_{i,n} = \frac{kp^4\rho}{n^2}\Lambda^t\Sigma_p\Lambda \longrightarrow \Lambda^t\Sigma\Lambda$$

so that $\sum_{i=1}^{k} u_{i,n}$ tends in distribution to $\mathcal{N}(0, \Lambda^t\Sigma\Lambda)$.
Note that the $v_{i,n}$ are k independent variables. Denote N_{i,r_d} the number of points X_l in the boundary region P_{i,r_d} of $P_{i,n}$ such that the ball $B(X_l, r_d)$ intersects Q and let $D(X_l)$ denote this intersection. Note that

$$\mathbb{E}N_{i,r_d} = \rho m(P_{i,r_d}) \leq Cpr_d.$$

$$\operatorname{var} v_{i,n} \leq \frac{C}{n^2}\mathbb{E}\left(\sum_{l=1}^{N_{i,r_d}}\sum_{m=1}^{N_Q} \mathrm{I}\{X_m \in D(X_l)\}\right)^2 \leq \frac{C}{n^2}(T_1 + T_2),$$

where

$$\begin{aligned} T_1 &= \mathbb{E}\sum_{l=1}^{N_{i,r_d}}\sum_{m=1}^{N_Q}\sum_{u=1}^{N_Q} \mathrm{I}\{X_m \in D(X_l)\}\,\mathrm{I}\{X_u \in D(X_l)\} \\ T_2 &= \mathbb{E}\sum_{l=1}^{N_{i,r_d}}\sum_{m=1}^{N_{i,r_d}}\sum_{u=1}^{N_Q} \mathrm{I}\{X_u \in D(X_l) \cap D(X_m)\}. \end{aligned}$$

$$\begin{aligned} T_1 &\leq \mathbb{E}N_{i,r_d}\mathbb{E}N_Q^2\mathbb{P}^2\{X_m \in D(X_l) | X_m \in Q\} \\ &\leq \rho^3 m(P_{i,r_d})(m^2(Q) + m(Q))\left(\frac{\pi r_d^2}{2m(Q)}\right)^2 = O(p). \\ T_2 &= \mathbb{E}\sum_{l=1}^{N_{i,r_d}}\sum_{m=1}^{N_{i,r_d}}\sum_{u=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, 2r_d)\}\,\mathrm{I}\{X_u \in D(X_l) \cap D(X_m)\} \\ &\leq \mathbb{E}N_{i,r_d}^2\mathbb{P}\{X_m \in B(X_l, r_d) | X_m \in P_{i,r_d}\}\mathbb{E}N_Q\mathbb{P}\{X_u \in D(X_l) | X_u \in Q\} \\ &\leq \rho^3(m^2(P_{i,r_d}) + m(P_{i,r_d}))\left(\frac{\pi r_d^2}{m(P_{i,r_d})}\right)m(Q)\left(\frac{\pi r_d^2}{2m(Q)}\right) = O(p). \end{aligned}$$

and var $\left(\sum_{i=1}^{k} v_{i,n}\right) = O\left(kp/n^2\right) = O\left(p^{-1}\right)$, so that this sum is negligible in ^{2}. Similarly

$$\text{var}\,(w_n) \quad \leq \quad \frac{C}{n^2}\mathbb{E}\left(\sum_{l\neq m=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, r_d)\}\right)^2 \leq \frac{C}{n^2}(T_1 + T_2).$$

where

$$\begin{aligned}
T_1 &= \mathbb{E}\sum_{l=1}^{N_Q}\sum_{m=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, r_d)\} \\
&\leq \mathbb{E}N_Q(N_Q-1)\mathbb{P}\{X_m \in B(X_l, r_d)|X_m \in Q\} \leq m^2(Q)\frac{\pi r_d^2}{m(Q)}. \\
T_2 &= \mathbb{E}\sum_{l=1}^{N_Q}\sum_{m=1}^{N_Q}\sum_{u=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, r_d)\}\,\mathrm{I}\{X_u \in B(X_l, r_d)\} \\
&\leq \mathbb{E}N_Q^2(N_Q-1)\mathbb{P}^2\{X_m \in B(X_l, r_d)|X_m \in Q\} \\
&\leq (m^3(Q) + 2m^2(Q))\left(\frac{\pi r_d^2}{m(Q)}\right)^2.
\end{aligned}$$

Then var $(w_n) = O\left(m(Q)/n^2\right) = O\left(p^{-1}\right)$ and w_n is negligible in ^{2}.
Consider now $K_{2,n}(r)$. Define $Z_2 = \sum_{t=1}^{d} \lambda_t K_{2,n}(r_t) = A_{N,n}Z_1$ where $A_{N,n} = \frac{n^4\rho^2}{N(N-1)}$. We have $\mathbb{E}(A_{N,n}^{-1}) = 1$ and from (5), var $(A_{N,n}^{-1}) = \dfrac{4}{n^2\rho} + \dfrac{2}{n^4\rho^2}$.
For $\delta > 0$, the Markov inequality gives

$$\mathbb{P}(|A_{N,n}^{-1} - 1| > \delta) \leq \frac{\text{var}\,(A_{N,n}^{-1})}{\delta^2}.$$

Then, with $\delta = n^{-1/4}$

$$\sum_{n=1}^{\infty} \mathbb{P}(|A_{N,n}^{-1} - 1| > n^{-1/4}) < \sum_{n=1}^{\infty} \frac{4}{n^{3/2}\rho} + \frac{2}{n^{7/2}\rho^2} < \infty.$$

From the Borel-Cantelli lemma, we get that $A_{N,n}^{-1}$ converges a.s. to 1. By the Slutsky lemma, $A_{N,n}Z_1$ converges in distribution to $\mathcal{N}(0, \Lambda^t\Sigma\Lambda)$. □

6.4. Computation of $\mathbb{E}h_1^2(U, r)$

Lemma 4.

$$\mathbb{E}h_1^2(U, r) \quad = \quad \frac{r^5}{n^5}\left(\frac{8}{3}\pi - \frac{256}{45}\right) + \frac{r^6}{n^6}\left(\frac{11}{48}\pi - \frac{56}{9}\right) + \frac{8}{3}\frac{r^7}{n^7} - \frac{1}{4}\frac{r^8}{n^8}.$$

Proof: From the computation of the bias, denoting $x_i = \frac{1}{r}(n - \xi_i)$, we get

$$\begin{aligned} h_1(\xi, r) &= \frac{\pi r^2}{n^2} - e_{r,n} \text{ on } A_n^1 \\ &= \frac{r^2}{n^2}(\pi - g(x_1)) - e_{r,n} \text{ on } A_n^2 \\ &= \frac{r^2}{n^2}(\pi - g(x_1) - g(x_2)) - e_{r,n} \text{ on } A_n^3 \\ &= \frac{r^2}{n^2}\left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2}\right) - e_{r,n} \text{ on } A_n^4 \end{aligned}$$

$$\mathbb{E}(h_1(X_j, r))^2 = \pi^2 \left(1 - \frac{2r}{n}\right)^2 \frac{r^4}{n^4} - e_{r,n}^2 + 4\left(1 - \frac{2r}{n}\right)\frac{r^5}{n^5} T_1 + 4\frac{r^6}{n^6}(T_2 + T_3)$$

$$T_1 = \int_0^1 (\pi - g(x_1))^2 dx_1 \tag{10}$$

$$T_2 = \int_0^1 dx_1 \int_{\sqrt{1-x_1^2}}^1 (\pi - g(x_1) - g(x_2))^2 dx_2 \tag{11}$$

$$T_3 = \int_0^1 dx_1 \int_0^{\sqrt{1-x_1^2}} \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2}\right)^2 dx_2. \tag{12}$$

To compute these three terms, we need integral computations on function g.

Lemma 5. *For $n \geq 1$,*

$$I_n = \int_0^1 u^{2n-1} \arccos(u) du = \frac{\pi(2n)!}{n 2^{2n+2}(n!)^2}.$$

$$J_n = \int_0^1 u^{2n}\sqrt{1-u^2} du = -(2n+2)I_{n+1} + 2nI_n.$$

$$\int_0^1 \sqrt{1-u^2}\arccos(u) du = \frac{\pi^2}{16} + \frac{1}{4}. \tag{13}$$

$$\int_0^1 \sqrt{1-u^2}\arccos^2(u) du = \frac{\pi^3}{48} + \frac{\pi}{4}. \tag{14}$$

Note: in the following, we use $I_1 = \pi/8$, $I_2 = 3\pi/64$, $J_1 = \pi/16$ and $J_2 = \pi/32$.

Lemma 6.

$$\int_0^1 g(u)\sqrt{1-u^2} du = \frac{\pi^2}{16}. \tag{15}$$

$$\int_0^1 g^2(u) du = \frac{2\pi}{3} - \frac{64}{45}. \tag{16}$$

$$\int_0^1 g^2(u)\sqrt{1-u^2} du = \frac{\pi^3}{48}. \tag{17}$$

$$\int_0^1 g(u)\ G\left(\sqrt{1-u^2}\right) du = \frac{\pi^3}{96} - \frac{5\pi}{48} + \frac{4}{9}. \tag{18}$$

Proofs are postponed in the appendix. Using these lemmas, we get

$$T_1 = \pi^2 - 2\pi G(1) + \int_0^1 g^2(x_1)dx_1 = \pi^2 - \frac{64}{45} - \frac{2\pi}{3}. \tag{19}$$

$$\begin{aligned} T_2 &= \pi^2\left(1-\frac{\pi}{4}\right) - 4\pi\int_0^1 g(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 dx_2 + 2\int_0^1 g^2(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 dx_2 \\ &\quad +2\int_0^1 g(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 g(x_2)dx_2. \end{aligned}$$

From the computation of the bias, $-4\pi\int_0^1 g(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 dx_2 = -\frac{8\pi}{3} + \frac{\pi^3}{4}$.

From (16), (17) and (18), we get

$$2\int_0^1 g^2(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 dx_2 = 2\int_0^1 g^2(x_1)dx_1 - 2\int_0^1 \sqrt{1-x_1^2}g^2(x_1)dx_1 = \frac{4\pi}{3} - \frac{128}{45} - \frac{\pi^3}{24}.$$

$$2\int_0^1 g(x_1)dx_1\int_{\sqrt{1-x_1^2}}^1 g(x_2)dx_2 = 2G^2(1) - 2\int_0^1 g(x_1)G\left(\sqrt{1-x_1^2}\right)dx_1 = -\frac{\pi^3}{48} + \frac{5\pi}{24}.$$

Adding these results, we obtain

$$T_2 = -\frac{\pi^3}{16} + \pi^2 - \frac{9\pi}{8} - \frac{128}{45}. \tag{20}$$

To compute T_3, we write

$$\begin{aligned} T_3 &= \frac{9\pi^3}{64} + \int_0^1 x_1^2 dx_1\int_0^{\sqrt{1-x_1^2}} x_2^2 dx_2 - \frac{3\pi}{2}\int_0^1 g(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} dx_2 \\ &+ \frac{1}{2}\int_0^1 g^2(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} dx_2 + \frac{3\pi}{2}\int_0^1 x_1 dx_1\int_0^{\sqrt{1-x_1^2}} x_2 dx_2 \\ &+ \frac{1}{2}\int_0^1 g(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} g(x_2)dx_2 - 2\int_0^1 x_1 g(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} x_2 dx_2. \end{aligned}$$

$$\int_0^1 x_1^2 dx_1\int_0^{\sqrt{1-x_1^2}} x_2^2 dx_2 = \frac{1}{3}\int_0^1 x_1^2(1-x_1^2)\sqrt{1-x_1^2}dx_1 = \frac{1}{3}(J_1 - J_2) = \frac{\pi}{96}.$$

From (15), $-\frac{3\pi}{2}\int_0^1 g(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} dx_2 = -\frac{3\pi^3}{32}$.

From (17), $\frac{1}{2}\int_0^1 g^2(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} dx_2 = -\frac{\pi^3}{96}$.

$\frac{3\pi}{2}\int_0^1 x_1 dx_1\int_0^{\sqrt{1-x_1^2}} x_2 dx_2 = \frac{3\pi}{4}\int_0^1 x_1(1-x_1^2)dx_1 = \frac{3\pi}{16}$.

From (18), $\frac{1}{2}\int_0^1 g(x_1)dx_1\int_0^{\sqrt{1-x_1^2}} g(x_2)dx_2 = \frac{\pi^3}{192} - \frac{5\pi}{96} + \frac{2}{9}$.

$$-2\int_0^1 x_1 g(x_1)dx_1 \int_0^{\sqrt{1-x_1^2}} x_2 dx_2 = \int_0^1 \left(x_1^3 - x_1\right) g(x_1)dx_1 = -\frac{3\pi}{64}.$$

Adding these results, we get

$$T_3 = \frac{\pi^3}{16} + \frac{19\pi}{192} + \frac{2}{9}. \tag{21}$$

Gathering (19), (20) and (21) gives the result. □

References

Bardet, J-M., Doukhan, P., Lang, G. & Ragache, N. (2008). Dependent Lindeberg central limit theorem and some applications, *ESAIM Probab. Stat.*, **12** , 154-172.

Bernstein, S. (1939). Quelques remarques sur le théorème limite Liapounoff. *C. R. (Dokl.) Acad. Sci. URSS*, **24**, 3-8.

Besag, J. E. (1977). Comments on Ripley's paper. *J. Roy. Statist. Soc. Ser. B*, **39** (2), 193-195.

Chiu, S. N. (2007). Correction to Koen's critical values in testing spatial randomness. *J. Stat. Comput. Simul.* 77(11-12), 1001-1004.

Cressie, N. A. (1993). *Statistics for spatial data.* John Wiley & Sons, New York. 900 p.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns.* Academic Press, London. 148 p.

Duranton, G. & Overman, H. G. (2005). Testing for localisation using micro-geographic data. *Rev. Econom. Stud.*, **72** (4), 1077-1106.

Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns.* Wiley-Interscience, Chichester.

Koen, C., (1991). Approximate confidence bounds for Ripley's statistic for random points in a square. *Biom. J.*,**33**, 173-177.

Marcon, E. & Puech, F. (2003). Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economical Geography*, **3** (4), 409-428.

Møller, J. & Waagepetersen, R. P. (2004). Statistical inference and simulation for spatial point processes. *Monographs on statistics and applied probability*, **100**, Chapman & Hall/CRC, Boca Raton, 300 p.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *J. Appl. Probab.* **13** , 255-266.

Ripley, B. D. (1977). Modelling spatial patterns. *J. Roy. Statist. Soc. Ser. B*, **39** (2), 172-212.

Ripley, B. D. (1979). Tests of randomness for spatial point patterns. *J. Roy. Statist. Soc. Ser. B*, **41** (3), 368-374.

Ripley, B. D. (1981). *Spatial statistics*, John Wiley & Sons, New York. 255 p.

Saunders, R. & Funk, G. M. (1977). Poisson limits for a clustering model of Strauss. *J. Appl. Probab.*, **14**, 776-784.

Stoyan, D., Kendall, W. S. & Mecke, J. (1987) *Stochastic geometry and its applications.* John Wiley & Sons, New York. 345 p.

Stoyan, D. & Stoyan, H. (2000). Improving ratio estimators of second order point process characteristics. *Scand. J. Statist.* **27**, 4, 641-656.
Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika* **36**, 18-25.
Ward, J. S. & Ferrandino, F. J. (1999). New derivation reduces bias and increases power of Ripley's L index. *Ecological Modelling*, **116** (2-3), 225-236.

