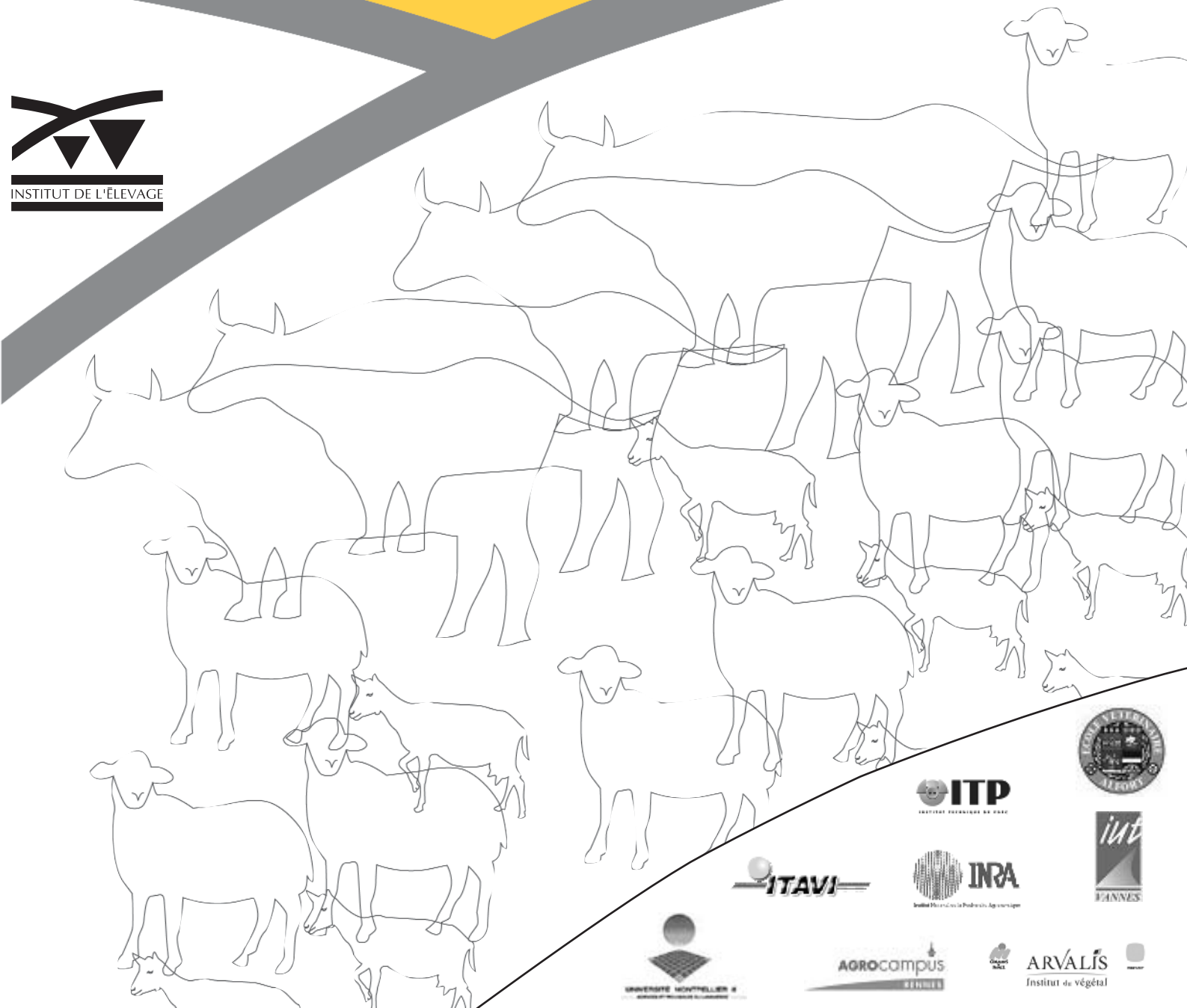
Juin 2005
Compte rendu 220517001
INRA
J. CHADOEUF, F. GOREAUD, J.Y. PONTAILLER et S. SOUBEYRAND

Contribution à l'élaboration de méthodes de statistique spatiale dans le traitement de données agricoles permettant de prendre en compte le contexte géographique et d'améliorer la précision des références fournies aux organismes de développement agricole

Annexe A1.3 : Tests non paramétriques d'indépendance de la répartition d'objets complètement observables distribués dans le plan

Cette étude a été conduite avec le soutien financier de l'enveloppe Recherche ACTA / BCRD du MAAPAR.

collection résultats



INSTITUT DE L'ÉLEVAGE

Juin 2005

Compte-rendu N° 220517001
Carlos LOPEZ
Service Biométrie

Contribution à l'élaboration de méthodes de statistique spatiale dans le traitement de données agricoles permettant de prendre en compte le contexte géographique et d'améliorer la précision des références fournies aux organismes de développement agricole.

(Dossier n°02/01-5 : Traitement statistique des données spatiales
Enveloppe Recherche ACTA/BCRD du MAAPAR)

Annexe A1.3 : Tests non paramétriques d'indépendance de la répartition d'objets complètement observables distribués dans le plan

J. CHADOEUF, F. GOREAUD, J.Y. PONTAILLER, S. SOUBEYRAND (INRA)

Collection Résultats

Tests non-paramétriques d'indépendance de la répartition d'objets complètement observables distribués dans le plan

Chadœuf J[*], Fady B[†], Goreaud F[‡], Pontailler JY[§] & Soubeyrand S[*]

July 6, 2004

travail effectué au sein du projet ACTA 02/01-5.

[*]INRA-Biométrie, Domaine St Paul,84914 Avignon, cedex9

[†]INRA-URFM, Avenue A. Vivaldi,84000 Avignon

[‡]Cemagref, Campus des Cézeaux, 24 Avenue de Landais, BP 50085, 63172 Aubière, cedex 1

[§]ESE / CNRS UMR C8079 (Ecologie, Systematique et Evolution) Batiment 362 - Universite Paris-Sud - F91405 Orsay cedex

Contents

1 Introduction

1.1 Quel objet d'étude ?

Nous nous intéressons dans ce texte à la répartition d'objets répartis plus ou moins au hasard, et ne s'automasquant pas. Dans le plan, deux types d'objets sont alors concernés, les points et les lignes. Le cas le plus simple et le plus étudié est le semis de points. On va le rencontrer dans des domaines très divers. En géographie par exemple, on pourra s'intéresser à la répartition des villes (Burghardt 1959). En écologie, ce sera la répartition d'espèces végétales ou animales que l'on regardera (Erschbamer *et al* 1983). En épidémiologie, on regardera la répartition de présence de maladie (Elliot *et al* 2001), cependant que l'on étudiera la dispositions d'objets célestes en astronomie (Martinez & Saar 2002).

On cherche alors à caractériser la distribution spatiale de ces points pour diverses raisons. Parmi elles on peut citer:

- se faire une représentation simplifiée. Considérons l'exemple d'un semis de points classique, la présence de la grippe en France à une date donnée. La carte brute du semis de points contient toute l'information disponible sur les foyers de la maladie. Elle est cependant peu utilisable telle qu'elle, car soit les points sont trop gros et on ne peut plus mesurer visuellement la densité de points dans les "hot spots", qui sont a priori les foyers les plus actifs, soit les points sont trop fins, et on ne verra plus les petits foyers, alors qu'ils peuvent être a l'origine des futurs hot spots. Avoir alors une représentation simplifiée permet alors une vision plus synthétique de l'extension de la maladie.

- connaître la densité locale de ces objets. En épidémiologie par exemple, savoir où apparaissent les cas les plus fréquents de maladie, à densité de susceptibles égale, permet de guider la recherche des causes et des facteurs de risque d'une maladie.

- caractériser les interactions entre ces objets. Ainsi, dans le cas des villes ou des nids, on cherchera à savoir si ceux-ci sont répartis par groupes, ce qui pourrait traduire un effet coopératif (défense collective par exemple chez les animaux), ou si ils tendent à se repousser, comme cela peut se produire si un territoire exclusif est défini autour de chaque point. Dans cet exemple, il est à noter que la statistique spatiale va analyser

les interactions spatiales entre individus en relation avec des hypothèses sur leur comportement social, mais pas la relation de causalité (le comportement agrégatif est-il une cause ou une conséquence d'une mode de répartition spatial ?).

- améliorer la précision d'estimateurs globaux. Le premier critère d'intérêt d'un processus ponctuel est son intensité, c'est à dire le nombre moyen de points par unité de surface. Une fois passée la première étape (la cartographie de l'intensité locale, c'est à dire le repérage des zone à faible ou forte densité), on peut s'intéresser à une approche plus globale, cherchant à connaître combien de points tombent en moyenne par unité de surface et avec quelle variabilité. L'intensité est estimée très naturellement comme le rapport entre le nombre de points dans une surface et son aire. Plus le semis est régulier et plus il va se rapprocher d'un processus non aléatoire, et la précision de l'estimation devient vite très bonne quand la taille de l'échantillon augmente. Si le semis est au contraire organisé en paquets, il sera nécessaire d'avoir une plus grande surface pour avoir une précision donnée (à même nombre de points par unité de surface) car l'estimateur variera beaucoup selon que les paquets de bordure sont dans la zone comptée, en dehors ou à moitié dedans.

- guider l'échantillonnage. Si la cartographie de toute une zone est relativement facile, le problème de l'échantillonnage se posera simplement en terme de taille de zones à cartographier. Ceci n'est cependant pas toujours facile à faire. Si comme dans le cas précédent, on s'intéresse à l'estimation de l'intensité du semis de points, connaître la distribution spatiale des points, et donc la dépendance spatiale (la corrélation) entre mesures faites entre zones échantillonnées va permettre dans un premier temps d'améliorer l'estimateur par la prise en compte de ces corrélations, puis d'optimiser le plan d'échantillonnage (en cherchant celui rendant la variance minimale, ou égale au niveau souhaité...)

Ces questions ne concernent pas en fait les seuls semis de points, mais les semis de tout type d'objet dans le plan, que ce soit des segments (de droite ou de courbe), des sous-ensembles du plan (semis de taches) (quelle est la longueur moyenne de segments par unité de surface ?)... Plus les objets auront une géométrie complexe et plus les questions que l'on se posera

vis à vis de ces objets pourront se multiplier. Ainsi par exemple, les questions liées à l'orientation viennent naturellement avec les segments (quelle est la l'orientation moyenne d'éléments de segments par unité de surface ?). Nous regarderons par la suite les deux premiers cas (semis de points et de segments), qui offrent l'avantage d'assurer une observation complète dans la zone d'étude, alors que, si les objets sont "opaques", des taches peuvent en masquer d'autres, et amènent à des développements qui débordent du cadre des test de permutation. Nous présenterons d'abord le cas standard des semis de points qui, dans la mesure où ils restent des objets très simples, nous permettront de préciser comment aborder ces questions de tests de permutation dans le cadre spatial. Dans une deuxième partie, tout en gardant le même canevas de travail nous aborderons avec les semis de segments les problèmes posés par l'étude d'objets plus complexes, liés à la présence d'objets partiellement observés car débordant de la zone d'étude.

1.2 Que veut-on faire ?

Dans ce texte, nous nous centrerons sur un aspect de la statistique spatiale, l'utilisation de tests de permutations en exploration de jeux de données:

- dans une première étape, face à un nouveau jeu de données, on cherche d'abord à classer le semis de points en trois grandes catégories: les points sont-ils plutôt regroupés en paquets, tendent-ils plutôt à se repousser, ou n'y a-t-il aucune structure particulière ? C'est l'objet des tests dits CSR (Complete Spatial Randomness) qui seront abordés en première partie.

- Les notions d'indépendance dans le cas de processus ponctuels bivariés ne sont pas toujours bien explicitées en relation avec les objets que ces processus modélisent. A-t-on un processus sur lequel se greffe en chaque point une valeur issue d'une variable aléatoire pouvant prendre deux marques ou a-t-on affaire à deux processus de points ? Ces deux définitions ne sont pas interchangeables, car elles sous-entendent deux types de fonctionnement différents et deux types de tests différents. Nous développerons ce point dans une deuxième partie.

- Le semis n'a pas toujours des propriétés constantes dans l'espace. Il peut présenter des gradients de densité. Si ces gradients sont connus, si par exemple ils sont liés à la distance à une frontière (semis de plantes

hydrophiles observé le long de canaux ou de marais par exemple), il est important de les prendre en compte dans les analyses si on veut éviter des conclusions erronées. Ainsi, dans le cas de l'étude d'un semis de plantes hydrophiles, l'utilisation brutale d'un test CSR conduira à conclure à un processus agrégé car ce qu'on détecte est le regroupement lié au gradient, sans qu'il n'y ait d'interaction entre elles. Dans ces cas, il est alors intéressant de modifier les tests CSR en conséquence. On en verra quelques développements dans le cadre d'un exemple.

1.3 Comment le faire ?

Les analyses statistiques comparent la distribution spatiale du semis de points à ce qu'elle aurait pu être sous différentes hypothèses. Elles représentent donc le semis de points comme la réalisation d'un processus aléatoire, sur lequel différentes hypothèses vont être faites.

Nous allons donc dans un premier temps rappeler ce qu'est un processus ponctuel dans le plan. Les hypothèses de bases généralement formulées pour permettre une approche statistique des processus spatiaux sont des hypothèses de stationnarité, d'isotropie et d'ergodicité. Nous allons dans le même temps regarder comment se déclinent ces hypothèses dans le cas des processus ponctuels.

Les tests que nous présenterons font partie des tests de permutation. Ils sont largement utilisés en statistique spatiale soit pour leur facilité d'utilisation dans le cadre de tests exacts (voir par exemple Peyrard *et al* 2004 pour les tests sur grille régulière), soit parce que le calcul analytique est lourd et difficile hors cas d'école (le cas des processus ponctuels). Nous exposerons rapidement leur principe. Nous développerons ensuite plus spécifiquement les tests correspondant aux questions posées plus haut.

2 Un processus ponctuel ?

2.1 Une définition

On appelle processus ponctuel un processus aléatoire dont une réalisation est un semis de points. Si on note $\mathbf{\Phi}$ un tel processus, ϕ sa réalisation, $\phi(B)$ est le nombre de points qui tombent dans un ensemble B.

Pour que ce processus soit défini, il faut lui imposer quelques conditions:

- $\phi(B)$ est presque sûrement fini. Cette hypothèse suppose qu'il ne peut y avoir un nombre infini de points localement sur les réalisations (presque sûrement ...).

- $\phi(B_1 \cup B_2) = \phi(B_1) + \phi(B_2)$ si $B_1 \cap B_2 = \emptyset$, c'est à dire que le nombre de points qui tombent dans un ensemble formé de deux parties disjointes est égal à la somme des points qui tombent dans chacune des deux parties. Cette condition assure la cohérence de la définition.

Le processus peut alors être formellement défini en utilisant le théorème de Kolmogorov à partir de la définition des lois des n-uplets $\Phi(B_1), \ldots, \Phi(B_n)$ soit des nombres de points tombant dans les ensembles $B_1, .., B_n$ pour tous les $B_1, .., B_n$ possibles.

On impose en général, et en particulier ici, que le processus soit simple, c'est à dire que deux points du processus ne tombent jamais à la même position presque sûrement.

L'exemple le plus simple de processus est le processus de Poisson. Pour une densité $\lambda > 0$, le processus Φ_λ est dit de Poisson si :

- $\Phi_\lambda(B)$ suit une loi de Poisson de paramètre $\lambda\nu(B)$, où $\nu(B)$ est l'aire de B. Autrement dit, la probabilité p_n d'observer un nombre de points n dans B est égale à $P(n) = \frac{(\lambda\nu(B))^n}{n!} \exp\{-\lambda\nu(B)\}$, et le nombre moyen de points dans B est $\lambda\nu(B)$. λ est alors le nombre moyen de points par unité de surface, i.e. l'intensité du processus.

- les $\Phi_\lambda(B)$ points sont répartis dans B indépendamment les uns des autres, et selon une loi uniforme sur B

Le processus Φ_λ est dit stationnaire car la loi de $\Phi_\lambda(B)$ ne dépend pas de la position de B (voir plus bas).

Si f est une fonction positive intégrable sur tout ensemble compact, on définit le processus de Poisson non-stationnaire Φ_f par

- $\Phi_f(B)$ suit une loi de Poisson de paramètre $\int_B f(u)du$,

- les $\Phi_f(B)$ points sont répartis dans B indépendamment les uns des autres, et tout point v inclus dans B est choisi avec la loi de densité $\frac{f(v)}{\int_B f(u)du}$.

Intuitivement, si x est un point du plan et dx une petite surface autour de x, $f(x)dx$ représentera le nombre moyen de points dans la surface dx, et donc $\int_B f(u)du$ sera le nombre moyen de points dans B. On retrouve le même principe que dans le cas stationnaire, mais le nombre moyen de points ne dépend pas uniquement de la surface, mais aussi de la localisation car l'intégrale de f dépend de B et non uniquement de $\nu(B)$. Par contre, dans les deux cas, le nombre de points tombant dans B est poissonnien.

