



LA MÉTHODE DE RIPLEY POUR L'ANALYSE DES STRUCTURES SPATIALES PONCTUELLES EN ÉCOLOGIE

© Jean-Michel Noël Walter
Université Louis Pasteur, Institut de Botanique
DEA de Géographie Physique et Aménagement
Strasbourg, 2000–2001
Mise à jour 2006

SOMMAIRE

INTRODUCTION	3
THÉORIE	4
RÉFÉRENCES	10
REMERCIEMENTS	10

INTRODUCTION

La présente notice a pour objectif d'introduire à l'utilisation de la méthode de Ripley afin d'analyser des structures spatiales de type « semis de points ». De tels semis de points peuvent représenter n'importe quel inventaire de populations d'êtres vivants ou d'objets dans un espace bi-dimensionnel. Ces populations peuvent être caractérisées par un attribut comme la couleur, l'espèce, etc., dont les individus sont représentés par leurs coordonnées x, y .

La méthode de Ripley fait partie d'une approche déjà ancienne basée sur les distances entre les individus. Elle permet d'analyser des structures complexes hiérarchisées qu'il n'était guère possible d'envisager avec autant d'efficacité et de possibilités d'interprétation avec d'autres méthodes. La méthode de Ripley s'est surtout développée dans les années 80, mais elle s'est enrichie plus récemment de la réflexion de nombreux chercheurs. En France, particulièrement, ce sont des écologistes forestiers qui en Inde, Guyane et France ont approfondi cette approche en la développant dans ses aspects théoriques comme dans l'interprétation biologique des analyses qu'elle permet. C'est donc sur ces travaux, qui au départ étaient des mémoires de DEA et des thèses, que s'appuie cette notice. Celle-ci ne fait qu'en résumer les principes premiers, afin de permettre à l'utilisateur novice de s'initier à cette méthode. C'est le logiciel ADE-4, développé à l'Université Claude Bernard de Lyon, qui est proposé pour les calculs.

Trois parties aborderont certains aspects théoriques, fourniront quelques exemples d'applications tirés de l'écologie forestière et, enfin, proposeront une étude de cas sous la forme d'un exercice concret. La bibliographie est limitée aux travaux des chercheurs français qui ont beaucoup contribué au développement de cette approche.



THÉORIE

Supposons que le jeu de données à notre disposition soit un « semis de points » sur un espace rectangulaire. Chaque point représente la position du centre d'une plante, d'un arbre, selon ses coordonnées (x, y) , l'espace rectangulaire étant la fenêtre utilisée dans l'analyse. À chaque point peut être associé un ensemble d'attributs quantitatifs (diamètre du tronc, hauteur de la couronne, etc.) et qualitatifs (taxon, modèle architectural, tolérance à l'ombre, vitalité, etc.). Ce type de données convient à un traitement statistique spatial utilisant la méthode de Ripley.

Cette approche a été appliquée aux données de dispositifs forestiers permanents de Guyane française (Paracou) et d'ailleurs (France ; Uppangala, Ghâts Occidentaux de l'Inde). Elle a été largement développée dans un certain nombre de thèses et de mémoires récents (Goreaud, 1995, 2000 ; Péliissier, 1995 ; Collinet, 1997 ; Gourlet-Fleury, 1997 ; Picard, 1999). Ces auteurs ont présenté en détail les bases théoriques de la méthode, accompagnées d'une abondante bibliographie. Ils ont ainsi pu proposer des interprétations biologiques aux structures spatiales observées, sur des bases quantitatives rigoureuses.

Il est clair que le motif de dispersion spatiale des individus dépend d'une multitude de facteurs : les uns exogènes, de nature physico-chimique, topographique, climatique, les autres endogènes, de nature biologique, liés à la compétition, à la reproduction, aux relations entre plantes et animaux, à la régénération des espèces (Chessel, 1978). Comme les motifs de dispersion spatiale observés (Figure 1) sont l'expression d'interactions complexes de tous ces facteurs à diverses échelles d'espace et de temps, la première tâche est d'établir, autant que possible, la part qui revient aux facteurs exogènes et endogènes.