Appendix A: Integration lemmas

A.1. Proof of Lemma 5

Integrating by parts

$$\int_0^1 u^{2n-1}\arccos(u)du = \int_0^{\pi/2} t\cos^{2n-1}(t)\sin(t)dt = \frac{1}{2n}\int_0^{\pi/2}\cos^{2n}(t)dt.$$

Using De Moivre formula

$$\cos^{2n}(t) = \frac{1}{2^{2n}}\left(2\cos(2nt) + 2\begin{pmatrix}2n\\1\end{pmatrix}\cos(2(n-1)t) + \cdots + \begin{pmatrix}2n\\n\end{pmatrix}\right).$$

Only the last term gives a non zero integral, giving the result for I_n.

$$\begin{aligned} J_n &= \int_0^1 (u^{2n+2}-u^{2n})(-(1-u^2)^{-1/2})du \\ &= \left[(u^{2n+2}-u^{2n})\arccos(u)\right]_0^1 - \int_0^1 ((n+2)u^{2n+1} - nu^{2n-1})\arccos(u)du \end{aligned}$$

and the term under brackets is zero, giving the result.

$$\begin{aligned} \int_0^1 \sqrt{1-u^2}\arccos(u)du &= \int_0^{\pi/2} t\sin^2(t)dt = \int_0^{\pi/2}\frac{t}{2} - \frac{t\cos(2t)}{2}dt \\ &= \frac{\pi^2}{16} - \left[\frac{t\sin(2t)}{4}\right]_0^{\pi/2} + \int_0^{\pi/2}\frac{\sin(2t)}{4}dt = \frac{\pi^2}{16}+\frac{1}{4}. \\ \int_0^1 \sqrt{1-u^2}\arccos^2(u)du &= \int_0^{\pi/2} t^2\sin^2(t)dt = \int_0^{\pi/2}\frac{t^2}{2} - \frac{t^2\cos(2t)}{2}dt \\ &= \frac{\pi^3}{48} - \left[\frac{t^2\sin(2t)}{4}\right]_0^{\pi/2} + \int_0^{\pi/2}\frac{t\sin(2t)}{2}dt \\ &= \frac{\pi^3}{48} - \left[\frac{t\cos(2t)}{4}\right]_0^{\pi/2} + \int_0^{\pi/2}\frac{\cos(2t)}{4}dt = \frac{\pi^3}{48}+\frac{\pi}{8}. \quad \square \end{aligned}$$

A.2. Proof of lemma 6

Equation (15) follows from equation (13).
Write $g^2(u) = \arccos^2(u) + u^2 - u^4 - 2u\sqrt{1-u^2}\arccos(u)$ and

$$\begin{aligned}\int_0^1 \arccos^2(u)du &= \int_0^{\pi/2} t^2\sin(t)dt = -\left[t^2\cos(t)\right]_0^{\pi/2} + 2\int_0^{\pi/2} t\cos(t)dt \\ &= 2\left[t\sin(t)\right]_0^{\pi/2} + 2\int_0^{\pi/2}\sin(t)dt = \pi - 2, \\ \int_0^1 (u^2-u^4)du &= \frac{1}{3} - \frac{1}{5} = \frac{2}{15}.\end{aligned}$$

$$\begin{aligned}\int_0^1 u\sqrt{1-u^2}\arccos(u)du &= \int_0^{\pi/2} t\cos(t)\sin^2(t)dt \\ &= \left[\frac{t}{3}\sin^3(t)\right]_0^{\pi/2} - \frac{1}{3}\int_0^{\pi/2}\sin^3(t)dt \\ &= \frac{\pi}{6} - \frac{1}{3}\int_0^{\pi/2}\sin(t)dt + \frac{1}{3}\int_0^{\pi/2}\cos^2(t)\sin(t)dt \\ &= \frac{\pi}{6} - \frac{1}{3} - \frac{1}{9}\left[\cos^3(t)\right]_0^{\pi/2} = \frac{\pi}{6} - \frac{2}{9}.\end{aligned}$$

Collecting the three parts yields to (16).

$$\begin{aligned}\int_0^1 g^2(u)\sqrt{1-u^2}du &= \int_0^1 \sqrt{1-u^2}\arccos^2(u)du \\ &\quad -2\int_0^1 (u-u^3)\arccos(u)du + \int_0^1 \sqrt{1-u^2}(u^2-u^4)du \\ &= \frac{\pi^3}{48} + \frac{\pi}{8} - 2\left(\frac{\pi}{8} - \frac{3\pi}{64}\right) + \frac{\pi}{16} - \frac{\pi}{32} = \frac{\pi^3}{48}.\end{aligned}$$

Write $G\left(\sqrt{1-x^2}\right) = \sqrt{1-x^2}\left(\frac{\pi}{2} - \arccos(x)\right) + \frac{x^3}{3} - x + \frac{2}{3}$

$$\begin{aligned}\int_0^1 g(x)G\left(\sqrt{1-x^2}\right)dx &= \int_0^1 \sqrt{1-x^2}\left(\frac{\pi}{2} - \arccos(x)\right)\arccos(x)dx \\ &\quad - \int_0^1 (x-x^3)\left(\frac{\pi}{2} - \arccos(x)\right)dx \\ &\quad + \int_0^1 \left(\frac{x^3}{3} - x + \frac{2}{3}\right)\arccos(x)dx \\ &\quad + \int_0^1 \left(-\frac{x^4}{3} + x^2 - \frac{2x}{3}\right)\sqrt{1-x^2}dx \\ &= \frac{\pi^3}{96} - \frac{5\pi}{48} + \frac{4}{9}. \quad \square\end{aligned}$$

Journal of Economic Geography (2009) pp. 1–18 doi:10.1093/jeg/lbp056

Measures of the geographic concentration of industries: improving distance-based methods

Eric Marcon and Florence Puech*,†*

Abstract

We discuss a property of distance-based measures that has not been addressed with regard to evaluating the geographic concentration of economic activities. The article focuses on the choice between a probability density function of point-pair distances or a cumulative function. We begin by introducing a new cumulative function, *M*, for evaluating the relative geographic concentration and the co-location of industries in a non-homogeneous spatial framework. Secondly, some rigorous comparisons are made with the leading probability density function of Duranton and Overman (2005), *Kd*. The merits of the simultaneous use of *Kd* and *M* is proved, underlining the complementary nature of the results they provide.

Keywords: Geographic concentration, distance-based methods, *K*-density function, Ripley's *K* function, *M* function
JEL classifications: C40, C60, R12, L60
Date submitted: 8 February 2008 **Date accepted:** 28 September 2009

1. Introduction

> Step back and ask, what is the most striking feature of the geography of economic activity? The short answer is surely *concentration*. Krugman (1991, 5)

As highlighted by Paul Krugman (1991) in the first pages of his book *Geography and Trade*, economic activities are definitely not homogeneously distributed. During the last decade, the appraisal of the degree of spatial concentration of economic activities has received an increasing amount of attention. Economists have improved the measurement of the geographic concentration of economic activities in several ways and various criteria have been suggested for a 'good concentration index' (Combes and Overman, 2004). Up to now, the measure that respects the largest number of these properties is the *K-density* function (denoted *Kd*) proposed by Duranton and Overman (2005). The *Kd* function (i) compares the geographic concentration results across industries, (ii) controls for industrial concentration, (iii) controls for the overall aggregation patterns of industries, (iv) tests the significance of the results and (v) keeps the empirical results unbiased across geographic scales.

The aim of this article is to discuss an additional important property of distance-based measures that has not yet been addressed in the economic literature.

*AgroParisTech ENGREF, UMR EcoFoG, BP 316, 97310 Kourou, French Guyana.
†Corresponding author: Florence Puech, LET (Université de Lyon, CNRS, ENTPE), Institut des Sciences de l'Homme, 14 av. Berthelot, 69363 Lyon Cedex 07, France. *email* <Florence.Puech@univ-lyon2.fr>

This is the choice between using a probability density function of point-pair distances or a cumulative function for evaluating geographic concentration. Surprisingly, the preference for one or the other has never been discussed in empirical economic articles.[1] The only exception is Duranton and Overman (2005) who claimed in the conclusion of their article that probability density functions reveal more information than cumulative functions. In this article, we shall demonstrate clearly that the two types of functions cannot be considered as substitutes for each other. To do this, we shall propose a new function, that we shall call the *M* function, that, like the *Kd* function, respects the five criteria listed previously. However, while *Kd* is a probability density function of point-pair distances because it is calculated on the basis of the average number of neighbours at each distance, the *M* function is cumulative, depending on the number of neighbours up to each distance.

We have reached two main conclusions. Firstly, both probability density functions and cumulative functions are viable approaches for analysing the geographic concentration of activities. Consequently, a switch to probability density functions is not compulsory in order to meet Duranton and Overman's criteria. Secondly, the simultaneous use of *Kd* and *M* is recommended because they provide complementary results concerning the distribution of economic activities.

Our study is organized as follows. Section 2 presents an overview of cumulative distance-based measures. The section after this introduces the *M* function and discusses its mathematical properties. Next comes a comparison with other similar distance-based methods in order to demonstrate the value of these two new measures (Section 4). The last section concludes.

2. An overview of distance-based measures

In order to evaluate the geographic distribution of establishments, economists have traditionally employed cluster-based methods, which measure the spatial concentration of economic activity according to pre-defined geographic limits (regions, counties ...). It is now widely accepted that the measures obtained with these methods, such as the Gini and the Ellison and Glaeser (EG) indices,[2] introduce a statistical bias resulting from the chosen concept of space. Cluster-based methods zone the area in question: dividing space into a set of geographical units raises the well-known Modifiable Areal Unit Problem (MAUP), which can be summarized as follows: 'the result will be sensitive to the shape, size, and position of the areal units chosen' (Morphet 1997, 1039). The use of cluster-based methods is therefore problematic as they violate property (v).[3] The solution to this problem is to use a continuous approach to space, thus switching from cluster-based methods to distance-based methods.

Distance-based measures are a relatively new way of gauging the geographic concentration of activities (Arbia and Espa, 1996; Marcon and Puech, 2003; Duranton and Overman 2005, 2008). The idea of these functions is simple. Unlike cluster-based

1 See for instance Fratesi (2008) for an empirical application of the *Kd* function or Arbia and Espa (1996), Barff (1987), Ó hUallacháin and Leslie (2007), Arbia et al. (2008), Jensen (2006) for different applications of cumulative distance-based methods.

2 For instance, see Combes et al. (2008) for a review of the main spatial concentration indices.

3 An empirical estimation of the shape and size bias resulting from different French area zonings may be found in Briant et al. (forthcomomg).

methods, distance-based methods do not zone the area in question in a specific manner but consider continuous space. Basically, each plant in the sample is localized by its coordinates (x,y) and the Euclidean distances between plants are considered. Unlike measures that only describe the location of economic activity on a single scale, distance-based methods can detect spatial structure at every scale. The main advantage of these methods is that they detect the distance at which significant geographic concentration or dispersion of establishments occurs. Two types of distance-based methods currently coexist in the economic literature:

(i) The probability density function of point-pair distances that is based on the average number of neighbours at each distance (r). This measure is then smoothed and normalized so that it sums up to 1.
(ii) Cumulative distance-based methods that describe geographic concentration by counting the average number of neighbours of plants on a disc, i.e. 'within' a circle of a given radius (r). This operation is then repeated for all possible radii.

The *Kd* function of Duranton and Overman (2005) is the only probability density function used in economic geography. Before their work, cumulative functions were used systematically in studies. Ripley's K function (1976, 1977), Besag's L function (1977) and their extensions based on the second-order property of point patterns are used in empirical applications.[4] However, while Ripley's functions are now widely applied in other scientific fields such as forestry and ecology,[5] Marcon and Puech (2003) have pointed out that they do not respect Duranton and Overman's five criteria (2005) for wide application in economics. Firstly, Ripley's function measures absolute concentration: it is based on the null hypothesis of a completely random spatial distribution of establishments (i.e. plants are distributed uniformly and independently). In spite of the longstanding debate about implementing absolute or relative measures (see Haaland et al., 1999), relative measures are still more widely used. Comparing a sector distribution to that of the whole of industry is even one of the theoretical criteria—property (iii)—defined by Duranton and Overman (2005). Relative measures detect whether each industry is overrepresented or underrepresented with respect to a baseline distribution, for example, the overall location pattern of industries. In other words, statistical tools based on relative concentration effectively measure the existence of specialized areas.[6] Secondly, Ripley's function does not control for industrial concentration, i.e. the productive concentration within an industry among plants belonging to the sector in question—property (ii)—as every establishment is considered to be a point on the plot, regardless of its size.

These two problematic points must be answered to permit direct comparisons between a cumulative function and *Kd*. In the next section, we propose two versions of a new statistical tool, the M function, for the measurement of intra- and inter-industry geographic concentration. We have called it the M function because it is an extension of

4 For instance see Barff (1987), Arbia (1989), Ó hUallacháin and Leslie (2007) and Arbia et al. (2008) for different applications of cumulative distance-based methods to describe the geographic concentration of industries.
5 A survey of empirical studies in ecology or forestry using the K or L function can be found in Puech (2003, 324).
6 A discussion on the limits of relative indices can be found in Appendix A of Mori et al. (2005).

the existing cumulative distance-based methods, namely Ripley's K function (1976, 1977) and Besag's L function (1977).

3. Improving Ripley's functions: an introduction to the *M function*

In what follows, we shall first give an intuitive presentation of the common framework. We shall then successively define the M function and discuss its properties.

3.1. An intuitive analysis

Our relative measure compares the location patterns of an economic sector to that of aggregate activity (represented by all sectors). For this, we develop a cumulative function that counts neighbouring points up to a chosen distance denoted r. Let us consider a map with points on it that represent plants. We choose:

(i) a reference point type, say a specific sector, and
(ii) a target neighbour type called T: the same sector for intra-industry concentration or another sector for inter-industry concentration.

The average number of target neighbours is compared to a benchmark to detect whether they are more or less frequent than if plants were distributed randomly and independently from each other. To control for variations in the local density of points, each number of target neighbours (T_i around a point i) is normalized by the number of all the neighbours in the same area (N_i). Around each reference point we obtain a ratio of target neighbours (T_i/N_i) within the distance r from each point i. The average of this ratio ($\overline{T_i/N_i}$) is compared to the global ratio of the target type (T/N) calculated for the entire area. If $\overline{T_i/N_i}$ is greater than T/N, we conclude that more plants of the target type are observed within a distance r around the reference points type than on average, if circles of radius r were drawn anywhere. In other words, target points are concentrated around reference points. The ratio $M = (\overline{T_i/N_i})/(T/N)$ will be used in our analysis for convenience because the benchmark is equal to one. M values are computed over a large range of distances and presented as a continuous function of r on a graph including confidence intervals for the null hypothesis of independence of plant locations (significance is checked by appropriate statistical tests). As a result of these successive normalizations any value of M can be interpreted immediately and compared across sectors and distances. Finally, points can also be weighted, counting, for example, the number of employees instead of the number of plants. Finally, significance tests must properly control for the non-independence of their distribution (i.e. industrial concentration).