2.2 Isotropie

Le processus est isotrope si sa loi est invariante par rotation, c'est à dire si, pour toute rotation r, $\Phi(r(B)) = \Phi(B)$. Considérons par exemple le cas de semis de plantes hydrophiles, chaque point représentant une plante. Si la zone d'étude est traversée de rigoles plus ou moins parallèles car le terrain est légèrement en pente, les plantes hydrophiles auront tendance à être alignées le long de ces rigoles. Le semis de points sera anisotrope. Cette hypothèse n'est pas fondamentale dans le traitement statistique des données, mais elle simplifie les calculs. Ainsi par exemple, la distribution de la distance d'un point du processus à son plus proche voisin ne dépendra pas de l'orientation, alors que ce n'est pas le cas si le processus est anisotrope. Dans un cas, on pourra se contenter d'une fonction unidimensionnelle fonction de la distance, dans l'autre cas, la fonction sera bidimensionnelle (ou plus si l'espace dans lequel se promènent les points est de dimension supérieure à deux).

2.3 Stationnarité

Intuitivement, on dira qu'un processus est stationnaire si la probabilité d'observer un évènement dans un ensemble A donné ne dépend que de la forme de A et pas de sa localisation. Par exemple, la probabilité d'observer n points dans un carré va dépendre de la taille du carré, mais pas de l'endroit où on le met. Formellement, on dira que le processus est stationnaire si sa loi est invariante par translation, c'est à dire si, pour tout n, pour tout n-uplet d'ensembles $B_1, \ldots, B_n$ et toute translation t, la loi de $\Phi(t(B_1)), \ldots, \Phi(t(B_n))$ est égale à la loi de $\Phi(B_1), \ldots, \Phi(B_n)$.

2.4 Ergodicité et mélange spatial

L'intérêt de la stationnarité est, comme dans le cadre de données indépendantes équidistribuées, de pouvoir estimer une caractéristique (par exemple le nombre moyen de points dans un cercle de rayon 0.5) en prenant des répétitions dans différentes zones de l'espace. En effet, si B_1 et B_2 sont deux zones de même forme, $\Phi(B_1)$ et $\Phi(B_2)$ ont même loi. Il est alors naturel, pour estimer la valeur moyenne du nombre de points tombant dans un ensemble B, de prendre des ensembles $B_1,..,B_n$ de même forme que B et de prendre la moyenne des nombres de points tombant dans chacun d'eux. Cependant, pour que cette moyenne aît un sens, il faut que les nombres de points tombant dans ces ensembles représentent des répétitions au sens statistique du terme, i.e. qu'ils soient suffisamment peu dépendants entre eux pour que la moyenne empirique converge vers l'espérance de la valeur. Un processus sera dit ergodique s'il vérifie une telle propriété.

Considérons les deux exemples suivants:

1 on tire au hasard la valeur d'une variable aléatoire U: $U = a$ avec probabilité $1/2$, $U = A$ sinon. Pour chaque tirage donné de U on considère alors le processus de Poisson d'intensité U. Le processus obtenu est un exemple (caricatural) un processus de Cox.

2 on considère la partition de $\mathbb{R}^2$ en carrés de taille unité, on affecte à chaque carré (i, j) une variable aléatoire $U_{i,j}$: $U_{i,j} = a$ avec probabilité $1/2$, $U_{i,j} = A$ sinon. Les $U_{i,j}$ sont indépendants entre eux. Pour un tirage donné des $U_{i,j}$, on jette au hasard sur chaque carré un nombre $N_{i,j}$ de points selon une loi de Poisson de paramètre $U_{i,j}$. Le processus obtenu est aussi de processus de Cox.

On trouvera en figure 1 un exemple de réalisation de tels processus. Nous avons pris $a = 1$ et $A = 10$. Dans cet exemple où les deux valeurs a et A sont très différentes, les deux réalisations sont très différentes dans le cas non-ergodique (cas 1), alors que l'on ne note pas de différence d'intensité ou de structure entre les deux réalisations du cas ergodique (cas 2).

Si nous disposons d'une réalisation des deux processus, on peut compter le nombre $m_{i,j}$ de points compris dans les cercles de rayon 0.5 et de centre les points de coordonnées $(i + 1/2, j + 1/2)$. Dans le premier cas, le nombre de points $m_{i,j}^{(1)}$ suit une loi de Poisson de paramètre $\frac{\pi}{4}U$ où U est le résultat du

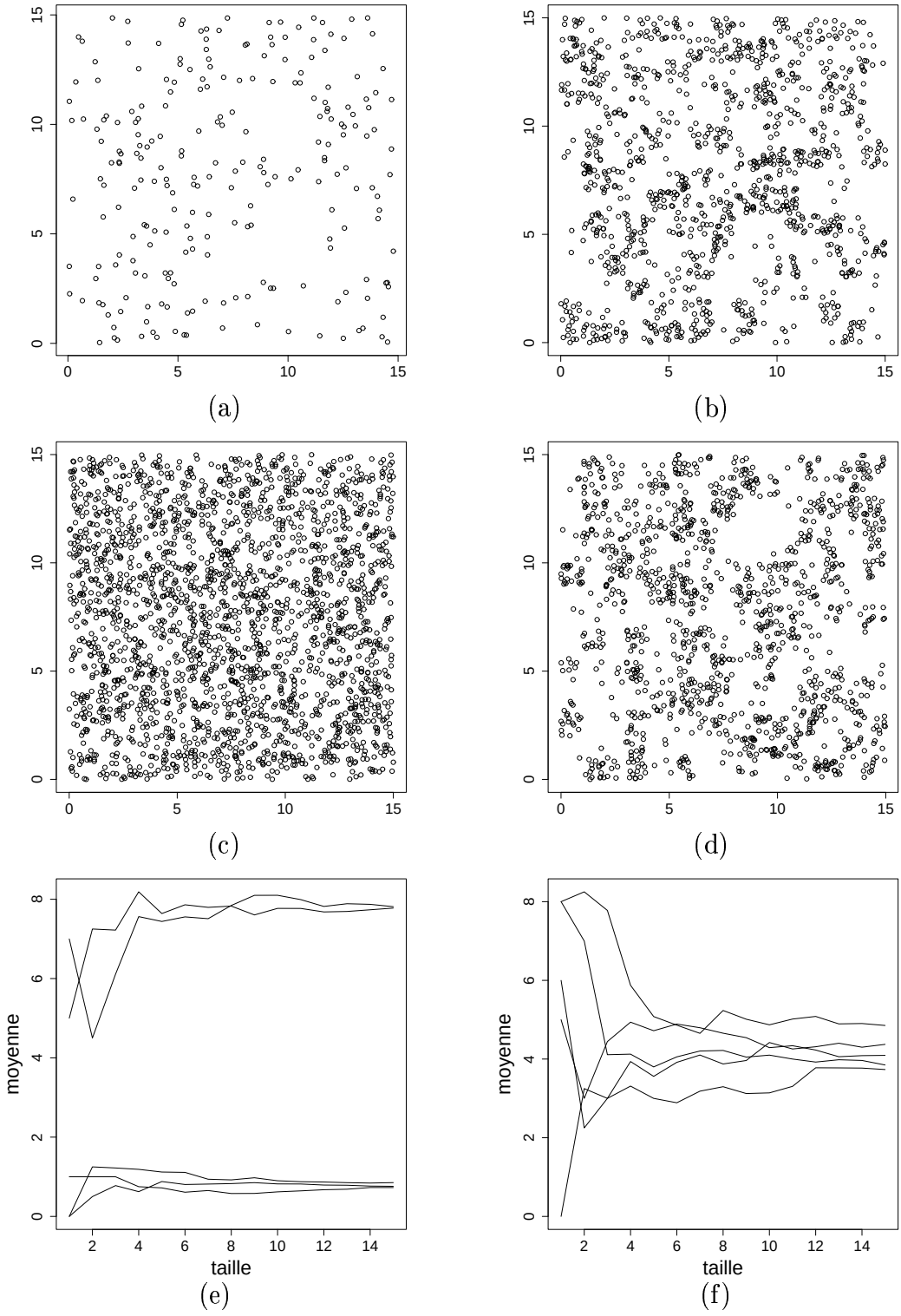


Figure 1: *Processus ergodique et non-ergodique. Première colonne : processus de Cox non-ergodique (intensité aléatoire, constante sur le plan, de valeur 1 ou 10 avec probabilité 0.5), deuxième colonne : processus de Cox ergodique (intensité aléatoire, constante par carré unitaire, de valeur 1 ou 10 avec probabilité 0.5). Figures (a) à (d) : exemples de réalisation, figures (e) et (f) : évolution de la moyenne du nombre de points à moins de 0.5 d'un point de la grille de pas 1 partant de (0.5,0.5) en fonction de la taille de la grille. Chaque ligne correspond au calcul sur une réalisation.*

tirage au hasard. la moyenne des $m_{i,j}^{(1)}$ tendra vers $\frac{\pi}{4}U$. Dans le deuxième cas, le nombre de points $m_{i,j}^{(2)}$ dans le cercle de centre $(i+1/2, j+1/2)$ suit une loi de Poisson de paramètre $\frac{\pi}{4}U_{i,j}$ où U est le résultat du tirage au hasard. la moyenne des $m_{i,j}^{(2)}$ tendra vers $\frac{\pi}{4}\frac{1}{2}(a+A)$. Dans le premier cas, la valeur limite de la moyenne dépend de la valeur de U, donc de la réalisation. Dans le deuxième cas, la valeur limite est indépendante de la réalisation. Si on veut estimer des caractéristiques des processus, on voit donc que la stationnarité ne suffit pas : il faut aussi l'ergodicité.

Ce résultat est illustré en figure 1. On a calculé cinq réalisations de chaque processus. Pour chaque réalisation, on a calculé la moyenne des $m_{i,j}$ sur des grilles de pas 1, de taille 1 à 15. L'évolution de cette moyenne pour chaque réalisation est représentée par une ligne des graphiques de la figure 1. Dans le cas ergodique, les lignes convergent, plus ou moins rapidement, vers $\pi(a+A)/4 \simeq 4.32$. Dans le cas non-ergodique, chaque ligne converge vers sa propre moyenne, (soit $\pi a/4 \simeq 0.78$ soit $\pi A/4 \simeq 7.85$)

Pour assurer que le "mélange spatial" est correct, deux voies sont possibles.

1)La première a cherché à quantifier directement ce mélange. Ainsi, on parlera par exemple de mélange exponentiel quand $|| P(U \cap V) - P(U)P(V) || \leq f(\nu(A), \nu(B)) \exp(-\alpha d)$ où U et V sont deux évènements sur deux parties A et B bornées de $\mathbb{R}^2$ distantes de d, $\alpha > 0$, $f(\nu(A), \nu(B))$ une fonction polynomiale en les surfaces de A et B. Si les évènements U et V étaient indépendants,$|| P(U \cap V) - P(U)P(V) ||$ serait nul. Dans le cas du mélange exponentiel, on impose que l'écart à l'indépendance entre les évènements U et V, mesuré par $|| P(U \cap V) - P(U)P(V) ||$ tende vers 0 à vitesse exponentielle quand les surfaces A et B sur lesquelles on observe le processus s'éloignent. De ce fait, si les évènements U et V ne sont pas strictement indépendants, ils en deviennent rapidement très proches.

2)Dans beaucoup de cas, vérifier directement des hypothèses de mélange, qui sont de plus suffisantes, mais pas nécessaires pour obtenir la convergence des statistiques, n'est pas possible. On fera alors une hypothèse d'ergodicité, qui s'exprime à partir d'une notion d'ensemble invariant. Pour toute réalisation Y, on note Y_x sa translation par le vecteur x. on note $Y \backslash Y_x$ les points de Y non inclus dans Y_x et $Y_x \backslash Y$ les points de Y_x non inclus dans Y. Alors $Y \backslash Y_x \cup Y_x \backslash Y$ désigne les points contenus dans Y ou Y_x, mais pas dans les deux. Y est dit un ensemble invariant si $P(Y \backslash Y_x \cup Y_x \backslash Y) = 0$ pour tout

x. Ceci traduit le fait "d'absence de mémoire" de ces ensembles. Le processus est dit ergodique si, pour tout ensemble invariant, $P(Y) = 0$ ou 1. Si le processus est ergodique, le nombre de points N_n observé dans une suite d'ensembles S_n tels que $S_1 \subset \ldots S_n..$ tendant vers $\mathbb{R}^2$ vérifie $\lambda = \lim_n \frac{N_n}{A_n}$ où A_n est l'aire de S_n et λ le nombre moyen de points du processus par unité de surface.

3 Les tests de permutation

Les tests de permutation sont des procédures statistiques développées pour éviter d'avoir à spécifier un modèle précis. Elles sont très proches du bootstrap (voir Efron & Tibshirani 1993) dans leur principe, mais on restreint les tirages de façon à conserver la distribution de la variable, ce qui nous donnera des tests exacts. Plus précisément, soit $X = (x_1, .., x_n)$ un échantillon, le principe du test est de :

- calculer une fonction d'intérêt sur l'échantillon, $T(X)$
- permuter au hasard les valeurs de l'échantillon, de telle sorte que l'échantillon X et l'échantillon permuté X_p aient la même probabilité sous l'hypothèse à tester,
- calculer la statistique $T(X_p)$ sur l'échantillon permuté,
- répéter N fois l'étape précédente pour obtenir un ensemble $T_1, ..., T_n$ de la valeur de la statistique sous l'hypothèse,
- calculer un intervalle de confiance de T à partir de cet échantillon en prenant les quantiles de niveau voulu, par exemple les quantiles de niveau $\alpha/2$ et $1 - \alpha/2$ si on veut un intervalle de confiance symétrique de niveau α
- regarder si $T(X)$ est dans l'intervalle de confiance.

Pour une taille n donnée, le nombre de permutations possibles est $N_{\max} = n!$. Si n est petit, on pourra faire l'ensemble des permutations possibles plutôt que de les tirer au hasard. Dès que n est un peu grand, le nombre de permutations possibles devient rapidement très grand . C'est dans ce cas surtout que le tirage au hasard se fera. Pour un test unilatéral d'un niveau α

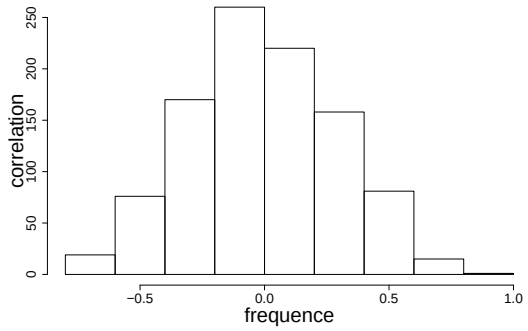


Figure 2: *Histogramme des corrélations sous permutation aléatoire des valeurs de chaque ligne du tableau 1.*

donné, il suffit de faire $N_0 = 1 - \frac{1}{\alpha}$ permutations pour obtenir un test exact de ce niveau. Cependant, On aura intérêt à monter N le plus possible (soit 5 à 10 fois N_0) pour augmenter la puissance du test (Hope 1968).

individu	1	2	3	4	5	6	7	8	9	10	11	12
X	6.62	3.22	0.32	1.55	8.60	6.55	2.12	0.05	7.86	5.47	7.37	9.56
Y	9.46	6.24	1.94	3.93	6.84	6.20	6.93	3.01	5.37	8.79	7.73	7.84

Table 1: *Exemple de deux séries de données de corrélation 0.702.*

Considérons par exemple le cas de deux variables Xet Y mesurées sur des individus indépendants. 12 points ont été échantillonnés. On a obtenu les valeurs données dans la Table 1 et un coefficient de corrélation entre les deux variables de 0.70. Sous l'hypothèse d'indépendance entre X et Y, les couples (X, Y) et (X, Y_p) (où Y_p représente le résultat d'une permutation aléatoire des données) sont équiprobables. Ainsi, pour deux permutations au hasard on obtient:

individu	1	2	3	4	5	6	7	8	9	10	11	12
Y_{p1}	7.84	6.93	7.73	9.46	1.94	3.93	5.37	8.79	6.20	6.84	6.24	3.01
Y_{p2}	6.84	7.73	5.37	8.79	3.01	6.93	6.24	9.46	6.20	3.93	1.94	7.84

ce qui donne les corrélations de -0.72 et -0.44

Sur 1000 permutations, on obtient l'histogramme des corrélations donné en figure 2, et un intervalle de confiance à 5% de $[-0.551, 0.592]$ qui ne contient pas la valeur observée. On rejette donc l'hypothèse d'indépendance des séries.

En conclusion, on remarquera que l'exemple choisi est un exemple que l'on traite typiquement en régression linéaire. En régression linéaire, on estimera une variance résiduelle mais on supposera connue la distribution des erreurs. En utilisant les tests de permutation, on ne se permet plus cette hypothèse. Le prix à payer sera alors une moins grande puissance du test de permutation.