L'approche de Ripley appartient à la classe des *méthodes de distances* entre individus ('point pattern analysis', littéralement « analyse des motifs, ou patrons, ponctuels »), méthodes déjà anciennes, puisqu'elles remontent au début du XX^e siècle (Urban, 2000). Ainsi, les méthodes d'*analyse de voisinage de premier ordre*, basées sur les distances au plus proche voisin, permettent-elles de détecter des motifs à échelle unique. Elles possèdent cependant des inconvénients : a) perte d'information sur les distances individuelles point par point, l'information ponctuelle étant noyée dans la moyenne, b) inadéquation pour détecter des structures complexes constituées de plusieurs motifs emboîtés à différentes échelles, c) manque d'élaboration numérique, de sorte que ces méthodes ne peuvent indiquer que la direction de l'écart à la distribution aléatoire, soit contagion, soit régularité. Par contre, les méthodes d'*analyse de voisinage de second ordre*, qui considèrent la distribution de toutes les distances point par point, sont à même de détecter des structures complexes et d'offrir des outils efficaces pour l'interprétation.

La fonction K de Ripley est utilisée pour comparer le motif ponctuel observé avec un motif ponctuel engendré par un processus connu, par exemple un processus de Poisson (aléatoire) homogène. Elle est la distribution cumulative des fréquences d'observations à une distance donnée point par point, ou à l'intérieur d'une classe de distances. C'est parce qu'elle préserve les distances à toutes les échelles que la statistique de Ripley peut quantifier l'intensité du motif à de multiples échelles.