3.2 Evaluating geographic intra-industry concentration

In mathematical terms, let us consider an area A containing a total of N plants belonging to a variety of industries. We shall focus on a particular industry S where N_S is the total number of establishments from that sector in area A. The description of the neighbourhood of the N_S plants follows. Consider a dummy variable $c_S(i,j,r)$ that is equal to 1 if the Euclidean distance between the two plants i and j from the sector S is less than the radius r ($c_S(i,j,r) = 0$ otherwise). The number of neighbouring establishments of plant i, belonging to the same sector and located within a distance

r from it, is thus $\sum_{j=1,i\neq j}^{N_S} c_S(i,j,r)$. In the same way, we define the dummy $c(i,j,r)$ as equal to 1 if plant j (whatever its industry) is located at a distance inferior or equal to r from the establishment i (the dummy's value is 0 otherwise). Consequently, the number of establishments located at most at a distance r from business unit i is: $\sum_{j=1,i\neq j}^{N} c(i,j,r)$. Plant size may now be included in our analysis. The weight associated with each dummy is that of the neighbouring plant j, and it is denoted w_j. The weight associated with the plants may, for example, be their number of employees. The average proportion of employees of industry S within a given radius r is clearly:

$$\frac{1}{N_S}\sum_{i=1}^{N_S}\frac{\sum_{j=1,i\neq j}^{N_S} c_S(i,j,r)w_j}{\sum_{j=1,i\neq j}^{N} c(i,j,r)w_j}$$

In the same way, we can define the ratio of employees in industry S in the entire area A compared to the whole of industry by: $\frac{1}{N_S}\sum_{i=1}^{N_S}\frac{W_S-w_i}{W-w_i}$ where W_S is the total number of industry S employees in the area A; W is the total number of employees in aggregate activity and w_i is the weight of plant i.[7] The ratio of the above quantities, averaged over all the establishments in sector S, defines the M function for the intra-industry geographic concentration of sector S as:

$$M_S(r) = \sum_{i=1}^{N_S}\frac{\sum_{j=1,i\neq j}^{N_S} c_S(i,j,r)w_j}{\sum_{j=1,i\neq j}^{N} c(i,j,r)w_j}\Bigg/\sum_{i=1}^{N_S}\frac{W_S-w_i}{W-w_i} \tag{1}$$

The numerator corresponds to the relative weight of sector S in comparison with the whole industrial within circles of radius r. The denominator represents the relative size of the considered sector in comparison with all activities in area A. The benchmark for the M function is 1 and this defines the same location pattern for the specific sector as for aggregate activity. This means that whatever the considered radius, there are proportionally as many employees who belong to sector S as there are in the whole of area A. Thus, M values superior to $1(M_S(r) > 1)$ indicate that there are proportionally more employees close to plants in sector S (within a distance r) than in the whole area. This corresponds to the definition of the relative geographic concentration of sector S at distance r. In contrast, the relative geographic dispersion of sector S at distance r is defined by $M_S(r) < 1$, indicating that there are relatively fewer employees in sector S within a distance r around the establishments than in the whole area. One can see that interpreting the M values is straightforward. For instance, $M_S(r) = 2$ indicates that, within a particular distance r, the relative density of employees in sector S is double that in the whole area. In the same way, $M_S(r) = 0.5$ indicates that within a given distance r around sector S plants the density of employees in this sector is on average half that of the whole area.

We shall now consider how the significance of the results can be tested and how we can control for the industrial concentration. Two types of confidence intervals for the null hypothesis are generated: local and global. The null hypothesis is that establishments belonging to sector S are located according to the same pattern as the others. To test this, we generate a series of random and independent distributions of the plant

7 w_i must be subtracted from W_s and W since we count the number of establishments around the plant i within a radius r: the reference establishment i itself should not be counted.

dataset based on the actual set of possible locations and the industry/plant size pairs (i.e. the industrial concentration as given). The local confidence interval is determined using the Monte–Carlo method. In practice, we generate a large number of simulations and choose a confidence level, say 5%. The 95% confidence interval of the M function for each value of r is bounded by the outer 5% of the randomly generated values. Nearly all empirical studies that use Ripley's K function (or one of its extensions) compute only the local confidence intervals to test the significance of the results. However, Duranton and Overman (2005) have recently criticized the computation of local confidence intervals on their own, considering them as too 'optimistic', and highlighted the need for the global confidence intervals of the null hypothesis as well. If we assume that the values of the M function at different radii are independently distributed, one would expect a proportion of them equal to the confidence threshold to be outside the confidence interval even though the point process corresponds to the null hypothesis. For instance at a 5% threshold, complete spatial randomness should not be rejected when 5 points in a 100-point plot are outside the confidence interval. Successive values of Ripley's functions are actually highly correlated: the risk of erroneous rejection of the null hypothesis is consequently reduced but cannot be quantified. A global confidence interval is defined such that the confidence threshold is the risk that the plot of a function generated by the null hypothesis exceeds the interval at least once. This may be chosen in many ways but it should have an equal weight at all distances. A simple method of computing global confidence intervals is to generate local confidence intervals at increasing confidence levels until the ratio of simulated plots that lie partly outside them reaches the predefined threshold. As an example, let us suppose that 1000 plots have been generated and the confidence level has been set at 5%. The outer values at each distance are eliminated, defining a local confidence interval at $2/1000 = 0.2\%$. The plots that lie partly outside this interval are counted. If there are 10 such plots, the global confidence level will be 1%. The process is repeated until the threshold is reached. If it is not reached exactly, interpolation is performed.

Five fundamental criteria characterizing a 'good' economic measure of geographic concentration were presented in the introduction. The M function respects them all. Moreover, one can note that an appreciable property of this index is that its values may be interpreted. Additionally, the M function can be calculated for any topology. Ripley's function and its developments (see Goreaud and Pélissier, 1999) require edge-effect correction for points that are close to borders. Complex geographical shapes are consequently intractable hence the domain must always be a polygon or a disc.[8] The M function provides an answer to this problem: comparing the number of neighbours in a certain industry to the total number of neighbouring establishments 'in the same area' avoids any need for correction. Working on complex geographical limits, such as national borders, is now possible. This last consideration justifies the somewhat complex computation of the M function. Software can be downloaded from the authors' website[9] to facilitate its implementation.

8 Sweeney and Feser (1998, 52) Figure 1, or Feser and Sweeney (2000, 361) Figure 2; Pancer-Koteja et al. (1998, 757) Figure 1; Rowlingson and Diggle (1993, 634) Figure 5.

9 Available online at: http://e.marcon.free.fr/Ripley (English and French versions).

3.3 Evaluating the co-location of industries

If the researcher suspects interactions between them it may be interesting to evaluate the co-location of different industries. The geographic concentration of different industries can be investigated using the inter-industry version of the M function that has the same properties as the intra-industry version. In what follows, we shall consider co-location between two sectors denoted S_1 and S_2. A complete description of the spatial distribution of the co-location patterns of these industries leads not to one but to two definitions of the M function. The first, M_{S_1,S_2}, depicts the spatial distribution of plants belonging to sector S_2 around those of sector S_1 in non-homogenous space. The second function, M_{S_2,S_1}, describes the spatial distribution of plants belonging to sector S_1 around those of sector S_2. The meaning of the co-location M functions is thus simple: they test whether the relative density of employees from one sector around establishments of another sector is on average greater or lesser than in the whole area.

Let us consider the same area A using the same notations as in the previous section. We shall now examine the Euclidean distances between plants belonging to two different industries. First, we shall consider the definition of M_{S_1,S_2} in which the reference plants (i.e. those at the centre of the circles) are from industry S_1. The definition of the $M_{S_1,S_2}(r)$ co-location function is thus:

$$M_{S_1,S_2}(r) = \sum_{i=1}^{N_{S_1}} \frac{\sum_{j=1}^{N_{S_2}} c_{S_2}(i,j,r)w_j}{\sum_{n=1,i\neq n}^{N} c(i,n,r)w_n} \Big/ \sum_{i=1}^{N_{S_1}} \frac{W_{S_2}}{W - w_i} \tag{2}$$

The value of Equation (2) shows whether the relative density of plants S_2 located around those of sector S_1 is greater ($M_{S_1,S_2}(r) > 1$) or lesser ($M_{S_1,S_2}(r) < 1$) than in the entire area A. In the same manner, we can define the function $M_{S_2,S_1}(r)$ that describes the spatial structure of the S_1 plants located around those of sector S_2. The alterations to the function are obvious:

$$M_{S_2,S_1}(r) = \sum_{i=1}^{N_{S_2}} \frac{\sum_{j=1}^{N_{S_1}} c_{S_1}(i,j,r)w_j}{\sum_{n=1,i\neq n}^{N} c(i,n,r)w_n} \Big/ \sum_{i=1}^{N_{S_2}} \frac{W_{S_1}}{W - w_i} \tag{3}$$

Concerning the significance of the results, both the local and global confidence intervals for the null hypothesis are computed but we shall pay particular attention to the null hypothesis. Monte–Carlo techniques are used to generate simulated distributions (the threshold and the number of simulations are exogenous). Nevertheless, the null hypothesis has to eliminate the sector-specific patterns in order to detect only interactions between the two industries. For instance, if S_1 is highly aggregated and S_2 completely randomly distributed, the relative importance number of S_2 plants around S_1 establishments is low and artificial segregation is detected. Under these conditions, the null hypothesis must control for the patterns of both S_1 and S_2. The solution is as follows. The null-hypothesis set of plants for M_{S_1,S_2} is generated by keeping the S_1 establishments fixed and redistributing all the other plant size/sector pairs between all the other locations, thus controlling for the pattern of S_1. To be sure that the structure of the industry S_2 is not responsible for any under- or over-estimation of the density of the plant's employees, we also need

to control for the structure of S_2: the same process applied to M_{S_2,S_1} controls for the pattern of S_2 plants. Lastly, it is accepted that there is a significant interaction if both values are significantly different from their respective null hypothesis. Note that the null hypothesis excludes the detection of a 'multi-concentration' phenomenon: a situation in which we could observe a significant co-location that does not result from an interaction between these two industries (this would be the case for instance if both industries locate around the plants of another industry). This is undoubtedly a limitation of the inter-industry M function and is shared by all other distance-based functions.[10]

4. Towards an unified framework for distance-based methods

At this stage, we can compare the statistical properties of distance-based methods and especially those of the two leading geographic concentration measures namely Kd and M.[11] Our aim is to reveal in which cases one measure should be preferred to the other. As underlined previously, our discussion focuses on the implications of using probability density functions rather than cumulative functions because this remains the main difference between Kd and M.

4.1 Common statistical framework

Historically, methods of characterizing the structure of point processes as a function of bilateral distances between pairs of points have been developed by Ripley (1976, 1977).

Ripley defined the function $g(r)$ as the ratio of the probabilities of finding two points at a distance r from each other to the product of the probabilities of finding each of them. If points are distributed independently, $g(r) = 1$; higher values show that point pairs at this distance are more frequent than under the null hypothesis of independence. The integral function $K(r) = \int_{\rho=0}^{r} g(\rho)2\pi\rho \, \mathrm{d}\rho$ is easy to estimate.[12] Assuming the point density is uniform on an area A and denoting by N the total number of points on the domain A, we find (like Sweeney and Feser, 1998, for example):

$$\hat{K}(r) = \frac{A}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j>i}^{N} c(i,j,r) \tag{4}$$

In a space where edge effects do not occur (say a torus), $c(i,j,r)$ is a dummy: its value is 1 if the distance between points i and j is less than r. In the real world, it is corrected for edge effects: when point i is close to the boundary of the domain, it has fewer neighbours because those outside the domain are not observed. After computing $\hat{K}$, $g(r)$ can be estimated by $\hat{g(r)} = \hat{K}(r + \Delta r) - \hat{K}(r - \Delta r)/2\Delta r$, taking Δr as arbitrarily small. $\hat{g}$ is proportional to the number of point pairs whose distance is close to r.

10 The term 'co-localization' is used by Duranton and Overman (2005) when co-agglomeration results from attraction on the part of both industries whereas they prefer the term 'joint-localization' when both industries exhibit significant co-agglomeration for another reason. However, they mention that they cannot disentangle these location patterns empirically.

11 As stated in the introduction, the debate about implementing absolute or relative measures for evaluating geographic concentration has recently been settled, as one of the criteria of a 'good' concentration measure states that relative indices must be preferred (Combes and Overman, 2004; Duranton and Overman, 2005). This article does not question this view.

12 See Marcon and Puech (2003) for a concise presentation of this function.

K is a cumulative function, while g is a local function. $\hat{K}$ has been widely used in the literature, but $\hat{g}$ has not. Both are restricted to homogenous point processes. Further mathematical developments are necessary in order to characterize inhomogeneous point sets and to control for the spatial distribution of the whole economy. Duranton and Overman (2005) chose to define the *Kd* function as the probability density function of point-pair distances. *Kd* is also proportional to the number of point pairs whose distance is close to r. The differences between *Kd* and $\hat{g}$ are: (i) *Kd* integrates appropriate smoothing, but this is only a technical improvement, and (ii) *Kd* does not correct for any edge effects. Its value is compared to those of point distributions with the same geometry, which have the same edge effects. We chose another approach. From Equation (4) and using the same definitions of T_i and T as those given in Section 2.1., it follows that the expression for $\hat{K}$ can be rearranged as:

$$\frac{\hat{K}(r)}{\pi r^2} = \frac{\sum_{i=1}^{N-1} \frac{\sum_{j>i}^{N} c(i,j,r)}{\pi r^2}}{N} \Big/ \frac{N-1}{A} = \frac{\sum_{i=1}^{N} T_i/N_i}{N} / T/N \tag{5}$$

It is thus a particular case of M.

To summarize, all these functions are derived from the raw data that is the number of point pairs at a given distance. Ripley's $\hat{g}$ is normalized so that its value is 1 when points are distributed independently. Duranton and Overman's *Kd* is normalized to be a probability density function. Ripley's $\hat{K}$ is the cumulative function of $\hat{g}$. It can be interpreted as the ratio of the observed number of neighbours to the number of neighbours there would be if the points were distributed independently. The M function is its generalization in non-homogenous space.

4.2 *Kd* and *M* functions as complementary measures of geographic concentration

In what follows, we need to re-examine a well-known dilemma, namely whether a probability density function or a cumulative function should be used. Some examples will be given to illustrate the advantages and limitations of both measures. We shall then explain why *Kd* and M can be considered as useful complements to each other in economic geography.

4.2.1 Fundamental differences between Kd and M

4.2.1.1. Property 1: like any probability density function, Kd *detects local density more precisely at different spatial scales.* The idea here is simple. A probability density function like *Kd* can detect structures at specific spatial scales and thus easily identifies local patterns. Like any cumulative function, M accumulates spatial information on the distribution of points up to a certain scale: its local density estimates then become less precise (Condit et al., 2000). Duranton and Overman (2005) underlined this important property when they asserted in the conclusion of their article that '*K*-densities are more informative than [Ripley's] *K*-functions with respect to the scale of localization'. In their working article, the reason for preferring a probability density function is even clearer:

> Traditionally, spatial statisticians have used the cumulative of the *K-density*, the *K*-function. [...] Given that a major objective of our analysis is to distinguish the spatial scale(s) at which

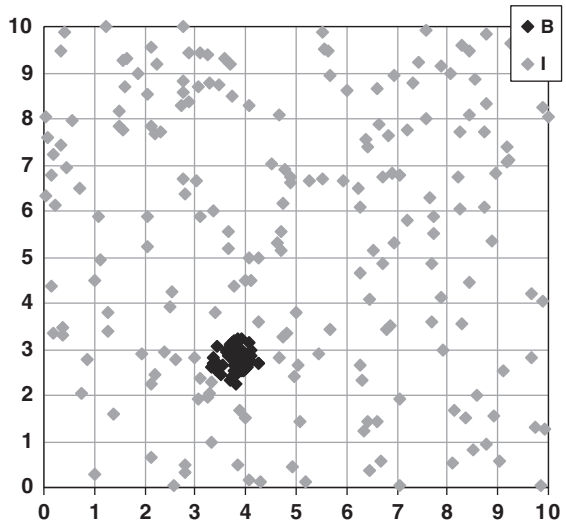


Figure 1. First theoretical distribution of establishments on a 10×10 area.