4 Statistiques d'intérêt

Deux grands types de statistiques ont été utilisées pour caractériser la répartition spatiale d'un semis de points. Le premier type, que nous n'aborderons pas ici, est basé sur le calcul d'indices. Il est présenté dans (Upton & Fingleton 1988). Le deuxième type est plus approprié aux données cartographiées. Présenté dans (Diggle, 1983), il repose sur une base empirique. Si on note $X_1, \ldots, X_n$ les N positions du semis observé dans une zone W, ces statistiques sont plus précisément:

- **la distribution de distance d'un point quelconque à un point du processus.** Cette fonction va mesurer la distribution des vides dans un semis. On va la calculer en prenant les M points Y_i dans W au nœuds d'une grille régulière, calculer pour chaque Y_i les distances aux points X_i du semis soit $|| Y_i - X_1 ||, ..., || Y_i - X_N ||$, prendre la plus petite $\min_{j \leq N} || Y_i - X_j ||$, puis en faire la distribution empirique pour l'ensemble des points Y_i, soit :

$$\widehat{F}(r) = \frac{1}{M} \sum_{i \leq M} \boldsymbol{I}_{\{\min_{j \leq N} ||Y_i - X_j|| < r\}}$$

  en notant $\boldsymbol{I}_v$ la fonction qui vaut 1 si l'assertion v est vérifiée, 0 sinon. Elle permettra de mesurer si des vides sont présents de manière excessive (ou trop présents) à une échelle donnée par rapport à ce qu'on s'attendrait sous le hasard.

- **La distribution de distance au plus proche voisin.** Cette dernière se calcule simplement en calculant la distance de chaque point du semis

à son plus proche voisin, puis en calculant la distribution empirique :

$$\widehat{G}(r) = \frac{1}{N} \sum_{i \leq N} \boldsymbol{I}_{\{\min_{j \neq i} ||X_i - X_j|| < r\}}$$

Elle se révèle particulièrement efficace pour analyser les écarts à l'indépendance sur les petites distances, dans le cas par exemple de processus réguliers où l'on interdit la présence de points à courtes distances l'un de l'autre (processus hard-core).

- **la distribution de distance entre points du processus.** On calcule simplement la distribution des distances entre tous les points du semis :

$$\widehat{H}(r) = \frac{1}{N(N-1)} \sum_{i \neq j} \boldsymbol{I}_{\{||X_i - X_j|| < r\}}$$

Elle se révélera plus particulièrement utile dans le cas où le semis ne s'écarte pas trop de l'indépendance sur les courtes distances, mais où l'écart va se faire sur des distances plus grandes.

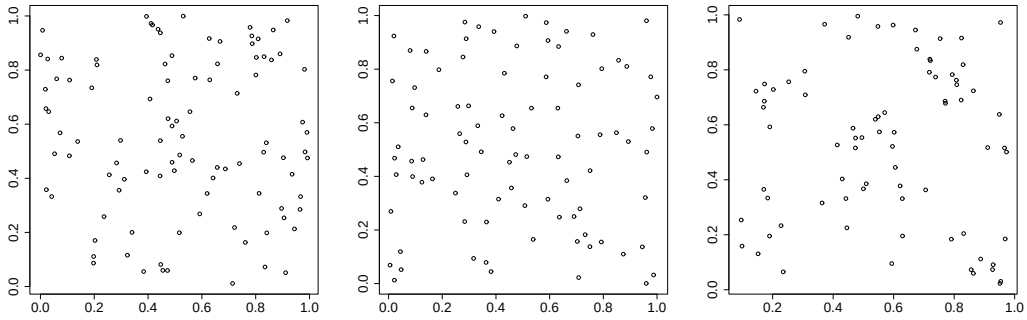
Ces statistiques résument rapidement les caractéristiques spatiales du semis de points aux trois grands types de distance rapidement calculables. Elles ne cherchent pas à estimer une caractéristique intrinsèque du processus. De ce fait, on ne fait pas de correction de bord, celle-ci est intégrée dans la statistique. Si on considère que ces statistiques ont été proposées pour explorer les semis de points en terme d'écart à l'indépendance totale, ce ne sera pas tant la valeur de ces statistiques que leur position par rapport à une bande de confiance construite sous l'indépendance qui prendra du sens.

5 Tests "Complete Spatial Randomness"

Si les points du semis ont été répartis dans la zone étudiée indépendamment les uns des autres et totalement au hasard, la réalisation $\phi = (x_1, \ldots, x_n)$ suit, conditionnellement au nombre de points n, la même loi que $(y_1, \ldots, y_n)$ où les y_i sont indépendants, identiquement distribués et de loi uniforme.

On va donc tester cette hypothèse en comparant les valeurs empiriques $\widehat{F}(r)$, $\widehat{G}(r)$ et $\widehat{H}(r)$ à leurs bandes de confiance sous l"hypothèse CSR, i.e. avec la densité

$$p(x_1, \ldots, x_n) = \frac{1}{\nu(W)}$$



(a) (b) (c)

Figure 3: *Trois semis de points obtenus : (a) sous hypothèse d'indépendance totale entre position des points, (b) sous hypothèse de régularité (la distance entre deux points est toujours supérieure à 0.033), (c) sous hypothèse d'agrégation (union de paquets de points aléatoires de 5 points en moyenne)*

si on note $\nu(W)$ la surface de W. Cette densité est la plus simple possible, mais le calcul analytique de la loi de distribution de F, G et H devient déjà difficile, surtout si la forme de W est un peu complexe.

Considérons l'exemple classique des trois semis de points suivants (voir Stoyan, Kendall & Mecke J. 1995):

- La figure 3(a) est obtenue en répartissant 100 points au hasard selon une loi uniforme dans le carré unité. Pour tirer un point au hasard, il suffit de tirer au hasard deux nombres entre 0 et 1. Le premier représentera son abscisse, le deuxième son ordonnée. On répète autant de fois que nécessaire pour obtenir le nombre de points désirés.

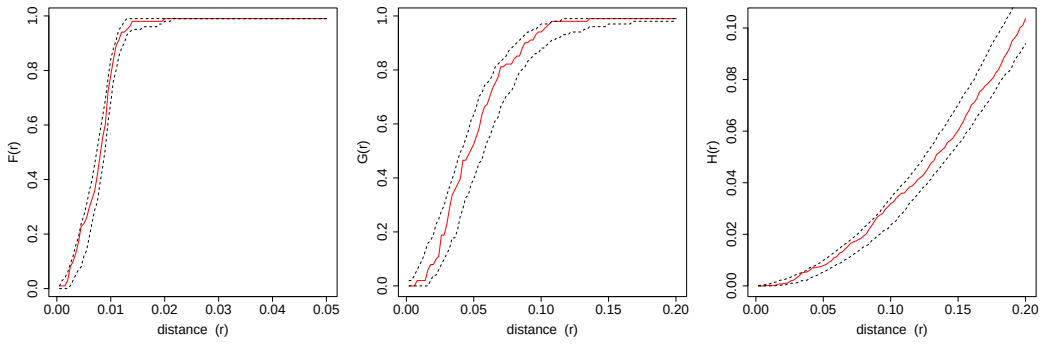
- La figure 3(b) représente la réalisation d'un processus hard-core, obtenu en semant au hasard 100 points dans le carré, en supprimant itérativement le point dont le nombre de voisins à moins de 0.033 est maximal, jusqu'à ce que tous les points restants soient à plus de 0.033.

- la troisième figure (3c) représente une réalisation d'un processus de Neyman-Scott. On sème au hasard 30 points "parents" dans le carré $[-0.1, 1.1]^2$, on sème autour de chaque point parent des points fils dans le carré de coté 0.2 centré sur le parent. Le nombre de points fils suit une loi de Poisson de paramètre 5. Le semis considéré est formé

des fils tombant dans le carré unité. Ce processus est donc formé de l'union de paquets de fils de taille moyenne 5, disposés au hasard et indépendamment les uns des autres. En tirant les parents dans le carré $[-0.1, 1.1]^2$, on évite les effets de bord sur la répartition des fils dans le carré $[0, 1]^2$.

Une distribution complètement aléatoire n'est pas incompatible avec la présence de paquets de points ou de vides, ainsi qu'on peut le peut noter sur la figure 3(a). De ce fait, il est difficile à l'oeil de distinguer entre le processus 3(a) et celui en 3(b) a priori plus régulier. le but des tests CSR est de faire rapidement le tri entre régulier/hasard/agrégatif.

Pour cela, on va prendre les trois statistiques précédentes, les calculer sur les figures observées, puis calculer leurs intervalles de confiance par simulation comme expliqué en section 3. L'hypothèse H0 est celle d'indépendance entre points, donc chaque simulation se fera en distribuant au hasard le même nombre de points dans la surface.

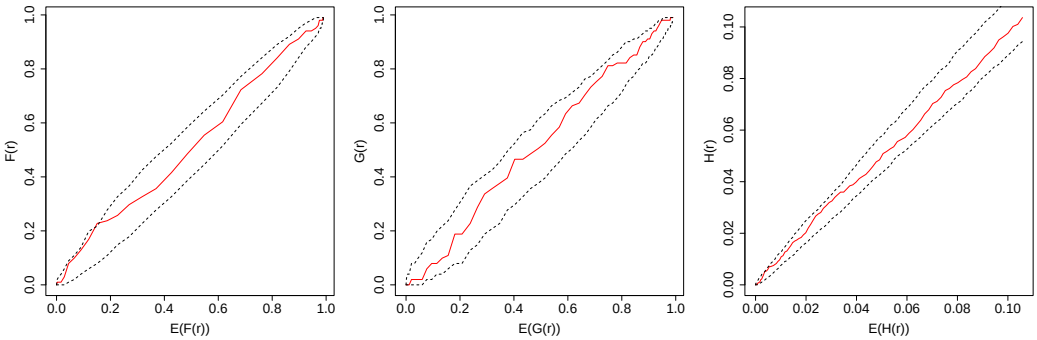


(a) (b) (c)

Figure 4: *Test d'indépendance spatiale du semis de la figure 3(a). (a) distribution de distance d'un point du carré au plus proche point du semis, (b) distribution de distance d'un point du semis à son plus proche voisin, (c) distribution de distance entre tous les points du semis. En abscisse : distance.*

Considérons le cas du premier semis de points (figure 3a). Les bandes de confiance de F (figure 4a), de G (figure 4b) et de H (figure 4c), tracées en noir tireté, correspondent pour chaque distance à un intervalle de confiance

individuel à 95%. [1] La courbe estimée sur le semis de la figure 3(a), en ligne continue, passe dans la bande de confiance. On remarquera sur ces courbes la montée relativement brutale avec la distance, ce qui tend à écraser les bandes de confiance et la courbe observée, rendant difficile l'observation des écarts.



(a) (b) (c)

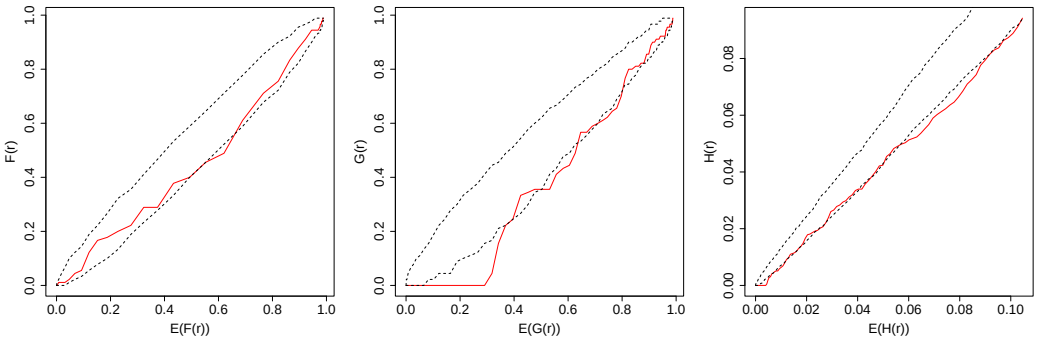
Figure 5: *Test d'indépendance spatiale du semis de la figure 3(a). Mêmes statistiques que la figure 4. En abscisse : valeur moyenne de la statistique correspondante sous l'indépendance.*

Pour éviter ce problème, et comme les courbes sont monotones, on utilise plutôt la moyenne de la statistique, ce qui donne, sur les mêmes simulations et les mêmes semis qu'en figure 3, les courbes suivantes en figure 5.

Ce changement d'abscisses permet une meilleure vision des courbes, cependant que les figures du type de celles de la figure 4 seront plutôt utilisées pour situer les échelles où ce que l'on observe se passe. Par la suite, on restera sur ce dernier type de graphique.

L'analyse du deuxième semis de points est présentée dans la figure 6. Si les trois statistiques sont affectées par l'écart au processus de Poisson, c'est la distribution au plus proche voisin qui est la plus sensible à un écart à l'indépendance totale dans le sens d'une régularité.

[1] A chaque distance, on construit un intervalle de confiance dont l'union nous fournit les bandes précédentes. Ces bandes ne constituent pas un test global de niveau 95%, mais plus faible. Pour construire un test global sur les distances $[d, D]$, on pourra par exemple reprendre un test de permutation basé sur l'écart entre les courbes (simulées et observée) et la courbe moyenne (voir Diggle, 1983).



(a) (b) (c)

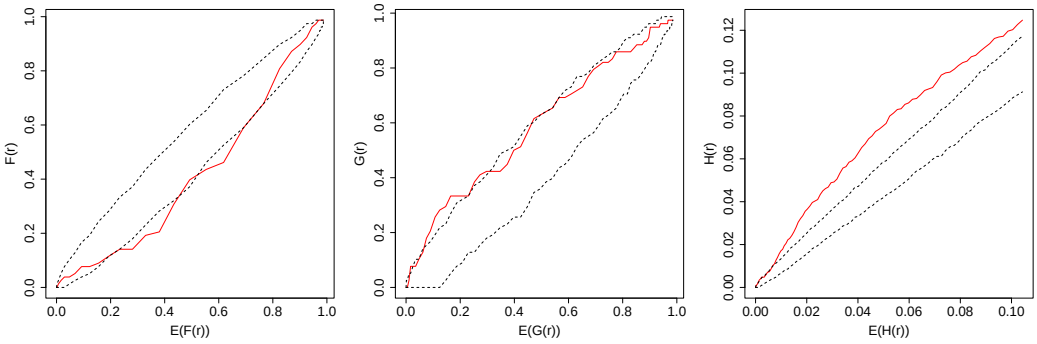
Figure 6: *Test d'indépendance spatiale du semis de la figure 3(b). Même axes que pour la figure 5*

Dans le cas du semis de type agrégatif, on voit sur la figure 7 que c'est surtout la distribution de distance entre points du processus qui est sensible à l'écart à l'indépendance, même si toutes les statistiques sortent de la bande de confiance ou en frôlent la limite.

Ces trois procédures restent les plus employées pour juger du caractère agrégatif, complètement aléatoire ou régulier d'un semis de points. Elles allient exploration d'aspects complémentaires du semis et intégration de l'effet de bord ce qui rend leur usage très général. On trouvera d'autres statistiques construites soit à partir des fonctions de base F, G et H (fonction J proposée par Van Lieshout et Baddeley (1997) par exemple), soit sur d'autres caractéristiques du processus (fonction K de Ripley (1976) en particulier). Dans ce dernier cas, la statistique offre l'avantage d'être directement interprétable, en dehors de tout aspect de test. En contrepartie, elle nécessite une correction de bord, ce qui alourdit les calculs et n'est pas toujours facile à mettre en œuvre dans le cas de fenêtres d'observations complexes.

6 Tests d'indépendance pour un processus marqué

On dit qu'un processus ponctuel est marqué quand on attache à chaque point du processus une variable supplémentaire appelée la marque. Les marques peuvent prendre des valeurs continues, ou des valeurs discrètes. Les cas pour



(a) (b) (c)

Figure 7: *Test d'indépendance spatiale du semis de la figure 3(c). Même axes que pour la figure 5*

lesquels ces processus sont utilisés peuvent être par exemple :

- le diamètre des arbres d'une parcelle forestière. Chaque point représentera la position d'un arbre, la marque son diamètre. Celle-ci est alors continue.
- La présence d'une maladie dans une parcelle forestière. Chaque point représentera la position d'un arbre, la marque la présence ou l'absence de maladie sur cet arbre.
- La répartition de deux espèces d'arbres dans une parcelle forestière. Chaque point représentera la position d'un arbre, la marque l'espèce à laquelle il appartient.

On voit sur les deux derniers exemples où les marques sont discrètes que le même cadre probabiliste englobe deux types de modèles très différents.

- Dans le premier cas, les positions des points sont fixées avant l'arrivée de la maladie. Le problème posé par l'indépendance va alors être de savoir si, une fois les positions connues, les marques se répartissent au hasard sur ces points ou si celles-ci forment des paquets (ou sont plutôt régulièrement réparties) parmi l'ensemble des points.
- Dans le deuxième exemple, les deux espèces peuvent s'être développées en même temps. Chaque espèce peut former des paquets ou non (cas

par exemple des espèces qui génèrent des rejets) et la question est alors de savoir si les deux processus sont indépendants entre eux, même si, individuellement, ils peuvent être non poissoniens.

A chacune de ces questions va donc correspondre un test particulier, i.e. des modes de tirages aléatoires différents, alors que la statistique utilisée sera toujours la même : une mesure de distance entre points de marques différentes.