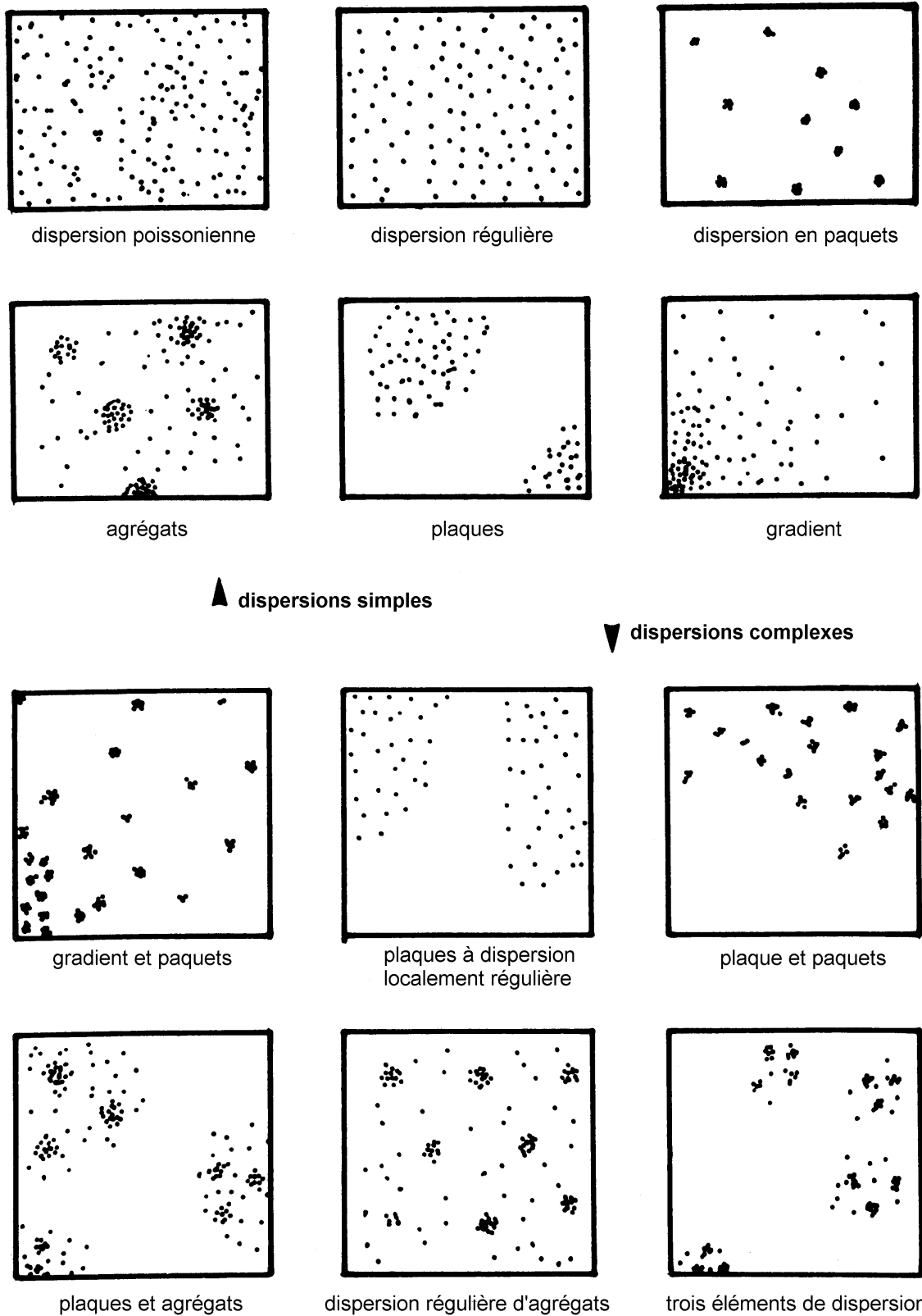


Figure 1. Typologie de quelques motifs de dispersion spatiale. D'après Chessel, 1978.

La méthode de Ripley (Urban, 2000) est basée sur le nombre de points (« semis de points », ensemble des individus, ou des arbres d'un peuplement, ou de n'importe quels objets) recensés à une certaine distance ou classe de distance :

$$K(r) = \lambda^{-1}E(r) \quad (1)$$

Où $E(r)$ est l'espérance du nombre de points à l'intérieur d'une distance r d'un point quelconque du semis de points. L'intensité moyenne des points λ peut être estimée par leur densité n/A , où n est le nombre total de points et A la surface totale échantillonnée.

La distribution cumulative $E(r)$ est estimée empiriquement, ce qui demande un inventaire exhaustif des points et non pas des échantillons dispersés de points. La statistique $K(r)$ de Ripley peut être estimée comme suit :

$$\hat{K}(r) = \lambda^{-1} \sum_{i=1}^n \sum_{j=1}^n (\delta_{ij} (\|x_i - x_j\| \leq r)) / n \quad (2)$$

Pour tout $i \neq j$, où δ_{ij} est une *fonction indicatrice* qui prend la valeur de 1 si la condition est vraie, sinon elle prend la valeur de 0. Ici, la condition est que la distance euclidienne entre les points soit $\leq r$.

Autrement dit, $K(r)$ est un recensement de la proportion des points qui tombent à l'intérieur de chaque classe de distances. L'inverse de λ dans l'équation convertit K en unités de surface.

Des effets de bordure peuvent sérieusement dégrader les statistiques basées sur les distances. Une approche due à Ripley permet une correction géométrique de l'effet de bordure (Figure 2). Dans ce cas, δ_{ij} prend des valeurs différentes de 1 dans la sommation de l'équation (2).

Le semis de points observé est engendré par un processus spatial ponctuel, le plus souvent inconnu, dont il n'est qu'une réalisation parmi d'autres dans l'espace de la parcelle (Collinet, 1997). Son analyse impose de faire les hypothèses de *simplicité* (il ne peut y avoir plus d'un arbre au même endroit), d'*homogénéité* ou de *stationnarité* (invariance par translation : les arbres ont la possibilité de se localiser sur toute la surface étudiée) et d'*isotropie* (invariance par rotation : absence de direction privilégiée dans les données) du processus spatial. Dans le cas d'une distribution de Poisson (processus poissonnien) d'une population, la valeur attendue de $K(r)$ est : $K(r) = \pi r^2$. Si $K(r) < \pi r^2$, les points sont mis à distance les uns des autres et la distribution est régulière. Si $K(r) > \pi r^2$, la distribution est contagieuse (agrégative). $K(r)$ s'interprète avec l'intensité de la population λ , avec pour le semis de points $\lambda = n/A$.