> excess-localisation takes place, the focus on the density distribution rather than its cumulative is warranted. (Duranton and Overman, 2001, 7)

Let us turn to an example to understand what issue is at stake. Consider two industries localized on a 10 x 10 territory. The first industry I has 200 production units randomly distributed (Poisson distribution) over the whole area. The second industry B is generated by a Matérn process (Matérn, 1960, cited by Stoyan et al., 1987): 50 points are uniformly generated in one cluster of radius 0.5. The location of establishments of both industries is shown in Figure 1. As in the other illustrative theoretical cases elsewhere in the article, the weight of each plant is equal to 1 to simplify the examples and the confidence intervals are computed at a 1% threshold, from 10,000 simulations (only global confidence intervals are shown in the figures). *Kd, M* and global intervals are computed at intervals of 0.5 up to a radius of 10.[13] The *Kd* and *M* functions for industry B are respectively given in Figures 2 and 3. Finally, note that Duranton and Overman (2005) recommend analysing the spatial pattern up to the median distance between all pairs of plants. However, in what follows, the results are given for all possible radii to describe completely the behaviour of the functions (even though this should not be done in empirical studies).

What can we learn from this example? As we can see in Figures 2 and 3, the first positive significant peak of the *Kd* and the *M* functions appears at a radius of 0.5.[14] At this distance, both measures successfully detect the circular cluster of the Matérn

13 It is interesting to note that $Kd(0)$ is defined by smoothing, whereas $M(r)$ may be undefined, if no plant has any neighbor less than r apart.

14 Note that contrarily to *Kd*, *M* values are not defined below *M* (0.5). As we previously underlined, this result is coherent: as any cumulative function, *M* does not use smoothing techniques (as *Kd* does) so it is not defined for distances lower than that of the closest point pair.

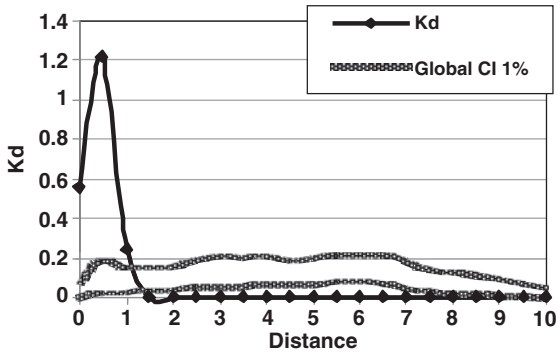


Figure 2. *Kd* function for industry B.

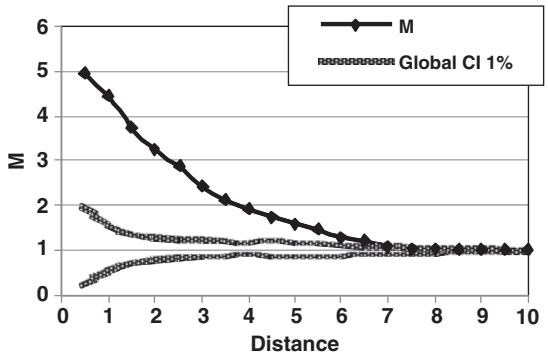


Figure 3. *M* function for industry B.

process. However beyond the distance of 0.5, the *Kd* and *M* plots are clearly different. Since there are no industry B neighbours beyond a radius of 0.5, the *Kd* values rapidly decrease while the *M* plot gradually (and not suddenly like *Kd*) returns to within the confidence interval bands. The *M* plot gives the impression of being less precise than the *Kd* plot because its slope is not steep beyond a radius of 0.5. However, the *M* plot is completely in accordance with the definition of a cumulative function and there is no misinterpretation. *M* perfectly detects the lack of industry B neighbours' because the *M* values decrease beyond a distance of 0.5. However, the *Kd* results seem more intuitive because for any cumulative function 'aggregation at smaller scales influences the estimate at larger scales' (Wiegand and Moloney, 2000, 220).

4.2.1.2. Property 2: like any cumulative density function, the M function identifies spatial structures of point patterns better. This property highlights the fact that only a cumulative function can reveal a superposition of different spatial point patterns. We illustrate this second property with two theoretical examples: independent distribution of clusters and spatial repulsion between clusters. In the first example, the clusters of an industry C are generated by a Matérn process: 50 plants are uniformly generated in 9 clusters of radius 0.5 randomly distributed on a 10×10 domain.

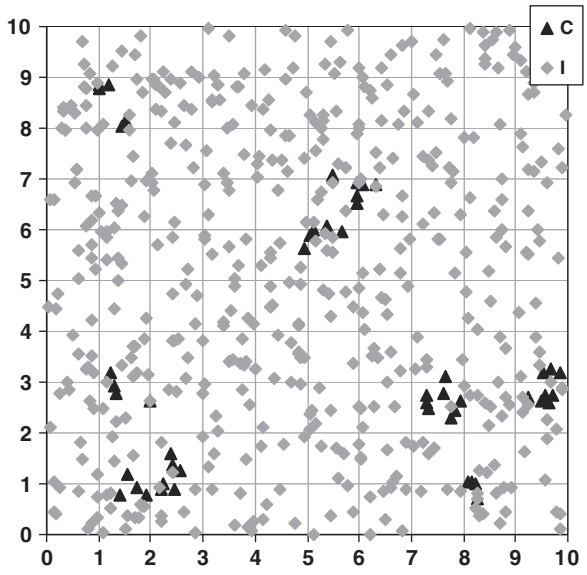


Figure 4. Second theoretical distribution of establishments on a 10×10 area.

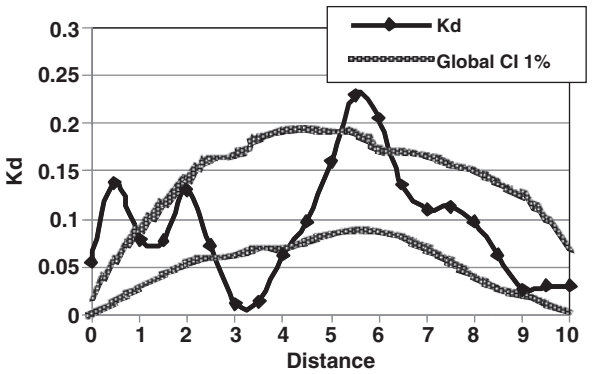


Figure 5. *Kd* function for industry C.

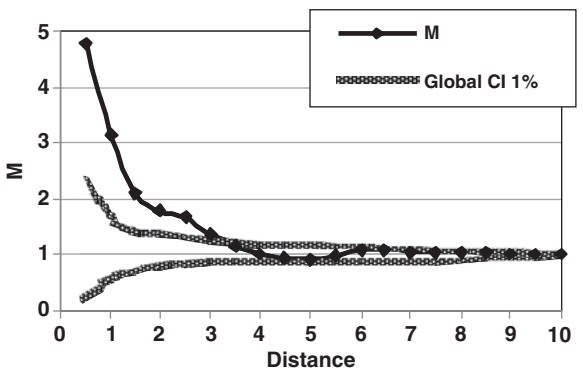


Figure 6. *M* function for industry C.

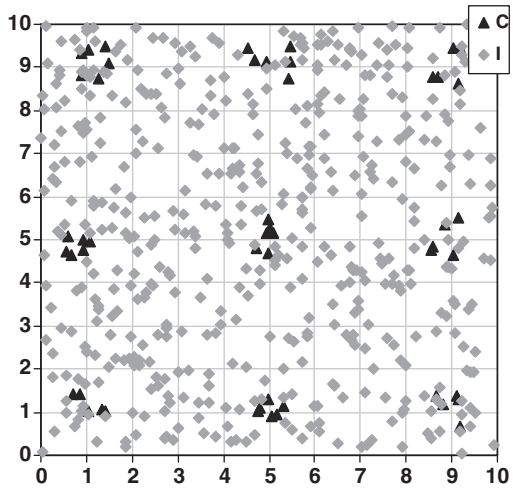


Figure 7. Third theoretical distribution of establishments on a 10×10 area.

500 establishments of another industry (called I) are randomly distributed according to a Poisson distribution. The map of the distribution of plants is given in Figure 4 and the results of *Kd* and *M* for industry C are shown in Figures 5 and 6. In the second example, the same industry I composed of 500 randomly distributed plants is located in the same area. However, the 50 establishments of industry C are now located in 9 clusters regularly distributed on a squared grid (Figure 7).

To begin with, let us consider the *Kd* plots for industry C (Figures 5 and 8). A first significant peak is detected within a radius equal to 0.5, corresponding to the size of the clusters. The *M* function corroborates the findings with the *Kd* function. At larger distances, differences between the plots appear. The positive significant peaks of both *Kd* functions reveal several excessive point-pair distances corresponding to the relative position of aggregates.[15] Between the clusters, both *Kd* plots indicate that there is a lack of industry C plants: it can be shown in Figures 5 and 8 that negative significant peaks emerge around a radius of 1.5–3. Without looking at the maps, the information given by the *Kd* values is insufficient for us to tell whether the observed lack of point-pair distances is caused by repulsion between clusters or a random distribution. This is a clear and important weakness of *Kd*. Estimations of the *M* function solve the dilemma: in the second case 'only', significant repulsion is detected between 2.5 and 4 (and between 7 and 8.5). The *M* plot is below the confidence interval at these distances, detecting the regular position of the clusters on the grid.

4.2.1.3 Property 3: M values are easier to interpret. *Kd* and *M* evaluate the spatial distribution of plants and, for every distance, summarize the location patterns at a certain level of concentration. It would be of value to a researcher for the results to be easy to understand. Ideally, the values that are obtained should be comparable (i) at several radii (ii) across sectors and (iii) for several points in time. In this respect, *M* is

15 For instance, in Figure 8, note that the first neighbours of industry C plants outside the cluster appear at a radius approximately equal to 4 (the grid size) and then around 8 (twice the grid size).

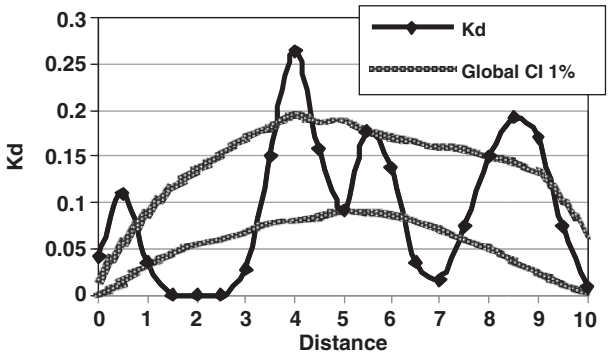


Figure 8. *Kd* function for industry C.

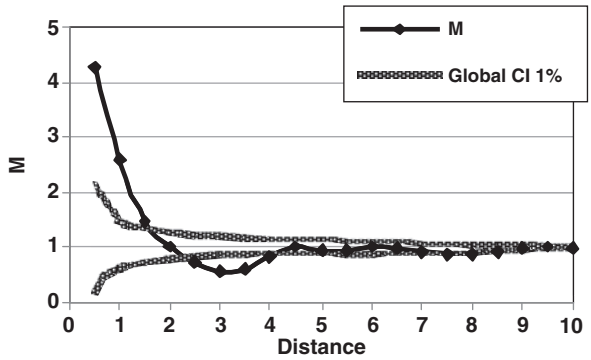


Figure 9. *M function* for industry C.

undoubtedly more useful than *Kd*. *M* function values can effortlessly be interpreted and compared whatever the distance. To give an example, let us consider the case where the *M* plot is outside the confidence interval (centered on 1) and *M* reaches the value of 5 at a given radius. This means that the relative density of neighbours from the considered industry 'within this radius' is 5-times higher than it would be if its plants were distributed as they are in the whole industry. This value is a measure of concentration, so it can be understood independently from the data it was calculated from. The *Kd* values are the probability density of a plant having a neighbour at a given distance. They depend both on the geometry of the plant distribution and the deviation of the sector under scrutiny from the benchmark distribution. The *Kd* values are consequently unsuitable. Duranton and Overman (2005) suggest using the difference between *Kd* and the upper $\overline{Kd}(r)$ or lower $\underline{Kd}(r)$ band of the confidence interval as a normalized measure. Nevertheless, the resulting normalized values still have no meaning.

4.2.2 Implications for measuring the geographic concentration of economic activities

We can now examine the implications of the above properties for the field of economic geography. Basically, like any geographic concentration measure, both *Kd* and *M* could

be useful for resolving two types of issues. The first concerns the characterization of the spatial distribution of economic activity. The second is related to identifying the determinants of agglomeration.

(i) *Kd* and *M* are useful for exploring the location patterns of economic activities.

As a general rule, for analysing a point pattern in economics we suggest first using *M* to have a global view of the spatial structure of the distribution, and then *Kd* to obtain more detailed information. There are three reasons for this. Firstly, *Kd* provides more precise results for the local density of establishments (Property 1). Secondly, a density function such as *Kd* is not able to evaluate the global effect of the superposition of spatial structures (Property 2). The absence of neighbour plants at a given distance may be the consequence of repulsion (Figure 7) or just compensation for very strong attraction at another scale (Figure 8).[16] Thirdly, we suggest in the theoretical examples that an overview of the map may avoid any misinterpretation of *Kd*. This technique is unhelpful for evaluating the geography of production because in real life things are more complex than in the theoretical cases considered above. Both the number of establishments and the number of sectors under study are higher, which means that looking at the map is less informative. Here too, the simultaneous use of both measures is advisable.

An exception to the general rule will be given by Property 3. If the only reason for using a geographic concentration index is to quantify the spatial deviation from locational randomness, the appropriate measure is undoubtedly the *M* function. The level of concentration of dispersion is certainly more intelligible for a decision-maker than any normalized results without concrete meaning.

(ii) *M* seems more appropriate for understanding the determinants of the geographic concentration of activities.

Should the influence of the determinants of agglomeration be analysed *at* a given distance or *up to* a certain distance? The answer is at the heart of the dilemma between choosing *M* or *Kd*.[17]

The vast majority of previous studies that have analysed the determinants of agglomeration clearly used a cumulative approach. Authors systematically regress the observed level of a cluster-based measure of spatial concentration with respect to different local factors (Kim, 1995; Ellison and Glaeser, 1999; Rosenthal and Strange, 2001; Co, 2002). Such studies are based on a particular zoning of the area, individual data is thus aggregated up to a certain level (such as a region or state). Is the availability of data the only reason for this? We do not think so: the proof is given by authors who deliberately avoid pre-defined zoning, preferring the geometric form of a disc in order to evaluate the global impact of the surrounding plants. For example, Holmes

16 Duranton and Overman (2005, 1086) suggest studying the geography of production only up to an economically pertinent distance. They define this maximum distance as the median radius between all pairs of plants. It is interesting to note that our recommendation of using both measures is all the more valid in the example that illustrates Property 2 because the problematic negative peak of *Kd* plots appears below the median distance for industry C in Figure 7.