6.1 Statistiques d'intérêt

Notons $(x_1, \ldots, x_n)$ les positions des points du semis, $(m_1, \ldots, m_n)$ les marques des n points, et supposons que $m_i = 1$ ou 2[2] . Notons n_1 le nombre de points de marque égale à 1; n_2 le nombre de ceux de marque égale à 2. Les fonctions utilisées doivent mesurer la distance entre les deux processus. On va donc retrouver des versions modifiées des fonctions précédentes. On va rencontrer par exemple:

- **la distribution de distance interpoints.** On calcule les distances entre tous les points de marque 1 et tous les points de marque 2 et puis on calcule la distribution de ces distances soit:

$$H_{12}(r) = \frac{1}{n_1 n_2} \sum_{i \leq n_1} \sum_{j \leq n_2} \mathbb{1}_{\{m_i=1\}} \mathbb{1}_{\{m_j=2\}} \mathbb{1}_{\{||x_i - x_j|| < r\}}$$

- **la distribution de distance d'un point d'un processus à son plus proche voisin de l'autre processus.** Pour chaque point de marque 0, on calcule la distance à son plus proche voisin de marque 1 et on calcule la distribution de ces distances soit:

$$G_{12}(r) = \frac{1}{n_1} \sum_{i \leq n_1} \mathbf{I}_{\{m_i=1\}} \mathbf{I}_{\left\{\min_{\{j/m_j=2\}}(||x_i - x_j||) < r\right\}}$$

Si la fonction $H_{12}(r)$ est symétrique en les deux processus, la fonction $G_{12}(r)$ ne l'est pas, on regardera aussi par conséquent la fonction $G_{21}(r)$.

[2]Le cas où la marque est à valeur continue, par exemple le diamètre des arbres, ne ressort pas strictement de ce paragraphe, dans le sens où elle ne définit pas des sous-ensembles de semis de points. Le cas 6.2 où l'analyse reposera sur l'identification des sous-ensembles ne s'appliquera donc pas, mais le 6.3 continuera de s'appliquer.

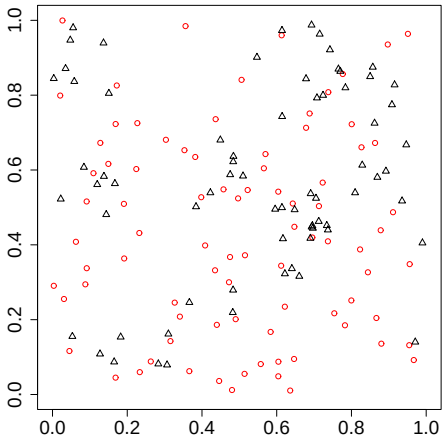


Figure 8: *semis formé de l'union (a)d'un semis de points régulier similaire à celui de la figure 3(b) (cercles) (b)d'un semis de points agrégé similaire à celui de la figure 3(c) (triangles)*

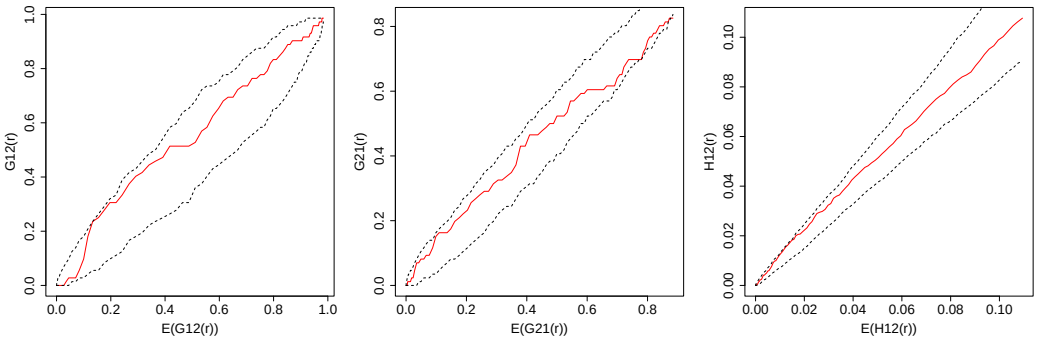
6.2 Test d'indépendance entre deux processus

6.2.1 Principe

On suppose que deux processus, chacun avec a priori sa propre structure spatiale, se superposent indépendamment l'un de l'autre. Notons $X^{(1)}$ et $X^{(2)}$ les deux processus. Si on suppose les deux processus stationnaires et isotropes, alors, on avait la même probabilité d'observer dans la fenêtre W les semis observés $W \cap X^{(1)}$ et $W \cap X^{(2)}$ que les semis $W \cap T(X^{(1)})$ et $W \cap T'(X^{(2)})$ où T et T' sont des isométries (i.e. des compositions de rotations et translations) dans $\mathbb{R}^2$.

Pour des fenêtres $W = [0, L] \times [0, l]$ rectangulaires, le principe du test est alors de refermer W sur lui même en utilisant la convention du tore (décrite plus bas), et d'utiliser le groupe des rotations-translations de ce tore. En pratique, si on se restreint au groupe des translations:

- on tire aléatoirement les deux coordonnées du vecteur de la translation aléatoire entre 0 et L pour la première, 0 et l pour la seconde. On obtient ainsi un vecteur v.
- on translate le deuxième processus pour obtenir $X_v^{(2)} = v + X^{(2)}$



(a) (b) (c)

Figure 9: *Test d'indépendance entre cercles et triangles du semis de la figure 8. Hypothèse de superposition aléatoire indépendante de deux processus. (a) distribution de distance d'un cercle au plus proche triangle. (b) distribution de distance d'un triangle au plus proche cercle. (c) distribution des distances entre cercles et triangles. Abscisses: valeurs moyennes des statistiques sous l'indépendance.*

- on utilise la convention du tore pour remettre dans W les points de $X_v^{(2)}$ qui ne tombent plus dedans. Si on note $A = (a, b)$ un point de $X_v^{(2)}$, la réalisation aléatoire sur le tore sera finalement formée des points

  $$A' = (a\boldsymbol{I}_{\{a<L\}} + (a-1)\boldsymbol{I}_{\{a>=L\}}, b\boldsymbol{I}_{\{b<l\}} + (b-1)\boldsymbol{I}_{\{b>=l\}}))$$

L'objectif de cette procédure, conserver la structure marginale de chaque processus, est donc atteint via l'usage des isométries, à l'effet de bord près dû au "recollement" sur le tore.

6.2.2 Exemple

Considérons l'exemple de la figure 8. Il est formé de l'union d'un semis de points issus d'un processus hard-core tel que décrit en section 5 et dont une réalisation est présentée en figure 3(b), (marque 1, représentée par un cercle) et d'un semis de points issu d'un processus de neyman-scott décrit en section 5 et dont une réalisation est présentée en figure 3(c), (marque 2, représentée par un triangle). Ces deux semis sont issus de processus indépendants.

Le résultat des tests proposés ci-dessus est tracé en figure 9. Comme précédemment, nous avons pris en abscisse de chaque graphique la valeur

moyenne de la statistique sous l'indépendance pour des raisons de lisibilité.

Pour les trois statistiques, la courbe observée (en traits pleins) reste à l'intérieur de la bande de confiance et on ne rejette pas l'hypothèse d'indépendance.

6.3 Test d'indépendance du marquage d'un processus ("random labelling")

6.3.1 Principe

On suppose ici que la marque est arrivée postérieurement au processus. Sous cette hypothèse, la répartition observée des marques est de même probabilité que toute autre répartition de marque sur les mêmes points ayant les mêmes nombres de marques à 1 et à 2. Le tirage aléatoire va donc se faire en jetant au hasard les marques observées parmi les points du semis. On n'aura donc pas ici d'effet de bord comme précédemment.

6.3.2 Exemple

Partant d'un semis de points issu du processus de neyman-scott comportant deux fois plus de points parents que celui présenté en section 5, on attache à chaque point une marque 1 ou 2 tirée au hasard indépendamment d'un point à l'autre, avec probabilité 1/2.

On obtient alors la figure 10 où les points marqués 1 sont représentés par des cercles, les points marqués 2 par des triangles.

L'analyse de la répartition aléatoire des marques au travers des trois statistiques proposées est présentée en figure 11. Là encore, on notera que les trois courbes calculées sur le semis réel sont incluses dans leurs bandes de confiance respectives, ne permettant pas de rejeter l'hypothèse d'un marquage aléatoire uniforme des marques (ce qui est bien le cas qui a été utilisé pour générer l'exemple).

6.4 Importance de la spécification de H0

Une mauvaise utilisation de l'hypothèse nulle peut entraîner des erreurs d'interprétation. Ainsi, sous l'expression "tester l'indépendance des marques d'un processus ponctuel marqué" se cache donc deux hypothèses. A chacune d'elles est associé un groupe de permutations spécifiques, laissant la loi des statistiques invariantes sous l'hypothèse testée.

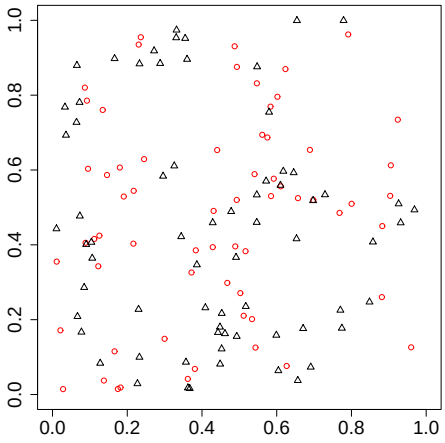


Figure 10: *semis formé par affectation aléatoire d'un cercle ou d'un triangle avec probabilité 1/2 à chaque point d'un processus agrégé, du même type qu'en figure 3(c) et d'intensité double.*

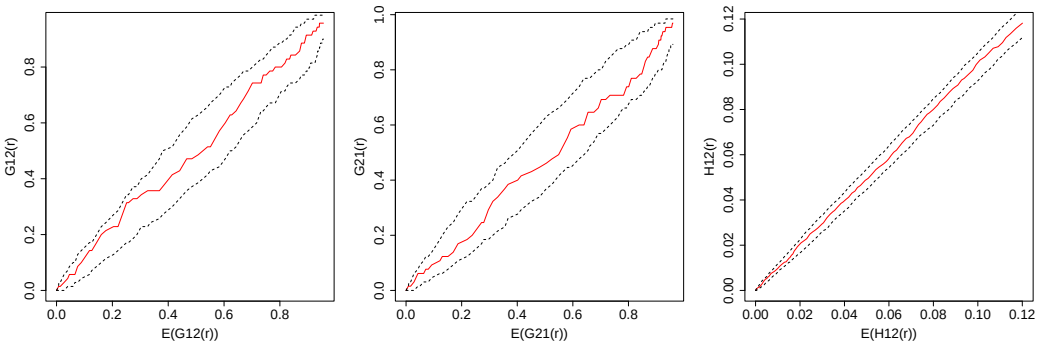
La figure 12 suivante présente le résultat du test appliqué à l'exemple de la figure 8, i.e. quand le marquage sert à désigner deux processus indépendants, si on lui applique la procédure adaptée au test d'une répartition aléatoire des marques parmi les points (soit le paragraphe 6.3). On rejette alors l'hypothèse.

Le même phénomène peut être observé si on analyse le semis de la figure 10 (issu d'un marquage aléatoire indépendant des points d'un processus agrégé) par les procédures testant l'indépendance de deux processus. On voit en figure 13 que l'hypothèse est rejetée.

Ceci se comprend si on revient aux configurations des semis observés.

- En figure 8, les points marqués à 1 (les triangles) sont regroupés en paquets, dus à la présence de paquets dans le processus de Neyman-scott. Lors du premier test(figure 9), on mesure si ces paquets sont ou non positionnés préférentiellement par rapport aux points marqués à 0 (les cercles). Dans le deuxième test (figure 12), on regarde si les triangles forment des paquets parmi les ronds et les triangles, ce que l'on détecte aisément.

- En figure 10, les marques sont placées au hasard parmi les points



(a) (b) (c)

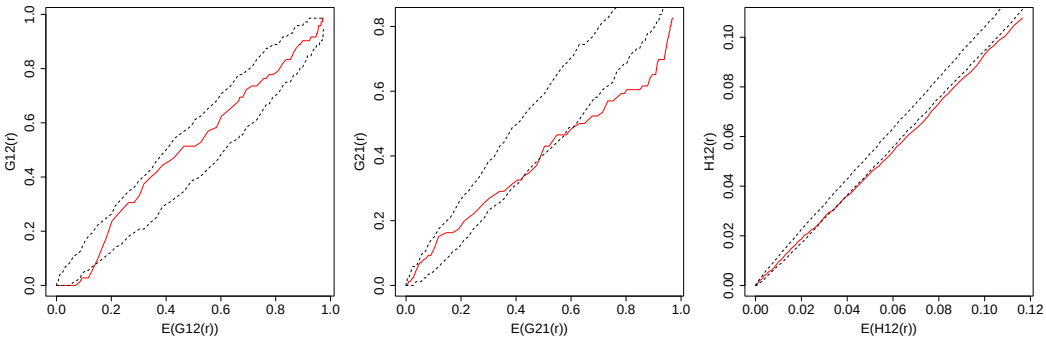
Figure 11: *Test d'indépendance entre cercles et triangles du semis de la figure 10. Hypothèse d'affectation aléatoire des marques. Même type de graphique qu'en figure 9.*

d'un processus de Neyman-Scott. On ne détecte donc pas de groupe spécifique de triangles parmi les triangles et les cercles (figure 11). En revanche, le premier test, en mesurant si les points marqués par les triangles sont placés au hasard par rapport aux points marqués par les cercles, repère la dépendance entre points du processus de neyman-scott.

7 Exemple : la dissémination du sapin concolor

L'objectif de l'étude de (Mahfoud, 2003), dont est issu le jeu de données présenté plus bas, est d'étudier la forte dynamique actuelle de reconquête des espèces d'arbres forestiers. Un des processus centraux de cette dynamique est la structuration des semis dans l'espace suite à la dispersion des graines. Le modèle choisi ici par l'Unité de Recherche Forestières Méditerranénnes (INRA) est un sapin originaire de l'ouest américain (Abies concolor), qui a parfois été utilisé en reboisement dans des zones où le sapin pectiné, autochtone en Europe, a du mal à s'installer.

Dans les modèles de dispersion, on suppose généralement que les semenciers émettent un nombre poissonnien de semis qui se distribuent autour



(a) (b) (c)

Figure 12: *Test d'indépendance entre cercles et triangles du semis de la figure 8. Hypothèse d'affectation aléatoire des marques. même type de graphique qu'en figure 9*

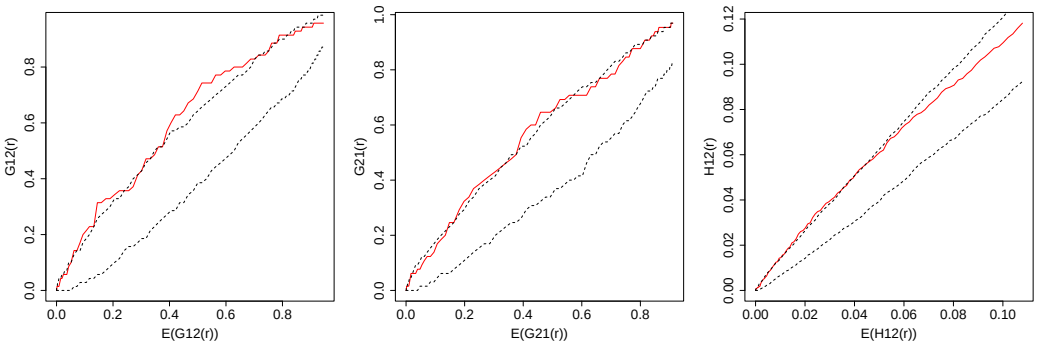
de leur semencier d'origine indépendamment les uns des autres, selon une fonction de dispersion le plus souvent isotrope. Le semis de jeunes autour d'un semencier forme alors un processus de Poisson non-stationnaire dont l'intensité est centrée sur le semencier.

On suppose également qu'il n'y a pas de compétition parmi les semis (par exemple en se mettant suffisamment loin des semenciers pour que la densité de semis soit telle que la distance entre semis ne permette pas de compétition), de telle sorte que le semis observé dans une zone est la superposition des semis venant de chacun des semenciers. Dans ce cadre, les semis venant de semenciers différents sont indépendants, et le semis final suit également un processus de Poisson non-stationnaire.

Une parcelle de la région de Lure a été choisie pour sa disposition privilégiée:

- elle est bordée d'une rangée de sapins concolors producteurs de même âge (plantation de bord de route en forme de haie)
- on ne trouve pas ces sapins dans l'environnement de la parcelle, de sorte que l'ensemble des semis dans la parcelle peut être attribué à cette haie.

Afin de couvrir la variabilité de semis présente dans la parcelle tout en évitant une cartographie complète, les positions des semis dans 7 transects



(a) (b) (c)

Figure 13: *Test d'indépendance entre cercles et triangles du semis de la figure 10. Hypothèse de superposition aléatoire indépendante de deux processus. Même type de graphique qu'en figure 9.*

équidistants de 150m de long et 6m de large perpendiculaires à la haie ont été repérées.