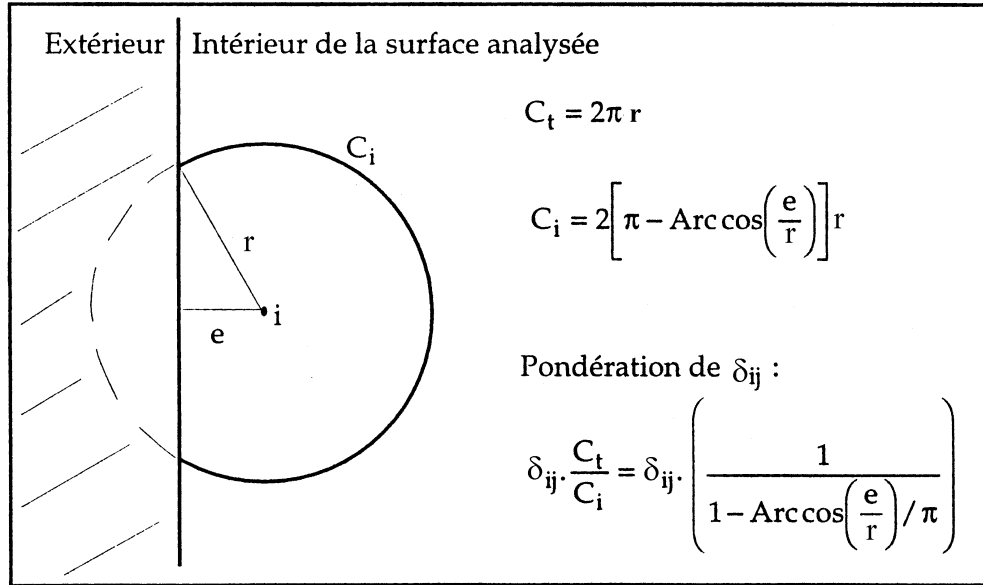


Figure 2. Principe de correction des effets de bordure pour tout point i situé à une distance $e < r$ de l'un des bords. Sous l'hypothèse que la densité des points est identique dans la fraction de cercle de rayon r exclue de la parcelle et dans l'intérieur de la surface testée, la correction consiste à pondérer, pour tout point i considéré, la valeur δ_{ij} par le rapport entre le périmètre total et l'arc de cercle inclus dans la parcelle. D'après Collinet (1997), Goreaud et Pélissier (1999), modifié.

Comme indiqué plus haut, $\lambda K(r)$ est l'espérance $E(r)$ du nombre de points distants d'une longueur $\leq r$ d'un point quelconque du semis. En pratique, on utilise un estimateur linéarisé de $K(r)$:

$$\hat{L}(r) = \sqrt{\hat{K}(r)/\pi} - r \quad (3)$$

Cette valeur s'annule pour une distribution aléatoire. Elle est positive pour une distribution contagieuse et négative pour une distribution régulière. Représentée graphiquement (Figure 3), cette transformation est plus facile à interpréter visuellement (Besag, 1977, *in* auteurs cités).

Pour tester l'hypothèse H_0 d'une distribution de Poisson, des intervalles de confiance locaux sont calculés, par exemple à un seuil de 5% pour chaque distance r , avec la même intensité que pour le motif observé. On utilise pour cela la méthode de Monte Carlo (Besag and Diggle, 1977, *in* Pélissier, 1995). Pour un semis de n points, on génère S fois une série aléatoire de n couples (x, y) . À chaque pas de distance dr , les valeurs de $\hat{L}(r)$ sont calculées pour les S séries et ordonnées de façon croissante. Au risque de $2k/S$, les $k^{\text{ème}}$ et $(S - k)^{\text{ème}}$ valeurs simulées de $\hat{L}(r)$ fourniront les limites supérieures et inférieures de l'intervalle de confiance. Les $5^{\text{ème}}$ et $95^{\text{ème}}$ valeurs simulées de $\hat{L}(r)$ définiront les limites de l'enveloppe de confiance à 90% de H_0 à la distance r . La précision des limites de l'intervalle de confiance ainsi que la qualité de lissage de l'enveloppe dépendent du nombre d'individus n et du nombre de simulations S (Collinet, 1997).