17 However, the way of testing the determinants of agglomeration is far from being settled. Several live debates still exist in the literature. For example, are concentration indexes the appropriate variables to answer this question (Combes et al., 2008, chapter 11)? If so, should a continuous-space framework or a discrete-space one be preferred (Ellison et al., forthcoming)?

(1999) estimates the link between vertical disintegration and the level of geographic concentration of plants by applying an integrative approach. However, probability density functions may also be useful but only in capturing the marginal effect of the factors of agglomeration. Some examples are given by Rosenthal and Strange (2003, 2008) who consciously use several concentric rings to estimate the spatial scope of externalities.

5. Conclusion

The aim of this article was to improve some existing relative statistical tools for testing the spatial concentration of industries. We have shown that *M* functions constitute first-class instruments for evaluating intra- or inter-industry geographic concentration. We have proved that the cumulative *M* function must be implemented with the probability density function *Kd* to give a complete and good description of the distribution of economic activities. Nonetheless, some intrinsic limits of these new tools suggest that other research approaches are needed to fill the gap between the theoretical and empirical literatures. Despite the considerable recent interest of researchers in an 'ideal' concentration index, and even though significant progress has been made, work still needs to be done to meet the most difficult criteria suggested by Combes and Overman (2004): the complete integration of the tools to economic theory, and the independence of geographic concentration measures from the industrial classification. In this article, we have enhanced existing distance-based methods but further investigations are still required.

Acknowledgements

We are grateful to Richard Arnott, Gilles Duranton, Pablo Jensen, John McBreen, many seminar and conference participants, the editor Henry Overman and three anonymous referees for their very helpful suggestions and comments.

References

Arbia, G. (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.

Arbia, G., Espa, G. (1996) *Statistica Economica Territoriale*. Padua: Cedam.

Arbia G., Espa G., Quah, D. (2008) A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics*, 34: 81–103.

Barff, R. A. (1987) Industrial clustering and the organization of production: a point pattern analysis of manufacturing in Cincinnati, Ohio. *Annals of the Association of American Geographers*, 77: 89–103.

Besag, J. E. (1977) Comments on Ripley's paper. *Journal of the Royal Statistical Society B*, 39: 193–195.

Briant, A., Combes, P.-P., Lafourcade, M. (forthcoming) Dots to boxes: do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*.

Co, C. Y. (2002) The agglomeration of US-owned and foreign-owned plants across the US States. *The Annals of Regional Science*, 36: 575–592.

Combes, P.-P., Mayer, T., Thisse, J.-F (2008) *Economic Geography: The Integration of Regions and Nations*. Princeton: Princeton University Press.

Combes, P.-P., Overman, H. (2004) The spatial distribution of economic activities in the European Union. In J. V. Henderson, J.-F. Thisse (eds) *Handbook of Urban and Regional Economics*. North Holland, Amsterdam: Elsevier.

Condit, R. et al. (2000) Spatial patterns in the distribution of tropical tree species. *Science*, 288: 1414–1418.
Duranton, G., Overman, H. G. (2001) Localisation in UK Manufacturing Industries: Assessing Non-Randomness Using Micro-Geographic Data. Working Paper.
Duranton, G., Overman, H. G. (2005) Testing for localisation using micro-geographic data. *Review of Economic Studies*, 72: 1077–1106.
Duranton, G., Overman, H. G. (2008) Exploring the detailed location patterns of UK manufacturing industries using microgeographic data. *Journal of Regional Science*, 48: 213–243.
Ellison, G., Glaeser, E. L. (1999) The geographic concentration of industry: does natural advantage explain agglomeration? *The American Economic Review, American Economic Association Papers and Proceedings*, 89: 311–316.
Ellison, G., Glaeser, E. L., Kerr, W. R. (forthcoming) What causes industry agglomeration? Evidence from coagglomeration patterns. *The American Economic Review*.
Feser, E. J., Sweeney, S. H. (2000) A test for the coincident economic and spatial clustering of business enterprises. *Journal of Geographical Systems*, 2: 349–373.
Fratesi, U. (2008) Issues in the measurement of localization. *Environment and Planning A*, 40: 733–758.
Goreaud, F., Pélissier, R. (1999) On explicit formulas of edge effect correction for Ripley's *K*-function. *Journal of Vegetation Science*, 10: 433–438.
Haaland, J. I., Kind, H. J., Midelfart-Knarvik, K. H., Torstensson, J. (1999) What determines the economic geography of Europe? Centre for Economic Policy Research. Discussion paper, 2072.
Holmes, T. J. (1999) Localization of industry and vertical disintegration. *The Review of Economics and Statistics*, 81: 314–325.
Jensen, P. (2006) Network-based predictions of retail store commercial categories and optimal locations. *Physical Review*, E 74: 035101(R).
Kim, S. (1995) Expansion of markets and the geographic distribution of economic activities: the trends in US regional manufacturing structure, 1860–1987. *The Quarterly Journal of Economics*, 110: 881–908.
Krugman, P. (1991) *Geography and Trade*. London: MIT Press.
Marcon, E., Puech, F. (2003) Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography*, 3: 409–428.
Matérn, B. (1960) Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut*, 49: 1–144.
Mori, T., Nishikimi, K., Smith, T. E. (2005) A divergence statistic for industrial localization. *Review of Economics and Statistics*, 87: 635–651.
Morphet, C. S. (1997) A statistical method for the identification of spatial clusters. *Environment and Planning A*, 29: 1039–1055.
Ó hUallacháin, B., Leslie, T. F. (2007) Producer services in the urban core and suburbs of Phoenix, Arizona. *Urban Studies*, 44: 1581–1601.
Pancer-Koteja, E., Szwagrzyk, J., Bodziarczyk, J. (1998) Small-scale spatial pattern and size structure of Rubus hirtus in a canopy gap. *Journal of Vegetation Science*, 9: 755–762.
Puech, F. (2003) Concentration géographique des activités industrielles: Mesures et enjeux. PhD thesis, Université de Paris I.
Ripley, B. D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13: 255–266.
Ripley, B. D. (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society B*, 39: 172–212.
Rosenthal, S. S., Strange, W. C. (2001) The determinants of agglomeration. *Journal of Urban Economics*, 50: 191–229.
Rosenthal, S. S., Strange, W. C. (2003) Geography, industrial organisation, and agglomeration. *Review of Economics and Statistics*, 85: 377–393.
Rosenthal, S. S., Strange, W. C. (2008) The attenuation of human capital spillovers. *Journal of Urban Economics*, 64: 373–389.
Rowlingson, B. S., Diggle, P. J. (1993) SPLANCS: Spatial Point Pattern Analysis Code in S-Plus. *Computers and Geosciences*, 19: 627–655.

Stoyan, D., Kendal, W. S., Mecke, J. (1987) *Stochastic Geometry and its Applications*. New York: John Wiley & Sons.

Sweeney, S. H., Feser, E. J. (1998) Plant size and clustering of manufacturing activity. *Geographical Analysis*, 30: 45–64.

Wiegand, T., Moloney, K. A. (2006) Rings, circles, and null-models for point pattern analysis in ecology. *Oikos*, 104: 209–229.

GENERALIZING RIPLEY'S *K* FUNCTION TO INHOMOGENEOUS POPULATIONS[1]

ERIC MARCON[2] AND FLORENCE PUECH[3]

Abstract:

In spatial statistics, Ripley's K *function (Ripley (1977)) is a classical tool to analyse spatial point patterns. Yet, it faces two major limits: it is only pertinent for homogeneous point processes and it does not allow the weighting of points.*

We generalize it to get a new function, M*, which oversteps these limits and detects spatial structures of inhomogeneous populations of weighted points.*

[1] We wish to thank Gabriel Lang for his helpful comments.

2 Corresponding author, AgroParisTech ENGREF, UMR EcoFoG, BP 316, 97310 Kourou, French Guyana. E-Mail: Eric.Marcon@agroparistech.fr

3 LET (Université de Lyon, CNRS, ENTPE), Institut des Sciences de l'Homme, 14 av. Berthelot, 69363 Lyon Cedex 07, France. E-Mail: Florence.Puech@univ-lyon2.fr

1 Introduction

The basic analysis of a point set relies on its first-order property, that is to say the average values of chosen variables. An example is given by foresters who classically describe a forest plot with the histogram of tree diameters. This is enough for forestry management, but some scientific fields (like ecology) study the way species interact with each others, tackles new questions: how do these trees occupy space? Do trees of the same species aggregate or repulse each other?

New tools have progressively been developed to rigorously answer these more complicated questions. A major milestone was established by Clark and Evans (1954). The general method was given: measure a pertinent variable (the distance from every object to its nearest neighbour) and compare its value to the one that would have been given by randomness. Next, a fundamental step was made by Ripley (1976) who wrote the full theory of the second-order properties of point processes, giving the framework for better tools, taking into account all neighbours rather than only the nearest. Ripley (1977) then introduced the *K* function to analyse point-set structures. This function has been widely used for 25 years and became a standard measure presented in spatial statistic handbooks (Ripley (1981), Diggle (1983), Upton and Fingleton (1985), Cressie (1993)). However, Ripley's *K* function faces two important limits: it supposes homogeneous space, and considers all points as equivalent (*i.e.* points' characteristics do not matter). Therefore, this tool seems to be inappropriate to analyse obviously non-stationary point sets or objects whose size matters, like the study of the spatial structure of manufacturing plants (Marcon and Puech (2003a)). A first solution was given by Cuzick and Edwards (1990) who developed a non-parametric test able to detect clustering in a non-homogeneous point set, followed by Diggle and Chetwynd (1991) who introduced the *D* function, defined as the difference between the *K* function for studied points (called *cases*) and the *K* function for the others (called *controls*). This is not completely satisfactory yet because, since both *K* functions are computed separately, all the data contained in the relative position of cases and control is lost.[4] Baddeley *et al.* (2000) generalize *K* to inhomogeneous point processes. They give a clean theoretical framework but practical applications are difficult, as we will se here. Therefore, the purpose of this study is to give a mathematical framework of a new measure, namely the *M* function, which oversteps these limits and actually detects spatial structures of inhomogeneous populations of weighted points.[5]

The paper is organised as follows. The next two parts recall the important features of point processes and that of Ripley's *K* function. Part 4 constitutes a discussion on some features of Ripley's *K* function to open the way (part 5) for its generalisation to inhomogeneous space and weighted points, introducing the *M* function. A comparison with Baddeley's K_{inhom} is developed.

2 Point processes

A point process is the equivalent of a random variable whose result is a point, defined by its coordinates *(x, y)* in a pre-defined area that we will call the *domain*, known and delimited.

[4] We will not detail hereinafter all tools derived from the *K* function but the reader should refer to the literature cited, including the definition of a standardised version of *D* proposed by Feser and Sweeney (2000).

[5] Note that an *ad hoc* economic version of the *M* function was introduced in Marcon and Puech (2003b).

Point processes are used as mathematical tools to characterize and eventually model events whose spatial repartition is known, such as trees in a forest.
An interesting way to describe an unknown-law process is through its first and second-order properties.

2.1 First-order property

Definition

Consider an area *A* supplying a realization of a point process. *N* is the actual number of points inside *A*. Each point is defined by its coordinates *(x, y)*. We denote *N(S)* the number of points inside a given sub-area *S*.
The process first-order property is its *density*, denoted $\lambda(x, y)$. Its definition is:

$$\lambda(x, y) = \lim_{dS \to 0} \frac{E[N(dS)]}{dS} \quad \textbf{(1)}$$

where *dS* is the elementary area around (*x*, *y*).
If $\lambda(x, y)$ is a constant, we will say the point process is **homogeneous or stationary,** and the density will just be denoted λ.

Probability to find a point in an elementary area

We will only consider *ordered* point processes (Diggle (1983), p.47), *i.e.* the magnitude of the probability to find several points in an elementary area *dS* is smaller than *dS*. In other words, we will be allowed to write that the probability to find several points in *dS* is almost equal to the probability to find one only.
This assumption is not restrictive. To get convinced, consider a process providing independent points. The probability to find two points in *dS* is $(P_{dS})^2$. According to the first-order property, it equals $[\lambda(x, y)dS]^2$. Since *dS* is small, $(dS)^2$ is negligible compared to *dS*.
This property establishes the linkage between probability and density. The existence of a point in *dS* is the result of a Bernoulli proof of parameter P_{dS}. The number of points in *dS* thus follows a Bernoulli law and its expectation is P_{dS}. According to equation (1), this expectation is $\lambda(x, y)$.
The probability to find at least one point in the elementary area *dS* around the point located at *(x, y)* is consequently:

$$P_{dS} = \lambda(x, y)dS \quad \textbf{(2)}$$

This relation is verified as long as *dS* is small enough for the probability to find two points remains negligible.

2.2 Second-order property

Definition

The second-order property of a point process, denoted $\lambda_2((x_1,y_1),(x_2,y_2))$, is defined by:

$$\lambda_2((x_1,y_1),(x_2,y_2)) = \lim_{dS_1,dS_2 \to 0} \frac{E[N(dS_1)N(dS_2)]}{dS_1 dS_2} \quad \textbf{(3)}$$

Probability to find two points in two elementary areas

The joint probability to find at least one point in each elementary area around (x_1,y_1) and (x_2,y_2) is denoted $P_{dS1,dS2}$. Once again, the probability to find more than one point in an area is negligible. The event "find both a point in dS_1 and in dS_2" realizes a Bernoulli proof with parameter $P_{dS1,dS2}$, so:

$$P_{dS_1,dS_2} = dS_1 dS_2 \lambda_2((x_1,y_1),(x_2,y_2)) \quad \textbf{(4)}$$

Introducing the first-order property:

$$P_{dS_1,dS_2} = P_{dS_1} P_{dS_2} \frac{\lambda_2((x_1,y_1),(x_2,y_2))}{\lambda(x_1,y_1)\lambda(x_2,y_2)} \quad \textbf{(5)}$$

The expression $\frac{\lambda_2((x_1,y_1),(x_2,y_2))}{\lambda(x_1,y_1)\lambda(x_2,y_2)}$, ratio of the second-order to the first-order property, is called *radial distribution function* (Diggle (1983)), or *point-pair correlation function* (Cressie (1993)). We follow Ripley (1977) and the following literature, denoting it $g((x_1,y_1),(x_2,y_2))$. Common usage (for instance Ripley (1977), Stoyan *et al.* (1987)) imposed g rather than λ_2 as the measure of the second-order property. We will follow it:

$$g((x_1,y_1),(x_2,y_2)) = \frac{P_{dS_1,dS_2}}{P_{dS_1} P_{dS_2}} \quad \textbf{(6)}$$

If the process is **isotropic**, $g(\bullet)$ only depends on the distance between points and it will be denoted $g(r)$. In the case of an **independent** point distribution, the joint probability is equal to the product of the individual ones, thus $g(\bullet)=1$. An independent point process is isotropic.

2.3 The homogeneous Poisson process particular case

Complete Spatial Randomness (CSR) is defined by homogeneity and independency.
The homogeneous Poisson point process gives completely random points. Inversely, a completely random point process is a homogeneous Poisson process (proof in Diggle (1983), p.50-51).

First-order property

A realization of a homogeneous Poisson process with parameter λA on the area A is a completely random point set of density λ. The number of points follows a Poisson law with parameter λA, that is to say that:

$$P(N = k) = e^{-\lambda A} \frac{(\lambda A)^k}{k!} \quad \textbf{(7)}$$

This property remains true for any area S chosen within A. The number of points inside it follows a Poisson law: its expectation is λS, its variance is λS (a completely random distribution is not regular). A non-homogeneous Poisson point process is defined similarly, with λ depending on the location. In what follows, we will only consider homogeneous Poisson processes (except if it is explicitly mentioned).