On peut noter a priori sur la figure 14, où les jeunes semis ayant plus de 15 verticilles sont représentés par des points noirs et les limites des transects dessinées en rouge, une forte décroissance de la densité de semis avec la distance à la haie, ainsi qu'une assez forte différence de densité entre transects.

Le premier test a été construit pour vérifier rapidement que la distribution spatiale des points n'est pas totalement au hasard dans les transects. On prend comme statistique de test la distribution de distance au plus proche voisin. Le test est un test de randomisation totale. Pour chaque simulation, on redistribue donc au hasard les points parmi les transects et uniformément dans les transects. La figure 15(a) présente une réalisation typique sous cette hypothèse. On voit sur la figure 15(b) que l'hypothèse est fortement rejetée.

Les semenciers étant répartis sur une haie le long de l'axe $y = 0$, les modèles de dispersion classiques conduisent en fait à un semis issu d'un processus de Poisson non-stationnaire le long de l'axe y si les contributions des semenciers sont équivalentes, ce qui parait plausible dans la mesure où tous les semenciers sont de même âge. Rejeter l'hypothèse d'indépendance totale est donc logique.

Dans un deuxième temps, on teste cette hypothèse d'indépendance des

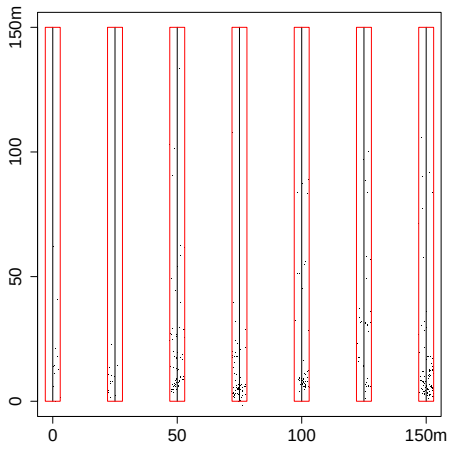
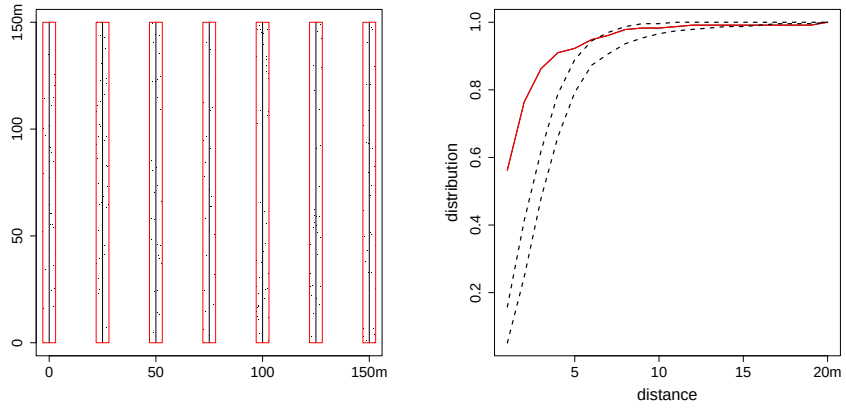


Figure 14: *Position de semis de sapins observés sur transects perpendiculaires à la haie semencière.*

semis conditionnellement à leur distance à la haie. Nous prenons la même statistique de test, la distance au plus proche voisin. Pour conserver la distribution de distance des semis à la haie, on ne va considérer que les randomisations conservant cette distribution. En pratique, pour chaque simulation, on conserve la coordonnée y_s de chaque point du semis et on tire son abscisse au hasard parmi les segments formés de l'intersection de la droite $y_s = y$ et des transects. La figure 16(a) illustre une réalisation possible d'un tel tirage. Sur la figure 16(b) est représentée la courbe observée ainsi que sa bande de confiance sous l'hypothèse d'indépendance conditionnellement à la distance à la haie. On note que l'écart entre la courbe observée et la bande de confiance est beaucoup plus faible qu'en 15(b), mais l'hypothèse reste rejetée.
L'hypothèse de simple non-stationnarité ne dépendant que de la distance à la haie est plausible si la haie est suffisamment homogène, de telle sorte que la densité de graines émise par unité de surface de la haie est constante. De fait, des arbres sont morts, leurs vigueurs sont variables, et celle-ci ne peut pas être vraiment considérée comme homogène, ce qui explique le rejet de l'hypothèse.

Dans un troisième temps on cherche alors à savoir si, conditionnellement à leur présence dans un transect donné et à leur distance à la haie, les semis

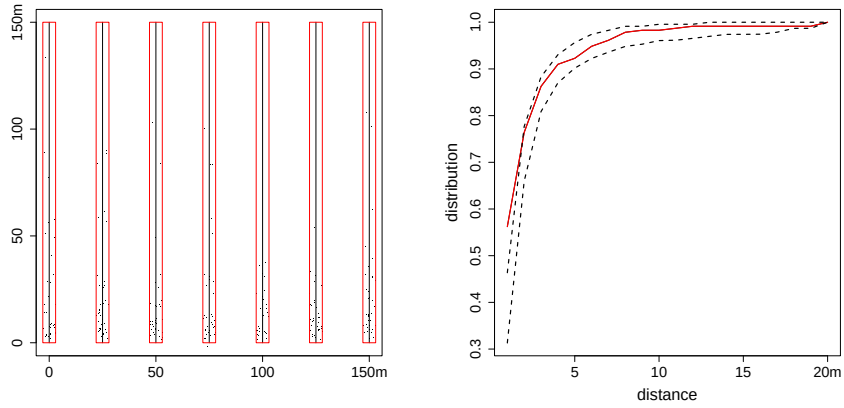


(a) (b)

Figure 15: *Test d'indépendance totale des semis. Figure (a) une réalisation typique sous l'hypothèse d'indépendance, figure (b) Distribution de la distance au plus proche voisin et intervalle de confiance à 95% sous indépendance.*

sont indépendants, i.e. si la variation de densité le long d'un transect est suffisamment faible pour considérer que l'intensité du processus de Poisson non-stationnaire est constante à y donné dans un transect. On part de la même statistique. Chaque simulation consiste à redistribuer chaque point (x_s, y_s) du semis dans son transect, avec la même distance à la haie. Seule son abscisse est donc tirée au hasard dans le segment formé de l'intersection de $y = y_s$ avec le transect d'origine du point. La courbe observée (figure 17(b)) est aux limites de l'intervalle de confiance et on ne rejette pas l'hypothèse à ce niveau.

En conclusion, on ne rejette pas l'hypothèse d'un processus de Poisson non-stationnaire unidimensionnel sur chaque transect, mais dont les caractéristiques dépendent du transect. Plusieurs raisons peuvent conduire au non-rejet de cette dernière hypothèse mais au rejet de la simple hypothèse d'un processus de Poisson non-stationnaire unidimensionnel sur l'ensemble des transects (deuxième hypothèse testée). Soit la fonction de dispersion, c'est à dire la probabilité pour un semis d'être à une distance donnée de la haie, ne dépend que du numéro du transect et de la distance à la haie, et le rejet de la deuxième hypothèse est du à la différence de nombre de semis par transect. Soit la forme de cette fonction de dispersion dépend en plus du



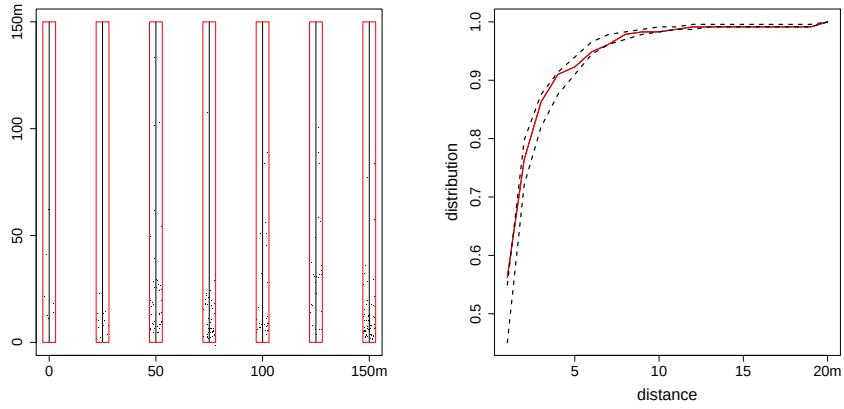
(a) (b)

Figure 16: *Test d'indépendance des semis conditionnellement à la distance à la haie. Figure (a) une réalisation typique sous l'hypothèse d'indépendance conditionnelle, figure (b) Distribution de la distance au plus proche voisin et intervalle de confiance à 95% sous indépendance conditionnelle.*

numéro de transect.

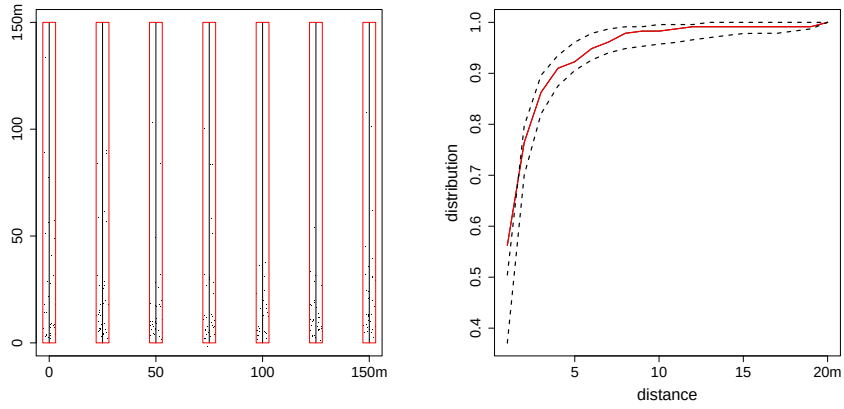
Dans un quatrième temps, nous testons donc cette hypothèse d'égalité des fonctions de dispersion. Nous souhaitons la tester sans faire d'hypothèse sur la forme de cette fonction. Nous allons alors travailler conditionnellement au nombre de semis par transect et conditionnellement aux ordonnées des semis. Nous allons donc (i) conserver les ordonnées des semis, (ii) les redistribuer dans les transects en respectant le nombre initial de semis par transect, (iii) tirer au hasard l'abscisse du semis dans son transect selon une loi uniforme. On travaillera toujours avec la même statistique, la distribution de distance au plus proche voisin. La figure 18(a) illustre une réalisation possible d'un tel tirage. Sur la figure 18(b) est représentée la courbe observée ainsi que sa bande de confiance sous l'hypothèse d'indépendance conditionnellement à la distance à la haie. On note que la courbe observée sort de la bande de confiance pour les petites distances, et l'hypothèse est rejetée.

Pour conclure on notera que, si les tests ont été présentés dans un cadre stationnaire, c'est surtout le mode de construction de ces tests en relation avec les hypothèses formulées qui reste intéressant. On peut alors



(a) (b)

Figure 17: *Test d'indépendance des semis conditionnellement à la distance à la haie et à l'appartenance au transect. Figure (a) une réalisation typique sous l'hypothèse d'indépendance conditionnelle, figure (b) Distribution de la distance au plus proche voisin et intervalle de confiance à 95% sous indépendance conditionnelle.*



(a) (b)

Figure 18: *Test d'égalité des fonctions de dispersion des différents transects. Figure (a) une réalisation typique sous l'hypothèse d'égalité, figure (b) Distribution de la distance au plus proche voisin et intervalle de confiance à 95% sous l'hypothèse d'égalité.*

débroussailler un problème spécifique via une série de tests adaptés.

8 Un processus de fibres ?

Les processus ponctuels sont les processus les plus simples qui n'engendrent pas d'auto-masquage que l'on peut rencontrer dans le plan. Cependant, les objets que l'on peut observer dans l'espace ne sont pas toujours des points mais peuvent être de dimension non-nulle. Dans le plan, ces objets peuvent être classiquement de dimension un (des segments de courbes ou assimilés comme tels), ou de dimension deux (des taches). Nous allons ici regarder le premier cas qui présente de fortes similarités avec le semis de points (toutes les courbes sont a priori observables) mais pose des questions différentes dues à la présence des segments de lignes. Ces processus sont appelés processus de fibres. Si le principe d'analyse reste le même pour un processus de fibres que pour un processus ponctuel, le passage d'un ensemble d'éléments de dimension 0 (les points) à un ensemble d'éléments de dimension 1 (les fibres) va enrichir considérablement la façon de mesurer les distances entre objets (i.e. les statistiques), mais aussi les hypothèses possibles à tester (on a rencontré deux hypothèses d'indépendance pour un simple processus marqué, on va rencontrer d'autre déclinaisons pour les fibres), ainsi que sur les difficultés engendrées (que faire des fibres coupées, avant ou après randomisation ?). Ce chapitre va donc balayer de la même façon que précédemment:

- quelques définitions et notions de base,
- les statistiques possibles permettant de mesurer la distance entre fibres,
- les problèmes de censure et de randomisation associés,
- les tests possibles. Comme avant, on cherchera à les organiser (indépendance totale, puis booléen anisotrope, etc).

8.1 Définition

D'un point de vue probabiliste de tels objets sont modélisés par des processus de fibres, i.e. des processus dont les réalisations sont formées de l'union de courbes rectifiables (c'est à dire localement de longueur finie) de longueur finies. Ces processus sont une classe particulière d'ensembles aléatoires fermés tels qu'étudiés par Matheron (1975). On peut relaxer les

hypothèses de régularité des fibres (Zähle 1982,1983), mais ce cas plus général ne sera pas considéré ici. Ils vont permettre de modéliser des processus pour lesquels une dimension est prépondérante par rapport aux autres. On peut citer par exemple le relevé des trajectoires d'animaux, les traces de fissures dans les coupes de sol, les positions d'arbres abattus par une tempête...

De façon similaire aux processus ponctuels, un processus de fibres Φ est défini par la mesure de la longueur $\Phi(B)$ des courbes observées dans tout ensemble B. On va demander les mêmes propriétés de finitude et d'additivité d'une mesure:

- la longueur $\Phi(B)$ est presque sûrement finie pour B compact,
- $\Phi(B_1 \cup B_2) = \Phi(B_1) + \Phi(B_2)$ si $B_1 \cap B_2 = \emptyset$.

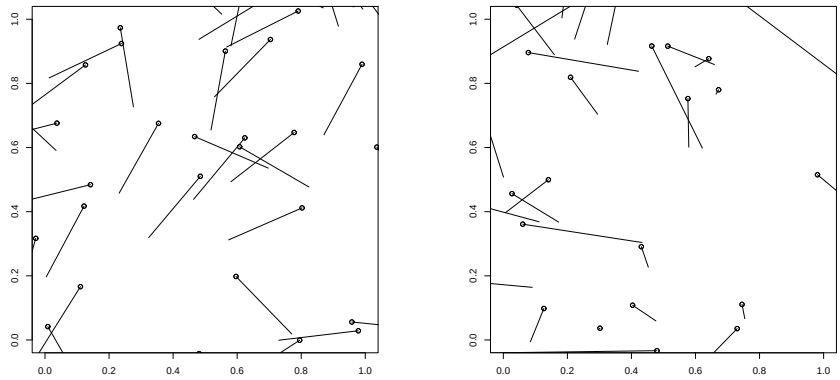
Les notions de mélange, d'ergodicité et de stationnarité présentées pour les processus ponctuels utilisent la notion de mesure indépendamment de sa nature (comptage ou longueur). On va donc les retrouver de façon identique pour les fibres.

La définition précédente est très large, et nous allons considérer par la suite un cadre plus restreint, pour lequel chaque courbe est identifiable et de longueur finie.

- Si on note X_i un point caractéristique de chaque courbe, par exemple son milieu, le processus X de ces points sera appelé processus de points d'origine.
- On notera γ_i la courbe attachée en X_i et centrée sur X_i de sorte que le processus de fibres devient $\Phi = \cup_i X_i \oplus \gamma_i$ (on note $x \oplus B = \{y/y = x + u, u \in B\}$ le translaté de B par x).
- chaque courbe γ_i devient alors une marque attachée au point X_i.

8.2 Un exemple : le processus booléen de segments

L'exemple le plus simple est le processus booléen de segments : partant d'un processus ponctuel de Poisson X, on attache à chaque point un segment dont les propriétés (longueur et orientation) sont tirées au hasard indépendamment d'un point à l'autre. Le point auquel chaque segment est attaché sera au choix le milieu du segment, son extrémité inférieure ou toute



(a) (b)

Figure 19: *Réalisations de processus booléens de segments. Dans les deux cas, le processus ponctuel sous-jacent est d'intensité 20, les segments de longueur moyenne 0.3 et l'orientation uniforme. (a):segments de longueur fixe, (b) segments de longueur aléatoire exponentielle.*

autre position identifiée de façon univoque sur le segment. Le choix n'aura aucun effet sur la structure du processus. La figure 19 présente deux réalisations typiques de processus booléens de fibres pour lesquels l'intensité du processus de Poisson est de 20. Dans le premier cas, les segments sont de longueur fixe, égale à 0.3, et d'orientation uniforme. Dans le deuxième cas, la longueur est distribuée selon une loi exponentielle de longueur 0.3.