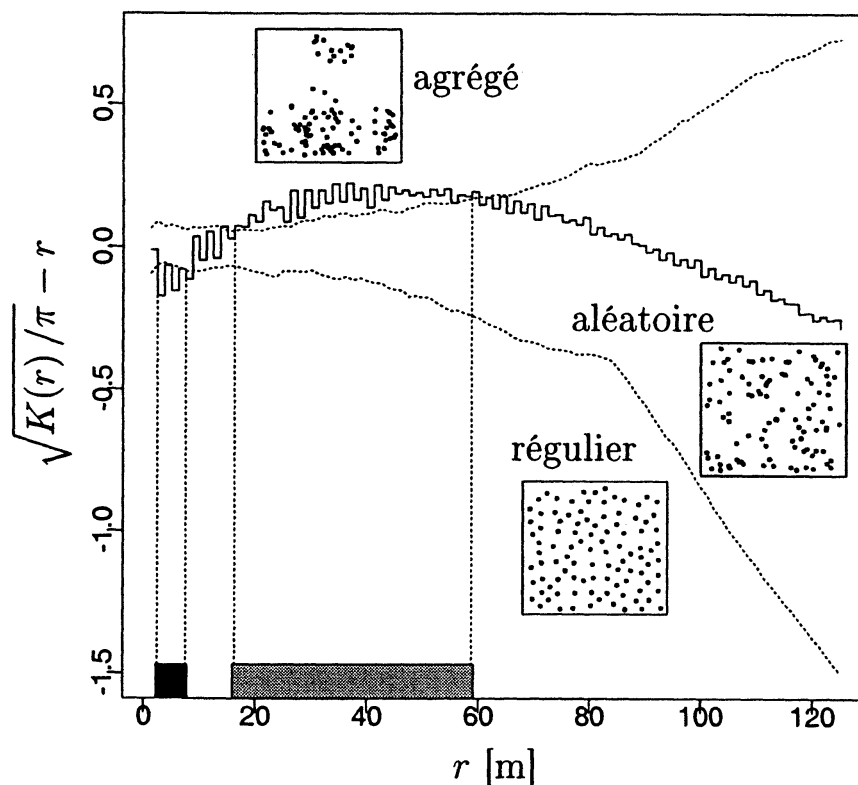


Figure 3. Fonction K de Ripley (trait continu) et enveloppes de 100 réalisations (trait pointillé) d'un processus de Poisson de même intensité. La fonction $\hat{L}(r) = \sqrt{\hat{K}(r)}/\pi - r$ est représentée. La bande sous-jacente indique les distances auxquelles un écart significatif à la répartition aléatoire se produit. D'après Picard (1999). En noir : régularité ; en gris : agrégation.

En général, il est recommandé que n soit > 30 , et que S soit au minimum 100, de préférence 1000. Compte tenu des effets de bordure, il est conseillé de ne calculer $\hat{L}(r)$ que pour des valeurs de r au plus égales à la moitié de la largeur la plus petite de la parcelle (Ripley, 1977, *in* Collinet, 1997). Cependant, de nouvelles formules de correction permettent d'étendre le champ de l'analyse sans perte de précision (Goreaud and Pélissier, 1999). En tout état de cause, il n'est guère recommandé d'utiliser l'approche de Ripley dans les situations de gradients ou lorsque la surface inventoriée est trop exiguë.

Jusqu'ici n'était considéré que le cas *univariable* de la fonction de Ripley, c'est-à-dire ne concernant que les coordonnées cartésiennes de position des individus d'une espèce donnée, ou d'autres catégories, comme indiqué dans l'introduction. Ces données étiquetées, ou affectées d'un label, sont appelées « motifs ponctuels marqués » ('marked point patterns', Urban, 2000). On définit ainsi des *processus ponctuels marqués*.

Lorsque deux catégories sont prises en compte, on se réfère à des processus ponctuels bimarqués (Picard, 1999). Cette approche *bivariable* (bivariée), ou *analyse intertype* de deux catégories (Diggle, 1983, *in* Pélissier, 1995), est une extension importante de l'approche de Ripley, qui a été particulièrement développée par Pélissier (1995), Goreaud (1995, 2000) et Picard (1999). Elle consiste à considérer les distances entre les individus des différents types. Elle permet, par exemple, de tester l'hypothèse de répartition aléatoire des étiquettes des catégories conditionnellement à une structure fixée (dans une population, la présence d'arbres

morts est-elle liée à la présence d'arbres vivants ?), ou de tester l'indépendance des sous-populations (par exemple, est-ce que la répartition spatiale d'une espèce dépend de la répartition spatiale d'une autre espèce ?). Un traitement alternatif de l'intervalle de confiance local peut se présenter pour les catégories considérées. Dans le premier cas, la substitution ('random labelling') consiste à permuter aléatoirement les étiquettes des catégories en gardant la même structure spatiale globale. Dans le second cas, le décalage ('random shifting') consiste à décaler d'un vecteur aléatoire l'une des catégories par rapport à l'autre en conservant la structure interne de chacune d'elles. Ces applications sont d'un grand intérêt général et peuvent être mises en œuvre en utilisant une extension remarquablement simple de la fonction de Ripley.

En effet, en rappelant que $\lambda = n/A$, et en réarrangeant l'équation (2) :

$$K(r) = \frac{A}{n^2} \sum_{i=1}^n \sum_{j=1}^n \delta_{ij}(r_{ij}) \quad (4)$$

Où la fonction indicatrice a été simplifiée mais garde cependant sa signification. Le cas bivariable du K de Ripley devient alors, pour des semis de points de classe 1 et 2 :

$$K_{12}(r) = \frac{A}{n_1 n_2} \sum_{i=1}^n \sum_{j=1}^n \delta_{ij}(r_{ij}) \quad (5)$$

Où la fonction indicatrice δ_{ij} est vraie (=1) si deux points i et j sont dans les limites d'une classe de distances et sont de types 1 et 2, respectivement ; n_1 et n_2 sont les tailles des sous-échantillons des deux groupes. Comme dans le cas univariable, la fonction indicatrice doit être pondérée pour tenir compte des effets de bordure.