Second-order property

Since point are distributed independently from each other, $g(\bullet)=1$.
The Poisson process will be used as the reference for complete spatial randomness (CSR), to compare the actual point distributions with.[6]

3 Ripley's *K* function

The *K* function, defined by Ripley (1976); Ripley (1977) is a good indicator for spatial structures (Besag (1977), Diggle (1983), Cressie (1993)). Here, we will only consider homogeneous and isotropic point processes.

3.1 Introduction: probability to find a neighbour at a given distance

We call a point *i*'s neighbours all the points located at a distance lower than or equal to a given value *r* (basically, it represents the count of neighbours in a circle of radius r centered on the point *i*). The number of neighbours' expectation is denoted $v(r)$. Its estimator, the observed number of neighbours, is denoted $V(r)$. Ripley (1977) showed that:

$$\frac{v(r)}{\lambda} = \int_{\rho=0}^{r} g(\rho) 2\pi\rho d\rho \quad \textbf{(8)}$$

3.2 Definition of the K *function*

Ripley (1977) defined the *K* function as:

$$K(r) = \int_{\rho=0}^{r} g(\rho) 2\pi\rho \, d\rho \quad \textbf{(9)}$$

If points are distributed independently from each other, $g(\rho)$=1 for all values of ρ, so $K(r)=\pi r^2$. This value is used as a benchmark:

- $K(r) > \pi r^2$ indicates that the average value of $g(\rho)$ is greater than 1. The probability to find a neighbour at the distance ρ is then greater than the probability to find a point in the same area anywhere in the domain: points are aggregated.

[6] Diggle (1983), p.50, calls it the "*cornerstone eon which the theory of spatial point processes is built*".

- Inversely, $K(r) < \pi r^2$ indicates that the average neighbour density is smaller than the average point density on the studied domain. Points are dispersed.

K(r) is estimated by the ratio of the average number of neighbours on the density, estimated itself by the total number of points divided by the domain area ($\hat{\lambda} = N/A$):

$$\hat{K}(r) = \frac{\hat{v}(r)}{\hat{\lambda}} = \frac{V(r)}{N/A} \quad \textbf{(10)}$$

The average number of neighbours can be expressed more explicitly by defining the indicator *c(i,j,r)*=1 if the distance between points *i* and *j* is at most *r*, 0 otherwise:

$$\hat{K}(r) = \frac{1}{\hat{\lambda} N} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} c(i, j, r) \quad \textbf{(11)}$$

3.3 Correction of the edge effects

Points located close to the domain borders are problematic because a part of the circle inside which points are supposed to be counted is outside the domain. Ignoring this edge effect results in underestimating *K*.

Ripley's correction

Ripley (1977) proposed to correct the indicator *c(i,j,r)* introduced in equation (11).
We denote L_{jr} the portion of the circle of radius *r* centred on the point *i* located inside the domain. If a part of the crown of width *dr* inside which a neighbour is counted is outside the domain, the neighbour is given a weight equal to the inverted ratio between the inside part of crown ($L_{jr}dr$) and the whole crown ($2\pi r dr$). The idea is that the outside part of the crown could have contained the same neighbour density than the inside part. The correction is:

$$\hat{K}(r) = \frac{1}{\hat{\lambda} N} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} \frac{2\pi r}{L_{ir}} c(i, j, r) \quad \textbf{(12)}$$

Besag's correction

Besag (1977), in his discussion of Ripley's paper, underlined that this correction gave an excessive weight to the farthest neighbours. The greater the radius *r*, the smaller L_{jr}, and the bigger the correction. He proposed an alternative: correct the edge effect not for each neighbour, but for all of them the same way.
We denote A_{ir} the part of the area of the circle of radius *r* centred on the point *i* located inside the domain. We count the number of neighbours inside the circle and we correct it by the ratio between the circle's area and its inside part. We suppose that the outside part of the circle would have contained the same neighbour density than the inside part. Finally:

$$\hat{K}(r) = \frac{1}{\hat{\lambda} N} \sum_{i=1}^{N} \frac{\pi r^2}{A_{ir}} \sum_{j=1, i \neq j}^{N} c(i, j, r) \quad \textbf{(13)}$$

Even though these edge-effect corrections methods are used alternatively, Ripley's is still widely applied in the literature (see for instance Haase (1995)).

Ward and Ferrandino's correction

Ward and Ferrandino (1999) introduced a global correction, arguing that local correction methods depend on the number and position of points close to the borders, thus introducing more variability in *K*'s estimator.
They proposed to evaluate the expectation of the number of points concerned with edge-effect correction for a given radius (this is not a problem since the point process is supposed to be homogeneous), compute the correction for them (by the Besag's method) and finally calculate the global underestimation of the number of point pairs which only depends on the domain's geometry. They denoted K_A (*A* for analytical) their estimator of *K* defined by:

$$\hat{K}_A(r) = \frac{1}{\hat{\lambda}(N-1)C(r)} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} c(i, j, r) \quad \textbf{(14)}$$

C(r) is the global correction factor. For a rectangular (*L* by *W*) domain, they found (as long as $r < W/2$):

$$C(r) = 1 - \frac{4}{3\pi}\left(\frac{r}{L} + \frac{r}{W}\right) + \left(\frac{11}{3\pi} - 1\right)\left(\frac{r^2}{WL}\right) \quad \textbf{(15)}$$

Note that they did not justify the replacement of *N* by *N-1* in the denominator of the estimator. We will explain the importance of this change, further.
Additionally, the authors calculated the variance of their estimator and its confidence interval. They claimed that their estimator both reduces *K*'s estimation bias and increases its efficiency and they justified this by several examples.

Other correction methods

The other correction methods are much more anecdotic. The most simple of all consists in using a buffer zone around the domain. The buffer is used to count neighbours but reference points (the points *i*) are never taken inside it. The buffer width is equal to the largest value of *r* so no edge effect ever appears. Since the buffer zone contains as much data as the domain, considering that most of the work is collecting data, the temptation is great to include the buffer into the domain: this method is very rarely used (examples can be found in Szwagrzyk and Czerwczak (1993), Kuuluvainen and Rouvinen (2000), fig.1a, p.803). The toroidal correction consists in treating the domain as a torus, that is to wrap it so that its opposite borders are in contact, supposing of course that its shape allows it. A good illustration is given by Haase (1995), fig.3, p.578. This solution is intuitively little satisfactory because it considers that the opposite points as very close. It was used by Peterson and Squiers (1995) and Kuuluvainen and Rouvinen (2000).
Empirical issues due to the edge-effect correction may be considerable. Getis and Franklin (1987) give formulas for a rectangular domain, Diggle (1983), p.72, for a rectangle and a

circle. Haase (1995) reviews and compares correction methods: Ripley's, the buffer zone, and the torus. He notices and corrects an error in Diggle's list of cases needing a correction, leading to a serious underestimation of *K* and a little error leading to a slight overestimation in Getis and Franklin's formulas. Goreaud and Pélissier (1999) developed algorithms to study more complex domains, implemented in ADE software (Thioulouse *et al.* (1997)). Treating complex geographical limits such as a country's boundaries is possible, but was never applied in the literature: the domain is always a polygon (Sweeney and Feser (1998), fig.1, p.52, Rowlingson and Diggle (1993), fig.5, p.634), or, more rarely, a circle (Pancer-Koteja *et al.* (1998), fig.1-3, p.757).

3.4 Besag's L *function*

Ripley's function is not very convenient to use. Comparing a computed value to its benchmark, πr^2, implies more computing and the hyperbolic chart is not very expressive.
Besag (1977) proposed to normalize the function to obtain a benchmark of zero:

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r \qquad \textbf{(16)}$$

3.5 Significance

The estimated values of *K* and *L* are compared to benchmarks given by a homogeneous Poisson process. To test whether a value of $\hat{L}(r)$ is significantly different from 0, the most common way is using the Monte Carlo technique (Diggle (1983)):

- A great number of random data sets is generated. Each of them is consistent with the null hypothesis tested.
- A confidence threshold α is chosen.
- For each value of *r*, $\hat{L}(r)$ values are sorted in increasing order. The *n*th value is denoted $\hat{L}_n(r)$.
- Extreme values are eliminated: the null hypothesis confidence interval limits are $\hat{L}_{N(\alpha/2)}(r)$ and $\hat{L}_{N(1-\alpha/2)}(r)$. For *N=1000* and α*=5%*, we retain the 26th and the 974th values.
- $\hat{L}(r)$ is considered significantly different from 0 if its value is outside the interval $[\hat{L}_{N(\alpha/2)}(r); \hat{L}_{N(1-\alpha/2)}(r)]$.

Attempts to directly calculate the confidence interval can be found in the literature. Ripley (1979) respectively proposed $\pm 1{,}42\frac{\sqrt{A}}{N-1}$ and $\pm 1{,}68\frac{\sqrt{A}}{N-1}$ as approximations of the interval limits at 5% and 1% thresholds. These values were obtained from simulations. Due to the lack of theoretical background, these values are very little used (for example by Szwagrzyk and Czerwczak (1993)).

4 Discussions on Ripley's *K* function

4.1 Global confidence intervals

4.2 Edge-effect corrections

Correcting the edge effects by Ripley's method, equation (12), is impossible if a single value of L_{jr} equals 0, that is, as soon as r is big enough for a circle around a point to be completely outside the domain. If the domain is a rectangle, K's computation is thus limited to half of its length. Diggle (1983), p.72, gave correction formulas applicable up to half of the width. Goreaud and Pélissier (1999) improved the edge-effect correction to allow computing K up to half of the rectangle's length.
We will rather use Besag's method, equation (13), which is not limited. Detailing the density estimator, we get the expression of K, corrected from the edge effect:

$$\hat{K}(r) = \frac{A}{N^2} \sum_{i=1}^{N} \frac{\pi r^2}{A_{ir}} \sum_{j=1, i \neq j}^{N} c(i, j, r) \quad \textbf{(17)}$$

4.3 Correction of K*'s bias*

Let us calculate $\hat{K}(r)$ according to equation (17) for a great value of r, such as the part of domain area included in each circle is the domain itself: $A_{ir}=A$ for all points i. Thus, every point's distance to any other is smaller than r: $c(i,j,r)=1$. We can calculate K:

$$\hat{K}(r) = \frac{A}{N^2} \sum_{i=1}^{N} \frac{\pi r^2}{A} \sum_{j=1, i \neq j}^{N} 1 = \pi r^2 \frac{N-1}{N} \quad \textbf{(18)}$$

This result is problematic: the point set structure is homogeneous by assumption, so K should tend to πr^2. This issue is rarely mentioned in the literature because Ripley's edge-effect correction method limits r to a fraction of the domain's size. Getis (1984) remarks that the number of point pairs is $N(N-1)$, so an unbiased estimator of the squared density is $N(N-1)/A^2$. Getis and Franklin (1987) use it without further explanations. Diggle and Chetwynd (1991) indirectly evoked it when they gave a different formulation for K "to get an unbiased estimator of K", without explaining the reason. Sweeney and Feser (1998) used the methods from Diggle and Chetwynd (1991) including their unbiased estimator. Moeur (1993) wrote that the estimator is biased, but only slightly, and used the same formula. Finally, Jones *et al.* (1996) used an unbiased formulation consistent with equation (19), below, justifying it by the loss of one degree of freedom.
The issue's cause must be searched in λ's estimator. The density estimator used in equation (19) is not the number of points divided by the area (N/A) because one of the points is necessarily at the centre of the circle and cannot be found in the crown. The unbiased density estimator is $(N-1)/A$. We can write an unbiased estimator for K:

$$\hat{K}(r) = \frac{A}{N(N-1)} \sum_{i=1}^{N} \frac{\pi r^2}{A_{ir}} \sum_{j=1, i \neq j}^{N} c(i, j, r) \quad \textbf{(19)}$$

4.4 Alternative point of view

Equation (19) can be rearranged:

$$\frac{\hat{K}(r)}{\pi r^2} = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1, i \neq j}^{N} c(i, j, r)}{A_{ir}}}{N} \Bigg/ \frac{N-1}{A} \quad \textbf{(20)}$$

This formulation of Ripley's function, without dimension, is easier to interpret.

Around each point *i*, $\frac{\sum_{j=1, i \neq j}^{N} c(i, j, r)}{A_{ir}}$ is the density of neighbours; its average value for all points is an estimator of D_r, the density of neighbours at the distance *r*. The density of neighbours on the whole domain, denoted D_A, equals $\frac{N-1}{A}$.

Thus *K* can be written as:

$$\frac{K(r)}{\pi r^2} = \frac{D_r}{D_A} \quad \textbf{(21)}$$

K(r), normalized by the area of the circle of radius *r*, is the ratio between the density of neighbours at the distance *r* and the density of neighbours on the whole domain.

The expression $\frac{K(r)}{\pi r^2}$ is an advantageous substitute to *L(r)*. The benchmark is 1. Its estimated value is a ratio of densities. $\frac{K(r)}{\pi r^2}$ peaks occur at distances at which the density of neighbours is the greatest.

5 Generalization of Ripley's *K* function

At this step, we are able to generalize Ripley's *K* function to non-homogeneous weighted point processes. We will first reconsider it from a probabilistic point of view instead of the classical geometric approach. Then, we will assume heterogeneity and different point weights by using appropriate probability laws.

5.1 Probabilistic estimator of K

Let us define a Bernoulli proof consisting in searching a neighbour around a point *i* in an elementary area *dS* in the circle of radius *r*. Its success probability is $\lambda_r dS$. The expectation of the number of neighbours in the circle is $\nu(r) = \lambda_r \pi r^2$ (obtained by summing the elementary

areas point number's expectation). Its estimator is the observed average number of neighbours around all points *i*.
Another Bernoulli proof can be defined by searching a neighbour around the point *i*, but this time, on the whole domain. Its success probability is $\lambda_A dS$. The expectation of the number of neighbours in the circle follows is $\lambda_A A$. Its estimator is *N-1*.
In equation (21), we put the stress on the fact that $K(r)/\pi r^2$ equals the ratio of density D_r to D_A. It comes immediately that:

$$\frac{K(r)}{\pi r^2} = \frac{\lambda_r dS}{\lambda_A dS} = \frac{P_r}{P_A} \quad \textbf{(22)}$$

$K(r)/\pi r^2$ can be estimated by the ratio of two Bernoulli-law probabilities that we will denote P_r and P_A.

5.2 Heterogeneous space

The definitions of Bernoulli proofs can be easily modified to take into account space heterogeneity, *i.e.* not to assume that the underlying point process is stationary. Rather than searching neighbours with an equal probability in a homogeneous space, we will search particular type neighbours among all existing points, whose locations are considered as given.
Following Diggle (1983), we call *cases* the N_{Sk} special points and *controls* the others. The Bernoulli proof consists in searching cases among all point *i*'s neighbours. Its success probability is estimated by the average ratio of cases to both controls and cases located inside the considered area (the circle of radius *r* or the whole domain). More precisely, we define the indicator $c_{Sk}(i,j,r)$=1 if both points *i* and *j* are cases and the distance between them is at most *r*, 0 otherwise:

- P_r is estimated by the average value (on all cases) of the ratio of the number of neighbour cases to the number of neighbour points (controls plus cases):

$$\frac{1}{N_{Sk}} \sum_{i=1}^{N_{Sk}} \frac{\sum_{j=1, i\neq j}^{N_{Sk}} c_{Sk}(i,j,r)}{\sum_{j=1, i\neq j}^{N} c(i,j,r)} \quad \textbf{(23)}$$

- P_A is estimated in the same way, but its expression is simpler: for any point *i*, the number of neighbour cases on the whole domain is $N_{Sk}-1$ and the number of neighbour points is $N-1$.