Une des différences fondamentales entre un processus ponctuel et un processus de fibres est clairement visible sur les deux figures. On y note en effet des segments interceptant les bords de la fenêtre, ce qui va entraîner des problèmes de censure car on ne connaîtra pas la longueur exacte de ces segments. Si formellement on peut regarder un processus booléen de segments comme un processus ponctuel marqué, le point étant un point d'origine du segment (par exemple son milieu ou son extrémité inférieure), on va rencontrer:

- des segments présents dans la fenêtre pour lesquels on ne peut pas situer le point d'origine,
- des points d'origine présents dans la fenêtre pour lesquels on ne connaît pas la longueur du segment.

On notera aussi que les deux figures, même si elles correspondent à des valeurs égales des paramètres généraux (intensité du processus de points d'origine, longueur moyenne des segments) et donc à la même longueur moyenne de fibre par unité de surface, offrent des patterns très différents et donc des propriétés probabilistes différentes. Dans le premier cas (longueur constante), les observations faites dans deux zones B_1 et B_2 distantes de plus de la longueur d'un segment sont indépendantes, car un segment ne pourra alors pas être à la fois dans les deux zones. La longueur observée dans chacune des zones B_i sera fonction du nombre et de l'orientation de segments dont les points d'origine sont dans des zones disjointes A_i (cette zone A_i est obtenue par l'union, sur les orientations θ possibles des segments, des dilations de B_i par les segments d'orientation θ à partir du point d'origine ; ainsi, si on prend le milieu du segment comme point d'origine, A_i sera formé de B_i et de sa couronne d'épaisseur la moitié du segment). L'indépendance des longueurs de fibre dans les deux zones est assurée par le caractère poissonnien des points d'origine et l'indépendance des orientations. Dans le deuxième cas, la distribution de longueur des segments étant exponentielle, la dépendance diminue avec la distance, mais ne disparaît jamais.

9 Statistiques d'intérêt

Les tests de permutation sur processus de fibres ne sont pas pratiqués de façon aussi systématique que sur processus ponctuels, vraisemblablement du fait du moindre usage de ces processus. On peut cependant construire des statistiques cherchant à explorer différentes facettes du processus, comme cela a été le cas avec les fonctions F, G et H pour les semis de points. On propose donc ici de regarder plus spécifiquement les fonctions suivantes :

9.1 Les fonctions associées au processus ponctuel des points d'origine

On va regarder au travers de ces fonctions les caractéristiques de la distribution des points d'origine. On retiendra par la suite les deux fonctions suivantes:

- $G_0(r)$ la fonction de distribution de distance d'un point d'origine à son plus proche voisin,

- $H_0(r)$ la fonction de distance entre points d'origine.

9.2 Les fonctions associées aux distances entre points appartenant aux fibres

On va chercher à généraliser dans une première étape les fonctions utilisées pour un processus ponctuel:

- $G_1(r)$ la fonction de distribution de distance d'un point quelconque d'une fibre au point le plus proche des autres fibres,
- $H_1(r)$ la distribution de distance entre points de fibres différentes.

9.3 Les fonctions associées aux distances entre fibres

On ne regarde plus la fibre comme le support d'un ensemble de points comme précédemment, mais comme des objets à part entière entre lesquels on peut définir une distance. Nous considérerons ici la distance du minimum, la distance entre deux fibres étant le minimum des distances entre deux points situés chacun sur une fibre. On trouve alors naturellement:

- $G_2(r)$ la fonction de distribution de distance d'une fibre quelconque à la fibre la plus proche,
- $H_2(r)$ la distribution de distance entre fibres.

9.4 Les fonctions associées aux longueurs de fibres

Nous nous restreindrons à une fonction notée ici $S_1(r)$, la longueur moyenne de fibres à distance inférieure ou égale à r d'un point quelconque du processus de fibres. Cette fonction est le correspondant direct de l'estimation de $\lambda K(r)$, le produit de l'intensité du processus ponctuel par la fonction de Ripley, dans le cas d' un processus ponctuel. En effet, $\lambda K(r)$ estime le nombre moyen de points à distance r d'un point quelconque du processus. Cependant, notre objectif étant un objectif de test et non d'estimation, nous ne ferons pas avec S_1 de correction d'effet de bord, celui-ci sera intégré dans la statistique.

9.5 Les fonctions associées aux caractéristiques des fibres

Enfin, on peut extraire une caractéristique a_i de chacune des fibres γ_i et l'utiliser comme marque des points d'origine X_i pour laquelle on pourra reprendre des critères classiques comme le variogramme. On pourra aussi traiter l'ensemble (X_i, a_i) comme un processus ponctuel multivarié si le critère s'y prête ainsi que cela a été proposé en 6.3.

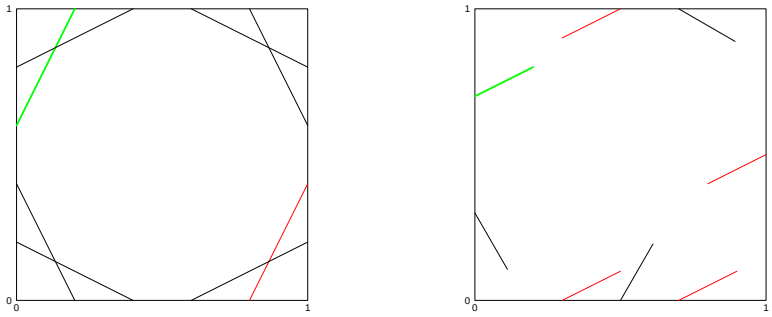
10 Problèmes de Monte-Carlo

Comme nous pouvons le constater dans le cas des processus ponctuels (paragraphe 5), les tests de Monte-Carlo reviennent à définir l'ensemble de n-uplets $\mathcal{Y}$ de W^n, chaque n-uplet correspondant à une redistribution des points, tel que tout $(y_1, ..., y_n) \in \mathcal{Y}$ ait la même probabilité que $x_1, .., x_n$ sous l'hypothèse à tester. Dans ce cadre le test revient alors à comparer la valeur observée de la statistique d'intérêt $f_0 = f(x_1, .., x_n)$ à la distribution de $f_y = f(y_1, ..., y_n)$ dans $\mathcal{Y}$ muni de la loi uniforme. Ainsi, dans le cas du test d'indépendance total, toutes les positions possibles de n points dans W étaient équiprobables, et $\mathcal{Y} = W^n$.

Dans le cas d'un processus de segments on dispose d'une information plus riche que la seule position de points. On dispose en effet de mesures d'angles et de longueurs. On dispose donc a priori d'un ensemble de réalisations possibles beaucoup plus vaste. A l'inverse, on a vu que l'on est confronté à un problème de censure, qui va intervenir dès que l'on souhaitera randomiser les fibres. En l'absence de la connaissance de la distribution de longueur des fibres, on cherche des réalisations de même probabilité que la réalisation originale, ayant le même nombre de fibres avec les mêmes longueurs. La censure, donc l'absence de connaissance de la longueur des fibres, va donc engendrer des difficultés que nous allons rapidement classer ici.

10.1 Fibres doublement censurées

Ces fibres interceptent deux fois le bord de la fenêtre d'observation. On ne connaît donc pas la longueur de ces segments ni aucune de leurs extrémités. Seules sont connues l'orientation θ_2 de chacun de ces segments ainsi que la longueur l_2 de son intersection avec la fenêtre. Un tel segment est représenté



(a) (b)

Figure 20: *Exemples de permutations possibles dans un carré d'un segment donné interceptant deux cotés (en a) ou un seul coté (en b) respectant la longueur. Le segment original est en trait gras vert. Les segments en rouge sont des exemples permettant de plus le respect de l'orientation du segment.*

en trait gras vert sur la figure 20(a).

Si on considère que ce segment est distribué aléatoirement uniformément, et que son orientation est uniforme, alors tout segment interceptant deux fois la bordure et dont la longueur de la partie interceptée vaut également l_2 est de même probabilité. Une fois l'une des intersections fixée, on aura

- 0 segment possible (si l_2 est plus petite que la distance de l'intersection aux sommets du carré),
- 1 segment possible (si l_2 est plus petite que la moitié du coté du carré et plus grande que la distance de l'intersection aux sommets du carré)
- 2 segments possibles (si l_2 est plus grande que la moitié du coté du carré).

De tels segments possibles sont dessinés en traits fins en figure 20(a).

Si on rejette l'hypothèse de distribution uniforme des angles, seul le segment tracé en rouge aura la même longueur l_2, et la même orientation θ_2. On restreint alors les possibilités de façon drastique.

10.2 Fibres simplement censurées

Ces fibres n'interceptent qu'une seule fois la fenêtre d'observation. Une telle fibre est représentée en gras dans la figure 20(b). Pour une telle fibre, on connaît une extrémité, la longueur l_1 de son intersection avec la fenêtre et son angle θ_1.

Sous l'hypothèse de répartition uniforme (position et orientation) d'une telle fibre dans le carré de coté $1 + 2l_1$ centré sur le carré original, tout autre segment, interceptant une fois la bordure et de même longueur l_1 dans le carré de coté 1, est de même probabilité d'observation que la fibre originale. L'ensemble $W(l_1)$ des segments de même probabilité que la fibre de départ et conservant la même longueur d'intersection l_1 avec le carré de départ est donc formé des segments interceptant la bordure de la fenêtre une seule fois et de même longueur. Sous cette hypothèse, et si on considère l'ensemble des points d'intersection des fibres de $W(l_1)$ avec le bord de la fenêtre, on voit que les points situés au milieu d'un coté de la fenêtre sont plus représentés que les points situés sur le bord de la fenêtre c'est à dire dans un angle, dans la mesure où plus de segments sont possibles dans le premier cas que dans le deuxième (dans un rapport de $\pi/(\pi/2)$). De tels segments sont représentés en traits fins sur la figure 20(b).

Si on rejette l'hypothèse de distribution uniforme des angles, seuls sont possibles les segments de même longueur l_1 parallèles au segment original. En particulier, on ne retrouvera plus de segments possibles dans certains angles (les coins supérieur gauche et inférieur droit par exemple en 20(b)). Quelques exemples de segments possibles sont donnés en rouge sur la figure 20(b).

10.3 Fibres non censurées

Ces fibres posent a priori moins de difficultés puisque l'on connaît leurs deux extrémités, leurs longueurs et leurs orientations.

Pour une telle fibre, sous l'hypothèse de répartition uniforme, tout autre segment de même longueur n'interceptant pas les bords sera de même probabilité que le segment d'origine, si on accepte une hypothèse d'orientation uniforme. On peut noter cependant que la mesure de l'ensemble des segments disponibles se rétrécit avec l'augmentation de la longueur de la fibre.

Si on rejette l'hypothèse d'orientation uniforme, les segments de même longueur et orientation n'interceptant pas les bords auront également la

même probabilité que le segment de départ.

11 Tests

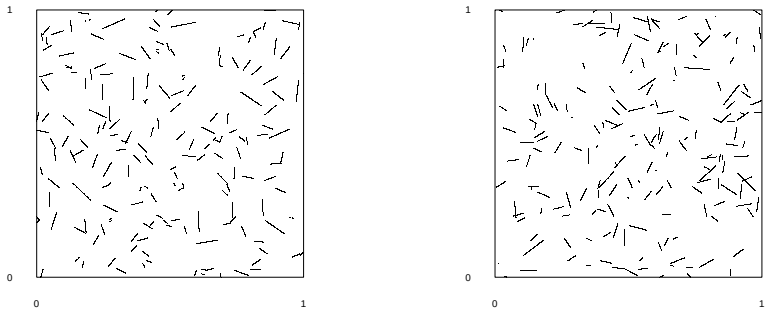
Comme précédemment, on va dans un premier temps chercher à tester si le processus est un processus booléen de segments d'orientation uniforme, l'équivalent du hasard total. Ensuite, on cherchera à tester différentes sous-hypothèses permettant de raffiner le test de départ s'il a été rejeté. On testera plus précisément:

- si le processus de points d'origine est poissonien
- si les segments sont affectés indépendamment entre points d'origine
- si les orientations sont indépendantes et uniformes
- si les orientations sont indépendantes

11.1 Test d'indépendance totale (processus booléen isotrope)

On veut tester si l'ensemble de fibres observées dans la fenêtre suit un processus booléen de segments où les segments sont orientés aléatoirement uniformément. On va donc randomiser les segments observés dans la fenêtre indépendamment les uns des autres, en respectant les contraintes signalées au paragraphe 10. Comme dans le cas des processus ponctuels deux approches sont possibles. Si l'on dispose d'une contre-hypothèse spécifique, on prendra les statistiques les plus sensibles à cette contre-hypothèse, de façon à prendre les tests les plus puissants possibles. Si on se situe dans un cadre plus exploratoire où aucune contre-hypothèse n'est a priori privilégiée, on cherchera un ensemble de statistiques permettant de couvrir au mieux un ensemble de propriétés du semis de fibres.

Considérons l'exemple de la figure 21. La figure 21(a) est une réalisation d'un processus de segments dépendants (et donc un processus de segments qui n'est pas booléen) construit de la façon suivante : le semis de points d'origine (X_i) est un processus de Poisson d'intensité 200. On calcule pour chaque point sa distance d_i à son plus proche voisin. Le processus de fibres est obtenu en attachant à chaque point un segment d'orientation aléatoire et de longueur $l_i = d_i$. La figure 21(b) est une réalisation d'un processus



(a) (b)

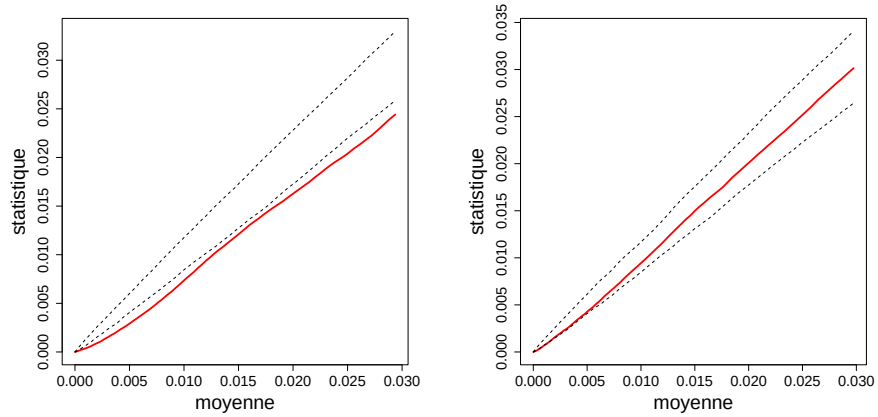
Figure 21: *Exemples de semis de fibres. (a) Semis de fibres dépendant. (b) Semis booléen de fibres de même distribution de longueur et d'angle.*

booléen de segments de même distribution de longueur et d'angle (obtenue par randomisation de la réalisation précédente).

La figure 22 présente les résultats pour la statistique $H_1(r)$ de distribution de distance entre points sur des fibres différentes. En figure 22(a), la courbe observée est en dessous de la bande de confiance, permettant de rejeter l'hypothèse booléenne, et de conclure que deux points pris au hasard sur deux fibres différentes sont en moyenne plus éloignés dans le semis de fibres observé que sous l'hypothèse booléenne. En 22(b), le test ne rejette pas l'hypothèse.

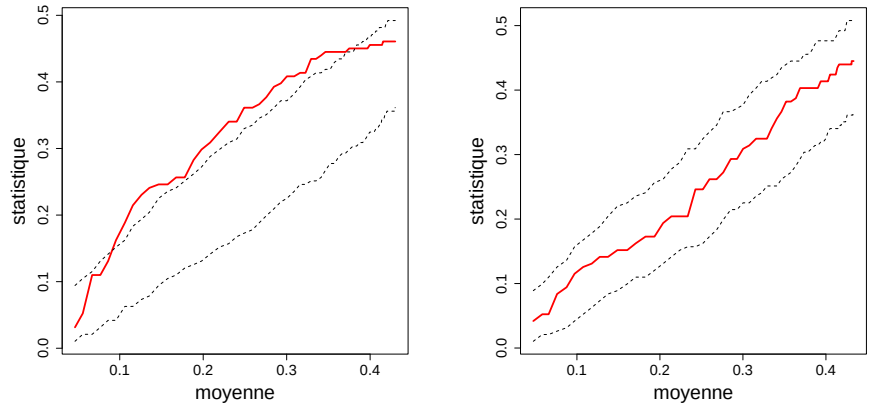
La figure 23 présente les mêmes tests effectués avec $G_2(r)$, la distance entre fibres. En 23(b), on ne rejette pas l'hypothèse d'indépendance totale. En 23 (a), l'hypothèse est rejetée, et l'on constate que la distance moyenne entre fibres proches est plus petite que sous le hasard.