L'interprétation de $\hat{L}_{12}(r)$ est basée sur les mêmes principes généraux que pour le cas univariable (équation (3)). On considère que dans les limites de l'enveloppe de confiance, le rapport spatial qui s'établit entre les deux catégories relève de processus purement aléatoire, autrement dit il n'y a *pas de dépendance*, de ségrégation spatiale catégorielle entre elles. Lorsque la courbe $\hat{L}_{12}(r)$ dépasse la limite supérieure de l'enveloppe de confiance, l'agrégation des semis de points devient significative, il y a *contagion* entre les deux catégories. Au contraire, lorsque la courbe $\hat{L}_{12}(r)$ dépasse la limite inférieure de l'enveloppe de confiance, il y a mise à distance des points et la répartition des semis est significativement régulière. Cette *régularité* est habituellement interprétée comme un effet de compétition, ou de répulsion, entre les individus. D'autres développements de l'analyse spatiale de Ripley ont été présentés ailleurs. Le lecteur intéressé trouvera des applications dans Haase (1995), Pélissier (1995), Goreaud et Pélissier (1999, et *in Thioulouse et al.*, 1997), Picard (1999), Goreaud (2000), Wiegand and Moloney (2004).



RÉFÉRENCES

Urban D (2000) *Spatial Analysis in Ecology*. Duke-Nicholas School of the Environment, Levine Science Research Center, Duke University Box 90328, Durham, NC 27708-0328, USA.

<http://www.env.duke.edu/landscape/>

Chessel D (1978) Description non-paramétrique de la dispersion spatiale des individus d'une espèce. In : Legay J.M. et Tomassone R. (Coordinateurs). *Biométrie et Écologie*. Société Française de Biométrie, p. 45–135.

Collinet F (1997) Essai de regroupement des principales espèces structurantes d'une forêt dense humide d'après l'analyse de leur répartition spatiale (Forêt de Paracou – Guyane). Thèse de Doctorat, Université Claude Bernard, Lyon I.

Goreaud F (1995) Étude et modélisation des peuplements hétérogènes : rôle des interactions dans la structuration spatiale du peuplement. Mémoire de DEA, Université Pierre et Marie Curie, Paris 6.

Goreaud F (2000) Apports de la structure spatiale en forêt tempérée à l'étude et la modélisation des peuplements complexes. Thèse ENGREF, Nancy.

Goreaud F, Pélissier R (1999) On explicit formulas of edge effect correction for Ripley's K-function. *Journal of Vegetation Science* 10, 433–438.

Gourlet-Fleury S (1997) Modélisation individuelle spatialement explicite de la dynamique d'un peuplement de forêt dense tropicale humide (dispositif de Paracou – Guyane française). Thèse de Doctorat, Université Claude Bernard Lyon I.

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.

Pélissier R (1995) Relations entre l'hétérogénéité spatiale et la dynamique de renouvellement d'une forêt dense humide sempervirente (Forêt d'Uppangala – Ghâts Occidentaux de l'Inde). Université Claude Bernard, Lyon I.

Picard N (1999) Passage d'un modèle individuel à un modèle de distribution de la dynamique forestière. Application à une forêt dense tropicale humide de Guyane française. Thèse de Doctorat, École Nationale du Génie Rural, des Eaux et des Forêts, Nancy.

Thioulouse J, Chessel D, Doledec S, Olivier J-M (1997) ADE-4: a multivariate analysis and graphical display software. *Statistical Computing* 7, 75–83. [Programmes ADSUtil, RIPLEY, DENSITY1, DENSITY2, INTERTYPE, PLOTS (Spatial Statistics) de Goreaud et Pélissier. (<http://pbil.univ-lyon1.fr/ADE-4/>). Cette section comprend tous les développements les plus récents sur l'approche de Ripley appliquée à l'écologie].

Wiegand T and Moloney KA (2004) Rings, circles, and null-models for point pattern analysis in ecology. *OIKOS* 104: 209–229.

REMERCIEMENTS

Que Raphaël Pélissier et François Goreaud soient vivement remerciés pour leur aide efficace et leurs programmes.