We define the function *K'*, generalizing *K* to heterogeneous space:

$$K'_{Sk}(r) = \frac{\sum_{i=1}^{N_{Sk}} \frac{\sum_{j=1, i\neq j}^{N_{Sk}} c_{Sk}(i,j,r)}{\sum_{j=1, i\neq j}^{N} c(i,j,r)}}{N_{Sk}} \Bigg/ \frac{N_{Sk}-1}{N-1} \quad \textbf{(24)}$$

5.3 Point weights

Point weights can be attributed to each realization of the Bernoulli proof to define the *M* function. This time:

- P_r is estimated by the average value (on all cases) of the ratio of the *weight* of neighbour cases to the *weight* of neighbour points (controls plus cases): $\frac{1}{N_{Sk}} \sum_{i=1}^{N_{Sk}} \frac{\sum_{j=1,i\neq j}^{N_{Sk}} w_{Sk}(i,j,r)}{\sum_{j=1,i\neq j}^{N} w(i,j,r)}$
- P_A is estimated in the same way but its expression is not as simple as that of identical points. For each point, the ratio is $\frac{W_{Sk} - w_i}{W - w_i}$, so its value changes according to *i*. Its average value is $\frac{1}{N_{Sk}} \sum_{i=1}^{N_{Sk}} \frac{W_{Sk} - w_i}{W - w_i}$

After simplifications, we will retain the following definition for *M*:

$$M_{Sk}(r) = \sum_{i=1}^{N_{Sk}} \frac{\sum_{j=1,i\neq j}^{N_{Sk}} w_{Sk}(i,j,r)}{\sum_{j=1,i\neq j}^{N} w(i,j,r)} \Bigg/ \sum_{i=1}^{N_{Sk}} \frac{W_{Sk} - w_i}{W - w_i} \qquad \textbf{(25)}$$

Points with no neighbour, verifying *w(i,j,r)*=0 cannot be taken into account: there are just ignored in the sums.

5.4 Case-Control design

A particular attention must be paid to case-control designs. For instance, spatial clustering of diseases is a major field of research (Diggle and Chetwynd (1991), Kingham *et al.* (1995), Gatrell and Bailey (1996), Gatrell *et al.* (1996) among others). All cases of disease are carefully referenced but the control point set, *i.e.* all the population, is just sampled. The aim is to characterise the structure of the cases compared to the controls. This approach is of course not limitative to geographical epidemiology.
The usual *M* function defined above could be slightly modified to take into account this feature. Since the controls are chosen to be a representative sample of the population at every scale, the weight of neighbours of any kind is replaced by the weight of controls. After simplifications, *M* can be rewritten as follows:

$$M_{cases}(r) = \sum_{i=1}^{N_{cases}} \frac{\sum_{j=1,i\neq j}^{N_{cases}} w_{cases}(i,j,r)}{\sum_{j=1}^{N_{controls}} w_{controls}(i,j,r)} \Bigg/ \frac{W_{Cases}(N_{cases} - 1)}{W_{controls}} \qquad \textbf{(26)}$$

Note that this holds if the weight of the controls is proportional to the weight of the neighbours anywhere in the studied area.

5.5 Significance

The null hypothesis to compare the *M* function with is, like before, a random distribution of points. However, space is no longer homogeneous, so the homogeneous Poisson process is no longer appropriate. The first-order property must be controlled for to allow the detection of the second-order property of the process. Thus, a point distribution generated according to the null hypothesis must respect, on the one hand, the first-order property (local values of the density) of the process the point distribution is a realisation of, and, on the other hand, its points must be distributed independently from each other.

The practical difficulty comes from the lack of knowledge of the point process that gave the point distribution, which is its unique available realisation. Its first-order property is consequently widely unknown. We can only assume that the actual set of point locations is a good approximation of it. Consequently, we will generate random data sets by randomly distributing the actual points set (type and weight couples) on the actual location set (coordinates). The confidence interval of the null hypothesis will then be computed by the Monte Carlo technique, as explained above.

5.6 Comparison with K_{inhom} (Baddeley et al. (2000))

K_{inhom} is a generalisation of Ripley's K to non stationary processes. Stationarity is required for the second-order property: this property is understated as soon as interactions between points are evaluated at a given distance.

Definition

K_{inhom} can be defined as an integration of the radial distribution function g on the circle of radius r., equation (9). When the density is not a constant, the result is (E denotes expectation and $\lambda(i)$ is the process density at the point i):

$$K_{in\,\mathrm{hom}}(r) = \frac{1}{A} E\left[\sum_{i=1}^{N} \sum_{j=1, i\neq j}^{N} \frac{c(i,j,r)}{\lambda(i)\lambda(j)} \right] \quad \textbf{(27)}$$

It can be estimated by:

$$\hat{K}_{in\,\mathrm{hom}}(r) = \frac{1}{A} \sum_{i=1}^{N} \sum_{j=1, i\neq j}^{N} \frac{c(i,j,r)}{\lambda(i)\lambda(j)} \quad \textbf{(28)}$$

The indicator $c(i,j,r)$ is then corrected by Ripley's method.
From a theoretical point of view, the problem is perfectly solved. Yet, applications are not straightforward. The difficulty arises in the estimation of the local densities. The natural solution consists in using a kernel estimation (Diggle (1985), Silverman (1986)).

Discussion

The authors mention a severe bias in K_{inhom}'s estimation when they apply this method to an aggregated process. The reason is quite clear: in the aggregates, the observed density is greater than the actual density of the process. It includes the effects of the aggregation process. The authors propose an improved technique for inhomogeneous Poisson processes but do not treat the other cases, including segregated processes.
Practically, the K_{inhom} computing software was developed by Baddeley under *R*. Its inputs are the point set and the associated local densities, which must be pre-processed.

The M function is also a generalisation of K, by a different approach. It compares the number of neighbours to that of all points around each case. This reference is analogous to $\lambda(j)$. We can ignore the question of K's bias, not taken into account by Baddeley, and consider a point set large enough for the denominator of M to be constant. Then, M can be rewritten in a formally close way to K_{inhom}: $M(r) = k \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, i\neq j}^{N} \frac{c(i,j,r)}{\lambda(j)}$, where k is a constant for a

given radius, including the denominator of *M*, the area of the circle of radius *r* and the domain area). Both functions are quite close, but with a few noticeable differences:

- K_{inhom} ignores the individual weights. The limit can not be by-passed considering each *w* weighted individual as a superposition of *w* points, or aggregation will be dramatically overestimated.
- The issue of the local density estimation is solved by *M* considering it as a constant within the circle of radius *r* around each point and computing it as simply as possible, counting all control points. If the cases are aggregated, the estimated density will not be biased.
- The indicator does not need to be corrected from edge effects when computing *M*. This is a decisive advantage to treat complex shapes such as country boundaries.
- Both functions are the average of reference point (the centres of the circles) values. Their weight is the same for all when computing *M* whereas points are weighted by the inverse of the local density $\lambda(i)$.in K_{inhom}. If the purpose is to evaluate the process properties, the *g* function for example, *M* overweights the points in dense areas. On the other hand, if one tries to characterize individual behaviours such as location choice, giving each individual the same weights seems more appropriate.

Conclusion

The function developped by Baddeley *et al.* (2000) constitutes a theoretical milestone in the effort for characterising non homogeneous point processes. However, as far as we know, it was never used in the empirical literature. Its fundamental issue is the great difficulty, both theoretical and practical, to estimate local densities. The *M* function keeps the real advantage to be easily tractable.

6 Examples

6.1 Theoretical examples

Three examples are given. Two of them illustrate very simple point patterns on a homogeneous space for a comparison of *L* and *M* functions. The third one computes a non-homogeneous, independent point process to show how the *M* function controls for the first-order property of point processes. No theoretical example is given with weighted points because they are not so easy to understand visually. Confidence intervals are computed at a 1% confidence level generated from 1000 simulations and all curves are computed at 0.1 intervals.

Aggregates

We consider a point set of three different kinds. The first two subsets (squares and circles) are made of 100 points completely randomly distributed. The last (triangles) is generated by a Neyman-Scott process: 5 aggregates (radius 0.5) of 5 points. Every point weight equals 1. The map is in Figure 1, the curves are in Figure 2.

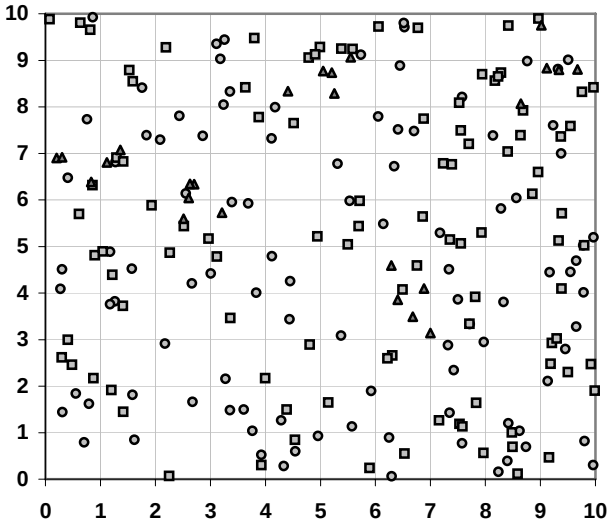


Figure 1: Aggregates, Point map

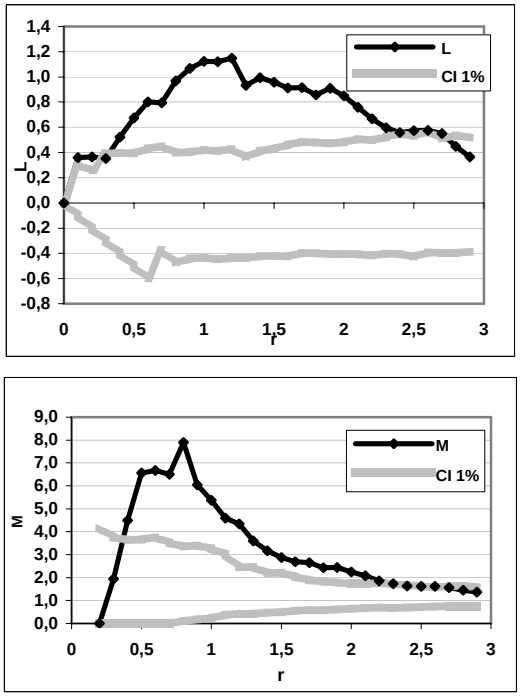


Figure 2: Aggregates, *L* and *M* functions for the aggregated point set

The *M* curve shape is similar to *L*'s: positive peaks denote concentration. Nevertheless, while *L* peaks approximately correspond to the aggregates' diameter (Goreaud (2000)), *M* peaks occur at distances at which the local density is the greatest, that is approximately the distance between points in the aggregates.

We consider a point set made of three different point types. The first two of them (squares and circles) are constituted of 100 completely randomly distributed points. The last one (triangles) is a perfectly even distribution of 100 points located on a square, 1 by 1 grid. All points' weights equal 1.

The first part of the *M* curve is made of 0 values, showing the absence of neighbours at any distance smaller than the grid size. Note that the *L* curve shape is different since its original value is 0 and its minimum slope is -1 by construction.

At the grid size, *M* value suddenly increases (the curve continuity is actually an artefact due to interpolation between points). It decreases again between each point-to-point distance ($\sqrt{2} \approx 1{,}44$ is the diagonal length, then 2, $\sqrt{5} \approx 2{,}24$ and so on).

Regularity

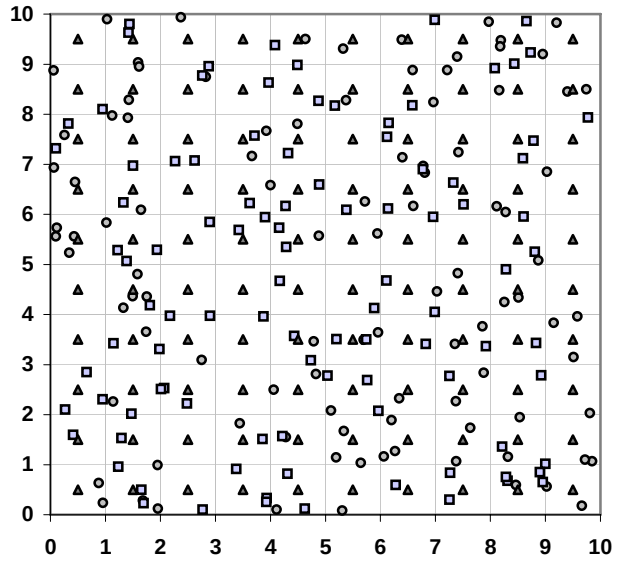


Figure 3: Regular point set, Map point

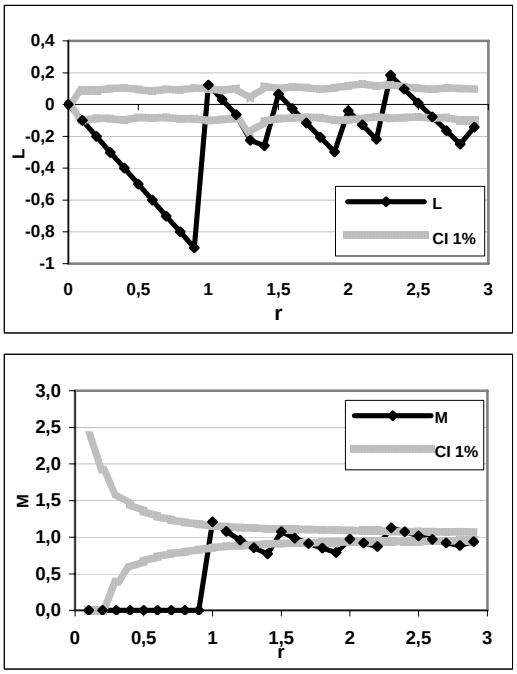


Figure 4: Regular point set, *L* and *M* functions for the regular point set

Inhomogeneous point set

We generated two completely random point sets (squares and circles) in a 10-by-10 domain.