Le deuxième test amène à penser que le semis de fibres est agrégé, deux points pris au hasard étant plus proches que sous le hasard. Le premier conclut plutôt à un semis de fibres régulier, car deux fibres au hasard sont plus éloignées que sous le hasard. Cette contradiction apparente s'explique d'un point de vue mathématique par la prise en compte des longueurs dans le premier test (les fibres y sont représentées proportionnellement à leur longueur) alors que les fibres ont toutes le même poids dans le second. Les notions d'agrégation et de régularité, évidentes pour un processus de points, perdent



(a) (b)

Figure 22: *Test d'indépendance totale : distribution de distance entre points situés sur des fibres différentes . Abscisse : valeur moyenne de la statistique sous indépendance totale, ordonnée : distribution observée (trait plein) et bande de confiance à 95%. (a) semis de fibres dépendant. (b) semis booléen de fibres de même distribution de longueur et d'angle.*



(a) (b)

Figure 23: *Test d'indépendance totale : distribution de distance au plus proche voisin entre fibres. Abscisse : valeur de la statistique sous indépendance totale, ordonnée : distribution observée (trait plein) et bande de confiance à 95%. (a) semis de fibres dépendant. (b) semis booléen de fibres de même distribution de longueur et d'angle.*

leur sens dans le cas de semis de fibres si on ne précise pas le sens que l'on donne à ces notions. Dès que l'on aborde l'étude de semis un peu complexes, il est donc important avant de tester l'existence d'une notion de ce type d'en donner une définition exacte, de formuler la question en conséquence et de bien préciser le choix de la statistique employée en fonction de l'hypothèse à tester.

11.2 Test d'indépendance des positions des fibres conditionnel aux angles observés (processus booléen anisotrope)

Comme mentionné au paragraphe 8.1, la notion d'isotropie se définit facilement pour un processus de fibres comme pour un processus de points. Par contre, plusieurs notions d'anisotropie sont possibles si on regarde un processus de fibres. En effet, l'anisotropie peut se manifester au niveau des points d'origine comme au niveau des fibres. Dans le premier cas, on pourra par exemple avoir un semis de fibres formé de paires de fibres placées l'une au dessus de l'autre, mais d'orientation indépendantes. Dans le deuxième cas, les paires de fibres peuvent être placées sans disposition préférentielle l'une part rapport à l'autre, mais parallèles. Remarquons au passage que l'on peut retourner le problème sur les semis de points, si par exemple on considère un processus ponctuel de Cox porté par un processus de fibres : partant des deux exemples ci-dessus, les semis de points sont obtenus en jetant au hasard des points le long des segments.

Nous nous intéressons ici à tester si un semis de fibres est issu d'un processus booléen de fibres anisotropes, c'est à dire si les points d'origine sont issus d'un processus de Poisson stationnaire, et si les fibres qui y sont attachées sont indépendantes, mais issues d'une distribution anisotrope. Un exemple de processus booléen anisotrope est donné en figure 24. Partant d'un semis de points issus d'un processus ponctuel poissonien d'intensité 200, on attache à chaque point un segment de longueur 0.05, d'orientation aléatoire uniforme dans le segment $[\pi/2, \pi]$.

Pour tester une telle hypothèse, on va alors restreindre l'espace des randomisations à celles qui conservent les angles des fibres, telles que décrites au paragraphe 10. En particulier, chaque fibre non censurée est redistribuée au hasard, en respectant sa direction, dans la sous-fenêtre où elle ne rencontrera pas les bords de la fenêtre d'observation. Chaque fibre interceptant une fois

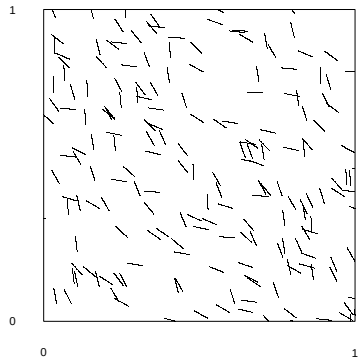


Figure 24: *Réalisation d'un processus booléen de fibres d'intensité 200, de longueur 0.05, d'orientation uniforme dans le segment* $[\pi/2, \pi]$.

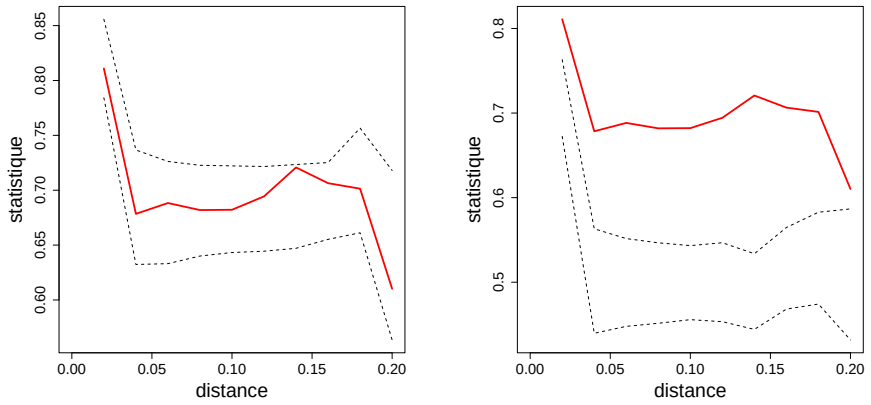
le bord est redistribuée au hasard sur le bord, en respectant l'orientation et la longueur de fibre considérée. Bien sur, en respectant l'orientation des fibres, on diminue l'espace des randomisations possibles et on perd donc de la puissance de test.

Le problème posé est aussi un problème d'orientation, on va donc choisir un critère qui intègre cette notion. Nous allons ici prendre un critère proche du variogramme, le carré moyen du cosinus des angles de deux segments à distance donnée. Plus exactement, on prendra comme critère

$$C(r) = \frac{1}{N_r} \sum_{(i,j)\in\mathcal{C}_r} \cos(\theta_i - \theta_j)^2$$

où N_r désigne le nombre de couples de segments dont la distance est dans l'intervalle $[r, r+\delta]$, $\mathcal{C}_r$ l'ensemble de ces couples et θ_i les orientations des fibres par rapport à un repère donné.

La figure 25 présente les résultats obtenus pour des classes de distance de $\delta = 0.02$ sur le semis de fibres de la figure 24. En figure 25(b), nous avons procédé à une randomisation totale (hypothèse booléenne isotrope) et on rejette fortement l'hypothèse booléenne isotrope. Le test de l'hypothèse booléenne anisotrope, effectué conditionnellement aux caractéristiques des



(a) (b)

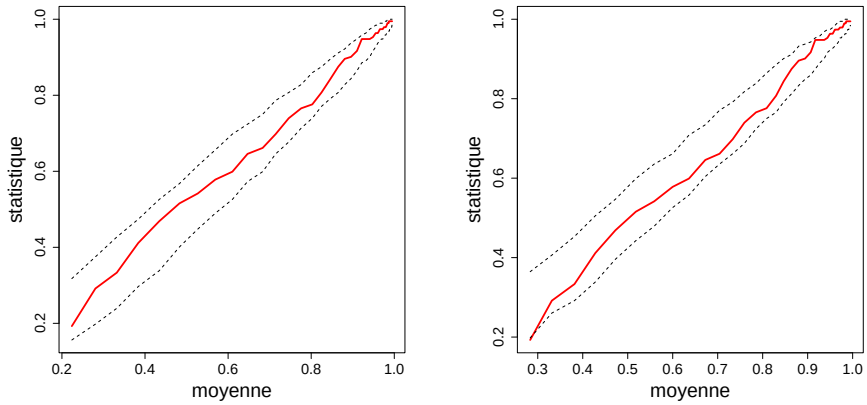
Figure 25: *Test de l'hypothèse booléenne anisotrope. Critère de test : cos^2 moyen de l'angle de deux fibres à distance donnée. (a) test conditionnel aux angles observés, (b) test d'hypothèse booléenne isotrope. Abscisse : distance, ordonnée : distribution observée (trait plein) et bande de confiance à 95%.*

fibres est accepté tel qu'indiqué en figure 25(a).

L'importance du choix du critère est illustré par la figure 25 et la figure 26 où nous avons repris la distribution de la distance du plus proche voisin entre fibres sur le même exemple. En Figure 25, le critère basé sur l'angle entre fibres permet de rejeter fortement l'hypothèse booléenne isotrope. Si les distances entre fibres sont fortement influencées par les angles de ces fibres pour des positions des points d'origine fixés, cette dépendance s'atténue fortement si les points d'origine ne sont plus fixés, mais varient uniformément indépendamment dans le carré. De ce fait, le critère de distance du plus proche voisin entre fibres (donné en figure 26) n'est pas discriminant, seul le tout début de la courbe observée s'écartant très faiblement de la bande de confiance.

11.3 Test d'indépendance des points d'origine

Les deux tests précédents permettent de tester si les fibres sont réparties indépendamment les unes des autres dans l'espace. Si ces tests sont rejetés, on va chercher à distinguer parmi différentes causes possibles, les principales



(a) (b)

Figure 26: *Test de l'hypothèse booléenne anisotrope. Critère de test : distribution de distance au plus proche voisin entre fibres. (a) test conditionnel aux angles observés, (b) test d'hypothèse booléenne isotrope. Abscisse : valeur de la statistique sous indépendance totale, ordonnée : distribution observée (trait plein) et bande de confiance à 95%.*

d'entre elles étant d'une part l'écart à la répartition poissonnienne des points d'origine, d'autre part la non-indépendance des affectations des fibres. D'un point de vue statistique, le choix du point d'origine de chaque fibre importe peu si le processus de fibres est un processus booléen. En effet, si un processus de segments est un processus booléen, le semis de points des milieux comme celui de l'une des extrémités (par exemple la gauche) sera poissonnien. Si par contre on sort de ce cadre, sa définition redevient importante. Prenons le cas extrême d'un processus de segments de longueur constante où le processus de semis d'origine est formé des nœuds d'une grille régulière. Si chaque fibre est attachée à son point d'origine par le milieu et si les orientations sont aléatoires uniformes, le processus des extrémités inférieures ne sera pas formé des noeuds d'une grille régulière, et s'en écartera d'autant plus que la longueur des segments est importante. On peut toujours définir le processus de points d'origine d'un point de vue géométrique (milieu, extrémité particulière...). Cependant, étant donné que la distribution des points d'origine est dépendante de leur nature, il pourra être intéressant par la suite que le processus de points d'origine ait un sens physique ou biologique que nous pouvons distinguer sur le semis de fibres. On pourra par exemple prendre la

position de la tête s'il s'agit d'animaux. Pour des processus physiques, on pourra considérer le centre de gravité si chaque fibre est un corps solide sur lequel s'appliquent des forces diverses.

La nature des points d'origine étant connue, tester l'indépendance des points d'origine correspond à tester un modèle où les points d'origine se répartissent au hasard, mais où la distribution des fibres peut être dépendante. Ainsi, dans le cas d'un troupeau, on peut tester si la distribution des têtes est poissonnienne, i.e. les animaux n'ont ni comportement grégaire ni répulsif au niveau de l'occupation du territoire observé, alors que l'orientation d'un animal peut dépendre de la position ou de l'orientation de ses voisins.

Techniquement, tester cette hypothèse relève du chapitre précédent, puisqu'il s'agit du test d'hypothèse CSR d'un processus ponctuel. Ainsi, dans la mesure où le test porte sur la distribution des points d'origine sans hypothèse sur la distribution des fibres à processus d'origine connu, le critère de test devra porter sur les points d'origine eux-mêmes (par exemple les fonctions $G_0(r)$ ou $H_0(r)$). La différence pourra venir de leur identification par rapport aux caractéristiques des fibres et des questions de censure que cela va engendrer:

- si on considère un processus de segments de longueur fixée, on saura situer le milieu de chaque fibre, que celle-ci soit complète ou non, dès que l'on connaîtra une extrémité dans le carré. Si les longueurs sont aléatoires, ce n'est plus le cas. On peut alors ne pas reconnaître la position de points d'origine situés dans la fenêtre.

- ce n'est pas parce qu'on connaît la position d'un certain nombre de points d'origine hors fenêtre que l'on connaît de façon complète la position des points d'origine dans la bordure de la fenêtre. Les fibres hors-fenêtre parallèles aux bords ne sont pas détectées, même si elles sont très proches.

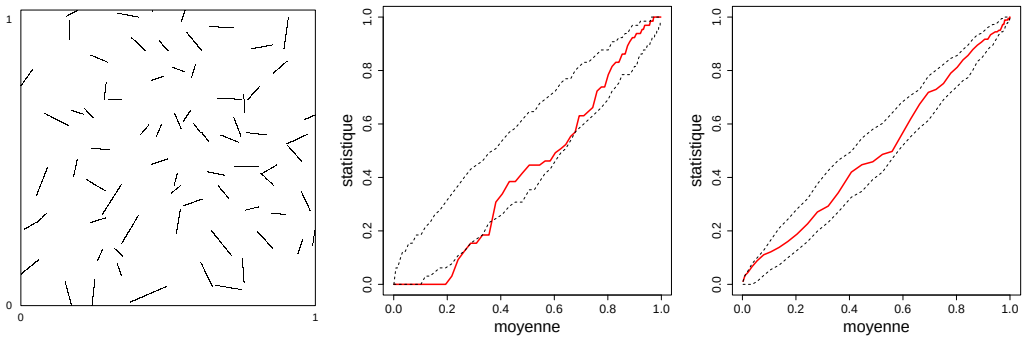
11.3.1 Si tous les points d'origine présents dans la fenêtre sont identifiés

Le cas va se présenter si les points d'origine sont définis hors propriétés géométriques (tête d'un animal par exemple), ou si on choisit comme point d'origine un point extrême comme le point inférieur (gauche si problème d'unicité). Ainsi que mentionné plus haut, ce ne sera pas le cas si on choisit comme point d'origine le milieu d'un segment car dans le cas de segments

censurés deux fois, on ne saura pas situer le milieu, et dans le cas de segments censurés une fois, on ne le pourra pas si la longueur des segments n'est pas connue.

Ce cas relève exactement du chapitre 5. On calculera les statistiques observées $F_0(r)$, $G_0(r)$ et $H_0(r)$, et on les positionnera par rapport à leurs bandes de confiance sous répartition aléatoire uniforme, obtenues par méthode de Monte-Carlo en redistribuant les point d'origine au hasard dans la fenêtre d'observation.

11.3.2 Si seulement une partie de ces points est identifiée



(a) (b) (c)

Figure 27: *Test de l'hypothèse de répartition aléatoire des points origine. Les points origine sont les milieux des segments. Critère de test : distribution de distance au plus proche voisin entre points origine calculé sur le plus grand rectangle inclus dans le carré unitaire où l'ensemble des points origine présent sont connus. (a) Processus de fibres dépendant basé sur un processus hard-core construit à partir d'un processus de Poisson d'intensité 100, de distance 0.035 auquel sont attachés des segments d'orientation aléatoire et de longueur proportionnelle (rapport de 10) à la distance au plus proche voisin. (b) Test appliqué à la figure(a) . Abscisse : valeur de la statistique sous indépendance totale, ordonnée : distribution observée (trait plein) et bande de confiance à 95%. (c) Même test appliqué sur la figure 21(a)*

Ce sera typiquement le cas si les points d'origine sont les milieux des

segments, ce que nous considérerons par la suite (les autres cas se règlent de manière similaire). Notons W la fenêtre d'étude. Sous l'hypothèse de répartition poissonnienne des points d'origine, chaque point X_i identifié est distribué, connaissant sa fibre associée γ_i, au hasard dans la zone W_i de la fenêtre telle que, si $y \in W_i$, $y \oplus \gamma_i$ ne rencontre pas le bord de la fenêtre. Notons par la suite $\{X_1, .., X_n\}$ l'ensemble des points d'origine, $X_1, .., X_p$ ceux dont les fibres attachées ne touchent pas le bord (et sont donc identifiables) et l'ensemble des conditions d'intersections des fibres $\mathcal{C}(X, \gamma) = \cap_{i \leq p}((X_i \oplus \gamma_i) \cap \partial W = \emptyset) \cap_{i>p} ((X_i \oplus \gamma_i) \cap \partial W \neq \emptyset)$ où ∂W dénote le bord de W. Soit $x_i \in W_i$ pour $i \leq p$, alors la probabilité d'observer $(X_1, .., X_p) = (x_1, .., x_p)$ connaissant $\mathcal{C}(X, \gamma)$ s'écrit

$$p((X_1, .., X_p) = (x_1, .., x_p) \mid \mathcal{C}(X, \gamma)) = \frac{p(\mathcal{C}(X, \gamma) \mid (X_1, .., X_p) = (x_1, .., x_p))}{p(\mathcal{C}(X, \gamma))} p((X_1, .., X_p) = (x_1, .., x_p))$$

et $p((X_1, .., X_p) = (x_1, .., x_p))$ est la loi uniforme sur le produit des W_i.