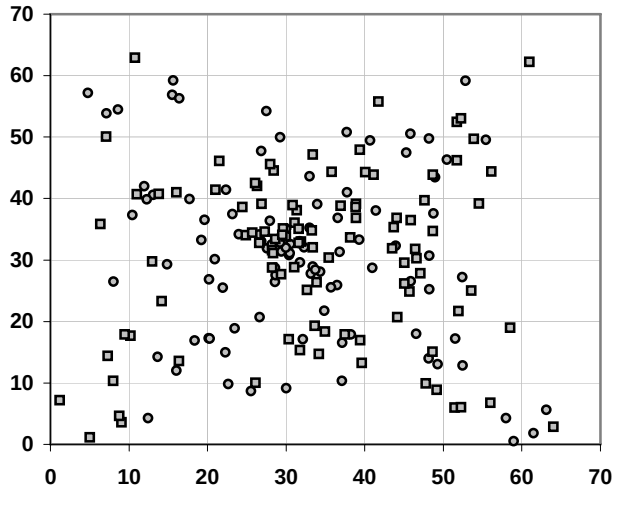


Figure 5: Inhomogeneous point set, Point map

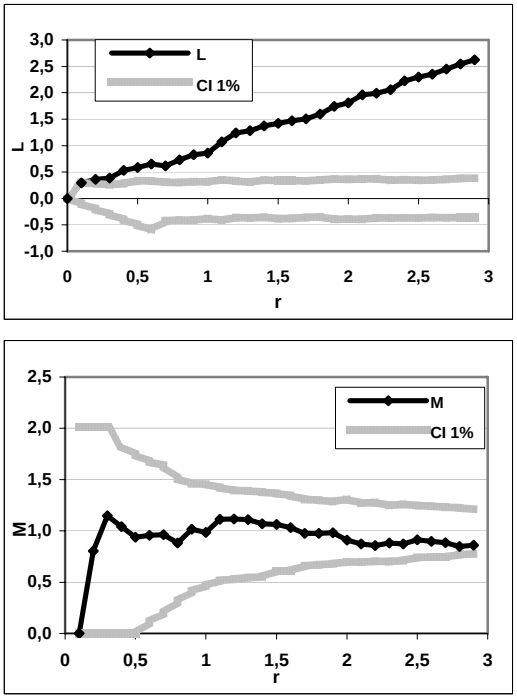


Figure 6: Inhomogeneous point set, *L* and *M* functions

Then, we transformed the points' coordinates: after having calculated the polar coordinates (r,θ) of each point from the centre of the point set, we squared the distance to get (r^2,θ). The result is a non-homogeneous Poisson pattern, in Figure 5. Both point types have the same random distribution, but the centre of the map shows a greater density, this pattern can be compared with a plants distribution around an industrial centre.
The *L* function is not applicable: assuming homogeneity, it will interpret the point distribution as a single big aggregate.
The *M* function is able to control for density variations. Figure 6 shows the *M* values for the first point kind: since its pattern does not differ from the other, its value is around 1.

6.2 Empirical examples

We retained two concrete examples of applications of the *M* function computed from real data in two different fields: spatial economics and geographical epidemiology. Hereinafter, weighted points are considered.

Evaluating the geographic concentration of industries

The first example is taken from Marcon and Puech (2003b)[7]. In this article, the location pattern of French manufacturing firms located in *the whole* metropolitan France in 1996 is studied. The sample is composed of more than 36,000 firms in fourteen sectors of activity. Every firm is weighted by its number of employees. In this case, the *M* function allows measuring the industrial concentration in France for a specific sector (intra-industry concentration).
In every manufacturing sector, significant concentration is detected. However, three main conclusions can be drawn. Firstly, the degree of industrial concentration measured by the *M* function noticeably differs from an industry to another. Secondly, the maximum concentration (significant concentration peak) does not appear at the same distance for each industry (the maximum concentration occurs at small distances, *i.e.*, in a radius of a few kilometres). And finally, the range of distances, on which an over-representation of the sector of activity compared to the whole area is significant, clearly varies across industries.
As an example, Figure 7 illustrates *M* function results for textiles (sector for which the highest peak is detected). The confidence interval is computed at a 5% confidence level from 20 simulations only, due to computing time and considering the clearness of the departure from the null hypothesis. Significant concentration is observed up to 200 kilometres. The peak reaches around 6.5 at a radius less than 1 kilometre.

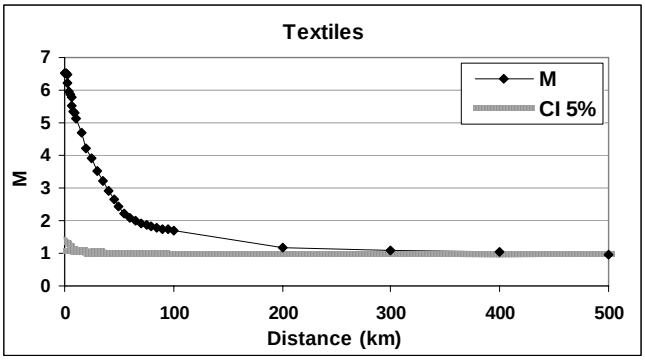


Figure 7: *M* function for textiles in France

[7] This example is not available in the paper itself, but on the authors' website as complementary results.

This indicates that at this radius, the relative density of employees in textiles is more than six times greater around textiles firms than in the whole area. It is worth noting that, in spatial economics, using distance-based methods like Ripley's *K* is quite new and cluster-based methods are more widely employed to evaluate the spatial agglomeration of the economic activity.

Cuzick and Edwards (1990) data set

Cuzick and Edwards (1990) introduced the first formal way to deal with non-homogeneous point processes. They used a data set (published with the paper) concerning the location of 62 cases of childhood leukaemia between 1974 and 1986 in the North Humberside area, England. A control set of 141 children representing the whole concerned population was chosen from the birth register. They could conclude that the cases were significantly clumped. We use this data set to go further. We are now able to corroborate their conclusion and also to precise the size of the aggregates.

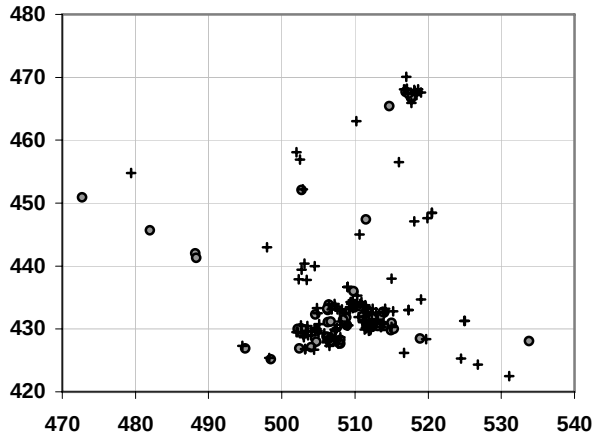


Figure 8: Childhood Leukemia map

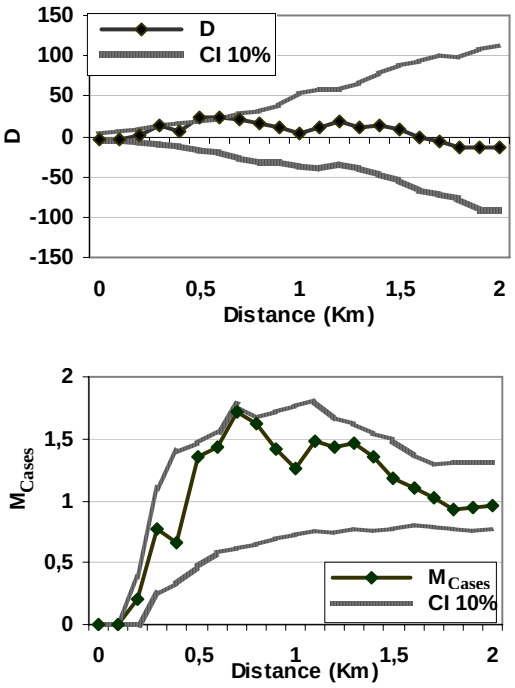


Figure 9: Cuzick and Edwards (1990) point set, *D* and *M* functions

The map is in Figure 8, cases are represented by circles and controls by crosses. Figure 9 shows *M* values for the cases. Note that it was computed according to the case-control design, equation (26). We can confirm clumping and precise it: in a 0.7 km radius around a case, the average case density is about 70% higher than it would be if the cases followed the control pattern (at this distance, the peak of the *M* function reaches 1.7).

In the discussion of Cuzick and Edwards (1990), Diggle (p.101) suggested that the *D* function, equal to K_{cases}-$K_{controls}$, would lead better results. The next year Diggle and Chetwynd (1991) published the mathematical framework of the *D* function and computed their new function on the same dataset. In figure 9, we recomputed *D* (considering the

rectangle domain shown in Figure 8[8]) and estimated the *M* function. It can be seen that the *M* and *D* functions give the same results if points are not weighted. Nevertheless, *D* values can not be interpreted easily and not compared across distances.
Both methods suffer here from a severe lack of power due to the very little number of controls. The confidence intervals are computed at 5% and 10% levels (from 1000 simulations). Increasing the number of controls would not have been a real problem if the experimental design had included a distance-based point pattern analyse.

7 Conclusion

The *M* function is defined as a generalization of Ripley's *K* function to allow its application to inhomogeneous point processes and to take into account point weights.
We had to reformulate the *K* function to understand it as a probability ratio, and by the way correct a bias remained in its definition despite occasional attempts to eliminate it. We also had to choose a definitive edge-effect correction method to make the whole theory consistent.
The probabilistic approach allows considering spatial heterogeneity. When using the *K* function, we know, or at least we hope, that the point process is stationary, *i.e.* the probability to find a neighbour scales with the area. However, using the *M* function, we suppose that the probability to find a neighbour of the good kind is given by the average proportion of good-kind neighbours combined with the local density of points. This assumption is very general and holds in most cases. Yet, this is an assumption and must be clearly kept in mind.
We think this is a significant improvement for spatial structure analysis:

- First of all because the number of situations in which the spatial structure can be analysed will dramatically increase (unfortunately, inhomogeneous point processes are not uncommon) if we compare it to the possible applications of *K*.
- *M* is more powerful than *D* because it does not ignore a part of the data.
- *M* is more convenient to use than *K* because no edge-effect correction is required. More than this, the domain limits do not have to be known, the point locations are enough. Therefore, complete geographical data sets can be treated without simplifying the domain shape and eliminating many border points.
- *M* does not require a good knowledge of the underlying point process and a pre-computation of local densities like K_{inhom} does.
- Neither *K* nor *D* nor K_{inhom} take into account the points' weight.

To allow effective use of the *M* function, we developed the necessary software, available on the authors' web site[9].

[8] Note that this data set was widely used and gave slightly different results according to the domain definition in Diggle and Chetwynd (1991), p. 1160, or Rowlingson and Diggle (1993), p. 634

[9] http://e.marcon.free.fr/Ripley/ (English, French and Italian versions).

References

Baddeley, A. J., Møller, J. and Waagepetersen, R. (2000). "Non- and semi-parametric estimation of interaction in inhomogeneous point patterns." *Statistica Neerlandica* **54**(3): 329-350.

Besag, J. E. (1977). "Comments on Ripley's paper." *Journal of the Royal Statistical Society* **B 39**(2): 193-195.

Clark, P. J. and Evans, F. C. (1954). "Distance to nearest neighbor as a measure of spatial relationships in populations." *Ecology* **35**(4): 445-453.

Cressie, N. A. (1993). *Statistics for spatial data*. John Wiley & Sons, New York. 900 p.

Cuzick, J. and Edwards, R. (1990). "Spatial Clustering for Inhomogeneous Populations." *Journal of the Royal Statistical Society* **B 52**(1): 73-104.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London. 148 p.

Diggle, P. J. (1985). "A Kernel Method for Smoothing Point Process Data." *Applied Statistics* **34**(2): 138-147.

Diggle, P. J. and Chetwynd, A. G. (1991). "Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations." *Biometrics* **47**: 1155-1163.

Feser, E. J. and Sweeney, S. H. (2000). "A test for the coincident economic and spatial clustering of business enterprises." *Journal of Geographical Systems* **2**(4): 349-373.

Gatrell, A. C. and Bailey, T. C. (1996). "Interactive Spatial Data Analysis in Medical Geography." *Social Science & Medicine* **42**(6): 843-855.

Gatrell, A. C., Bailey, T. C., Diggle, P. J. and Rowlingson, B. S. (1996). "Spatial point pattern analysis and its application in geographical epidemiology." *Transactions of the Institute of British Geographers* **21**: 256-274.

Getis, A. (1984). "Interaction modeling using second-order analysis." *Environment and Planning A.* **16**: 173-183.

Getis, A. and Franklin, J. (1987). "Second-order neighborhood analysis of mapped point patterns." *Ecology* **68**: 473-477.

Goreaud, F. (2000). *Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes*. Thèse de doctorat, ENGREF. Nancy.

Goreaud, F. and Pélissier, R. (1999). "On explicit formulas of edge-effect correction for Ripley's K-function." *Journal of Vegetation Science* **10**(3): 433-438.

Haase, P. (1995). "Spatial pattern analysis in ecology based on Ripley's K function: Introduction and methods of edge correction." *Journal of Vegetation Science* **6**(4): 575-582.

Jones, A. P., Langford, I. H. and Bentham, G. (1996). "The Application of K-Function Analysis to the Geographical Distribution of Road Traffic Accident Outcomes in Norfolk, England." *Social Science & Medicine* **42**(6): 879-885.

Kingham, S. P., Gatrell, A. C. and Rowlingson, B. S. (1995). "Testing for Clustering of Health Events within a Geographical Information System Framework." *Environment and Planning A.* **27**(5): 809-821.

Kuuluvainen, T. and Rouvinen, S. (2000). "Post-fire understorey regeneration in boreal *Pinus sylvestris* forest sites with different fire histories." *Journal of Vegetation Science* **11**(6): 801-812.

Marcon, E. and Puech, F. (2003a). "Evaluating the Geographic Concentration of Industries Using Distance-Based Methods." *Journal of Economic Geography* **3**(4): 409-428.

Marcon, E. and Puech, F. (2003b). *Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods*. Cahiers de la MSE, **2003.18:** 22 p.

Moeur, M. (1993). "Characterizing spatial patterns of trees using stem-mapped data." *Forest Science* **39**(4): 756-775.

Pancer-Koteja, E., Szwagrzyk, J. and Bodziarczyk, J. (1998). "Small-scale spatial pattern and size structure of *Rubus hirtus* in a canopy gap." *Journal of Vegetation Science* **9**(6): 755-762.

Peterson, C. J. and Squiers, E. R. (1995). "An Unexpected Change in Spatial Pattern Across 10 Years in an Aspen-White Pine Forest." *Journal of Ecology* **83**(5): 847-855.

Ripley, B. D. (1976). "The Second-Order Analysis of Stationary Point Processes." *Journal of Applied Probability* **13**: 255-266.

Ripley, B. D. (1977). "Modelling Spatial Patterns." *Journal of the Royal Statistical Society* **B 39**(2): 172-212.

Ripley, B. D. (1979). "Tests of 'randomness' for spatial point patterns." *Journal of the Royal Statistical Society* **B 41**(3): 368-374.

Ripley, B. D. (1981). *Spatial statistics*. John Wiley & Sons, New York. 255 p.

Rowlingson, B. S. and Diggle, P. J. (1993). "SPLANCS: Spatial Point Pattern Analysis Code in S-Plus." *Computers & Geosciences* **19**(5): 627-655.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall. 175 p.

Stoyan, D., Kendal, W. S. and Mecke, J. (1987). *Stochastic Geometry and its Applications*. John Wiley & Sons, New York. 345 p.

Sweeney, S. H. and Feser, E. J. (1998). "Plant Size and Clustering of Manufacturing Activity." *Geographical Analysis* **30**(1): 45-64.

Szwagrzyk, J. and Czerwczak, M. (1993). "Spatial patterns of trees in natural forests of East-Central Europe." *Journal of Vegetation Science* **4**(4): 469-476.

Thioulouse, J., Chessel, D., Dolédec, S. and Olivier, J.-M. (1997). "ADE-4: a multivariate analysis and graphical display software." *Statistics and Computing* **7**(1): 75-83.

Upton, G. J. G. and Fingleton, B. (1985). *Spatial Data Analysis by Example, volume 1: Point Pattern and Quantitative Data*. John Wiley & Sons, New York. 410 p.

Ward, J. S. and Ferrandino, F. J. (1999). "New derivation reduces bias and increases power of Ripley's L index." *Ecological Modelling* **116**(2-3): 225-236.