- Soit la distribution des fibres associées γ dépend de la position des points d'origine X, et la loi de X connaissant $\mathcal{C}(X, \gamma)$ dépend effectivement de γ. Dans ce cadre, on se restreindra à étudier la distribution des points d'origine connus dans une zone W' où tous ces points sont connus. Si la fenêtre d'observation W est rectangulaire, on peut prendre par exemple le plus grand rectangle W_1 tel que W_1 ne contienne aucun point d'origine possible des fibres pour lesquelles on ne connaît pas le point d'origine et on appliquera aux seuls points contenus dans W_1 un test d'indépendance totale dans W_1. La figure 27 présente le résultat d'un tel test sur deux exemples. Le premier exemple est illustré en figure 21(a), un processus de fibres dépendant basé sur un processus ponctuel poissonnien. Le test correspondant à cette figure est donné en figure 27(c), où l'hypothèse de distribution aléatoire uniforme du semis de points d'origine n'est pas rejetée. Le deuxième exemple est donné en figure 27(a), où le processus de points d'origine est un processus hard-core de rayon 0.035, construit à partir d'un processus ponctuel d'intensité 100. Chaque fibre est proportionnelle à la distance du point d'origine à son plus proche voisin (10 fois plus grande). Le test correspondant en figure 27(b) rejette bien l'hypothèse d'une distribution aléatoire malgré le faible nombre de fibres présentes.

- Soit la distribution des fibres associées ne dépend pas des positions X. On peut par exemple imaginer que l'orientation de chaque animal d'un troupeau ne dépend que de l'orientation des animaux qui leur sont hiérarchiquement supérieurs. Dans de tels cas, $p(\mathcal{C}(X,\gamma) \mid X) = p(\mathcal{C}(X,\gamma))$, et $p(X \mid \mathcal{C}(X,\gamma)) = p(X)$, En particulier, les positions des points d'origine conditionnellement aux caractéristiques des fibres sont indépendantes entre elles. On va donc travailler sur l'ensemble de points d'origine identifiés dans W dont les fibres ne rencontrent pas le bord de la fenêtre. Conditionnellement aux fibres observées, toute réalisation obtenue en redistribuant au hasard la position de chacun de ces points d'origine X_i dans W_i est de même probabilité que la réalisation d'origine, ce qui permettra de bâtir la bande de confiance sur l'ensemble des points d'origine identifiés et dont les fibres ne touchent pas le bord de la fenêtre. On va ainsi, lorsque cela est possible, travailler sur un échantillon plus important que dans le cas précédent et disposer d'un test plus puissant.

11.4 Et les points d'origine hors-cadre ?

Reprenons le cadre fourni par le schéma présenté ci-dessus

- dans le cas où les caractéristiques des fibres et les positions des points d'origine ne sont pas indépendantes, les points hors-cadre ne sont pas utilisés. En effet, les prendre en compte signifierait prendre une zone W' plus grande que la fenêtre, alors que tous les points d'origine n'y sont pas connus (ceux dans $W' \backslash W$ parallèles au bord de W).

- dans le cas où il y a indépendance, on pourra les intégrer, après avoir redéfini W_i comme l'ensemble des y tels que $y \oplus \gamma_i$ intercepte la fenêtre et permet la reconnaissance de y.

11.5 Test d'indépendance de l'affectation des segments

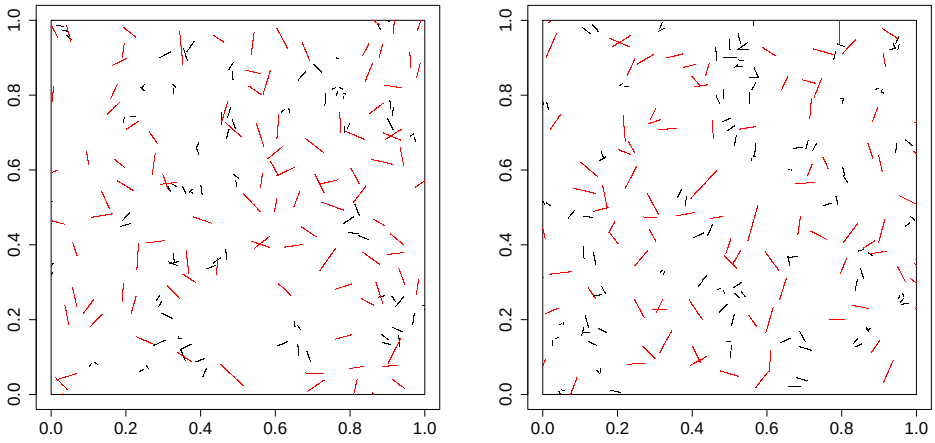
Si les tests d'hypothèse booléenne et si l'hypothèse de distribution aléatoire des points d'origine est également rejetée, on veut savoir si le rejet de l'hypothèse booléenne n'est due qu'à la distribution des points d'origine, (i.e. les fibres attachées aux points d'origine sont tirées au hasard indépendamment l'une de l'autre), ou si une dépendance entre fibres existe également. On peut

distinguer trois cas représentant trois tests faits sous des hypothèses de plus en plus générales :

- La distribution des fibres est isotrope. On peut alors travailler conditionnellement à l'affectation des longueurs et tester la dépendance angulaire entre fibres. Sous l'hypothèse d'isotropie et d'indépendance des marques, le semis de fibres original et le semis de fibres obtenu en tirant au hasard selon une loi uniforme l'orientation de chaque fibre totalement observée ont la même probabilité, si on conditionne la loi uniforme à l'ensemble des angles tels que la fibre reste totalement observée. On pourra alors utiliser un critère comme le carré moyen du cosinus de l'angle entre deux fibres à distance donnée (cf paragraphe 11.2).

- La distribution de longueur et la distribution angulaire sont indépendantes mais la distribution angulaire n'est a priori pas uniforme. Si $(l_1, ..., l_n)$ sont les longueurs observées, $(\theta_1, ..., \theta_n)$ les angles observés des fibres n'interceptant pas le bord, tout couple de vecteurs $l_\phi = (l_{\phi(1)}, .., l_{\phi(n)})$ et $\theta_\psi = (\theta_{\psi(1)}, .., \theta_{\psi(n)})$ où ϕ et ψ sont deux permutations aléatoires indépendantes de $\{1, ..n\}$ appartenant au sous-ensemble des permutations tel que les fibres de longueur l_ϕ et d'angle θ_ψ ne rencontrent pas le bord de la fenêtre d'observation (les fibres touchant les bord restant inchangées) a même probabilité d'être observé que le couple original. En pratique, on procédera par un algorithme d'acceptation/rejet: on prendra deux permutations au hasard et on calculera les fibres correspondantes. Si l'une de ces fibres coupe le bord de la fenêtre, on recommence l'opération, sinon, on accepte la simulation. Tous les critères jugés pertinents pour juger de la dépendance angulaire (cosinus carré par exemple) ou de la dépendance des longueurs (variogramme par exemple) pourront ensuite être employés comme critères de test.

- On veut tester si les fibres sont indépendantes, mais on ne peut rien supposer de leur distribution. Si $(\gamma_1, ..., \gamma_n)$ est le vecteur de fibres incluses dans la fenêtre, alors tout vecteur $(\gamma_{\phi(1)}, ..\gamma_{\phi(n)})$ où ϕ est une permutation aléatoire de $\{1, ..n\}$ a même probabilité d'être observé que le vecteur original, si on contraint les permutations au sous-ensemble tel que ces fibres ne rencontrent pas le bord de la fenêtre d'observation (les fibres touchant les bords restant là aussi inchangées). La même



(a) (b)

Figure 28: *Exemples de réalisations de processus de fibres bivariés. (b) processus dépendants : semis de points d'origine poissonien d'intensité 200. On construit un premier processus de fibres de longueur égale à la distance au plus proche voisin. Chaque processus final est formé de l'ensemble des fibres les plus grandes (en rouge) ou les plus petites (en noir). (a) processus indépendants. On effectue deux réalisations indépendantes du processus décrit en (b), et on superpose de façon indépendante le noir de la première réalisation avec le rouge de la seconde réalisation.*

procédure et les mêmes tests que précédemment peuvent ensuite être employés.

11.6 Test d'indépendance entre deux semis de fibres

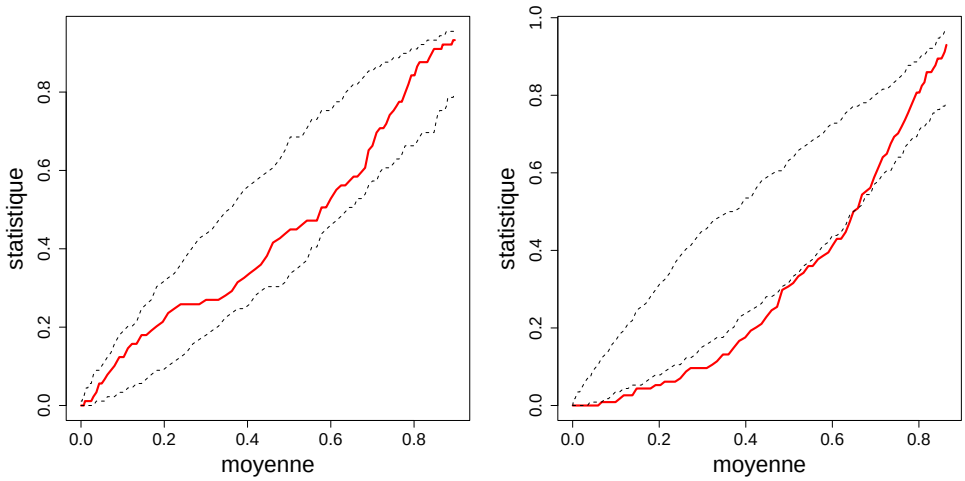
Le problème envisagé dans ce paragraphe est l'analogue direct du problème posé sur semis de points et présenté au chapitre 6. Deux types de fibres étant identifiés, on cherche à savoir si ces deux processus sont indépendants. On retrouve bien sûr les mêmes hypothèses, soit le marquage des fibres d'un processus de fibres est un marquage aléatoire, soit deux processus de fibres, chacun avec sa structure propre, ont été superposés de façon indépendante. Ramenés au niveau des points d'origine, ces deux cas se traduisent de la même

façon : soit le marquage du processus de points d'origine par les fibres est un marquage aléatoire ,soit deux processus de fibres, chacun avec sa structure propre, ont été superposés de façon indépendante. Une fois défini un critère approprié de mesure inter-processus, par exemple la distribution de distance au plus proche voisin entre fibres de processus différents, on va tester les deux hypothèses sur un schéma similaire au cas de semis de points :

- Marquage aléatoire. On veut tester que les marques ont été réparties aléatoirement entre les deux processus. On est donc dans le cadre du paragraphe précédent, seul change le critère de test, car il va mesurer la ressemblance entre les deux semis. On pourra par exemple utiliser comme critère le cosinus carré de l'angle entre deux fibres venant de chacun des semis.

- Superposition indépendante. Chacun des deux processus ayant sa propre structure, on veut tester si la superposition des deux processus est indépendante. On aimerait pouvoir reprendre la procédure envisagée pour les semis de points, en translatant au hasard l'un des processus par rapport à l'autre. Si la taille des fibres est faible par rapport à la taille de la fenêtre, on pourra reprendre la convention du tore et effectuer des translations aléatoires. L'approximation faite dans le cas des fibres sera comparable à celle que l'on fait pour un processus ponctuel. Si la taille des fibres n'est pas négligeable, il faut envisager d'autres procédures. Si l'un des deux processus est isotrope, on va utiliser le cercle inscrit dans la fenêtre. On restreint le processus à son intersection avec le disque, et l'on fait subir au processus isotrope une rotation aléatoire autour du centre du disque. On construira ainsi une bande de confiance du critère par rapport à laquelle on situera la valeur du critère sur le processus observé. Dans cette procédure, les points au centre du cercle se déplacent peu l'un par rapport à l'autre. On peut envisager de modifier la statistique de distance, par exemple en la pondérant selon la distance au centre, pour donner un poids plus égal à l'ensemble des fibres observées dans le disque. Si aucun des deux processus n'est isotrope, on peut restreindre le calcul du critère à une sous fenêtre distante de l des bords, et translater l'un des processus d'une translation aléatoire de module inférieur à l.

La figure 28 illustre deux réalisations de semis de fibres bivariés, l'un indépendant en Figure 28(a), l'autre dépendant en Figure 28(b). Partant



(a) (b)

Figure 29: *Tests d'indépendance entre les semis de fibres rouge et noir présentés en figure 28. Test effectué sur la distance au plus proche voisin d'une fibre rouge à une fibre noire restreintes au cercle de centre (0.5,0.5) de rayon 0.5. Randomisation par rotation aléatoire. (a) test sur figure 28(a), (b) test sur réalisation 28(b).*

d'un semis de points d'origine poissonnien d'intensité 200, on calcule pour chaque point sa distance au plus proche voisin et on attache à chaque point d'origine une fibre d'orientation aléatoire uniforme et de longueur dix fois la distance au plus proche voisin. Le processus bivarié obtenu en Figure 28(b) est obtenu en regroupant les fibres les plus longues (en rouge) ou les plus courtes (en noir). Le processus en Figure 28(a) est obtenu en calculant deux réalisations du processus bivarié précédent, puis en superposant de façon indépendante le semis de longues fibres de la première réalisation avec le semis de courtes fibres de la seconde. Dans les deux cas, chaque semis rouge ou noir n'est pas booléen de par la dépendance entre fibres et la distribution de son semis de points d'origine. De plus, les semis de fibres rouges (resp noires) sont issus du même processus et suivent donc la même distribution.

Le test d'interaction décrit en 11.6 (superposition) est effectué pour chacune de ces réalisations en restreignant les semis au cercle de centre (0.5,0.5) et de rayon 0.5. On prend comme critère de comparaison la distribution de distance d'une fibre rouge à la plus proche fibre noire. Les figures 29(a)

et 29(b) illustrent les résultats de ce test. L'hypothèse d'indépendance est rejetée en figure 29(b) . Elle ne l'est pas en Figure 29(a), la courbe observée oscillant à l'intérieur de la bande de confiance.

12 Exemple : les chutes d'arbres en forêt

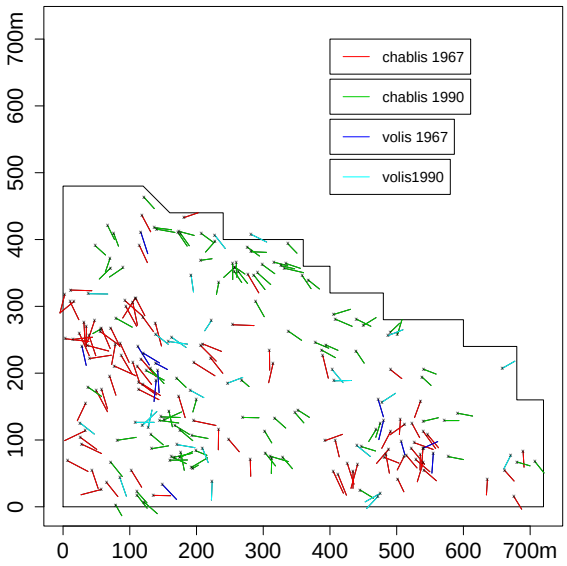
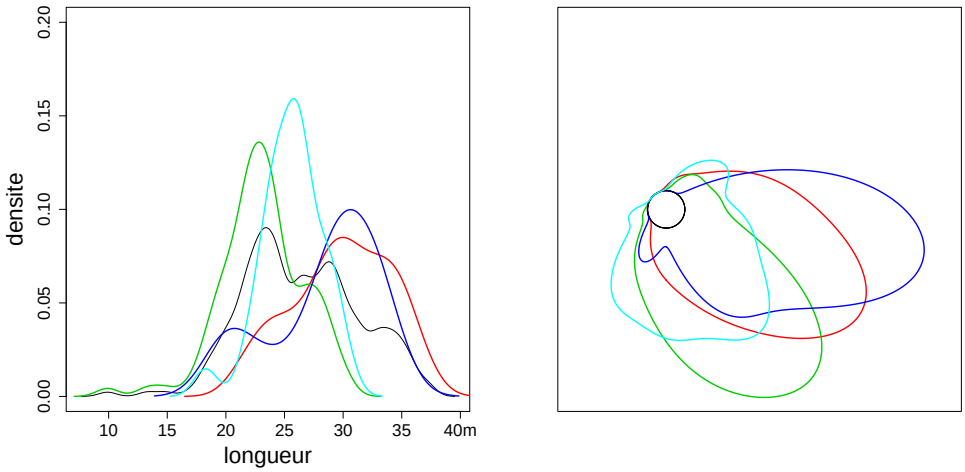


Figure 30: *Chute d'arbres en forêt de la Tillaie. Chaque point marqué en bout de fibre correspond à la position de la souche.*

La réserve de la Tillaie est une réserve naturelle en forêt de Fontainebleau, c'est à dire une zone dans laquelle on ne pratique aucune gestion forestière pour pouvoir étudier la dynamique naturelle (Faille *et al* 1984; Bénédeau 2003). Les chablis (arbres renversés par le vent) et les volis (arbres cassés par le vent) ont été positionnés sur cette parcelle de 720m $\times$ 480m à l'occasion de deux évènements venteux importants (Pontailler *et al* 1997), les années 1967 et 1990 (figure 30). Ces chutes se produisent préférentiellement en bordure de clairière, et l'un des problèmes de gestion forestière lors des opérations de

martelage est d'éviter la formation de telles trouées qui peuvent être le point de départ de chutes en domino.



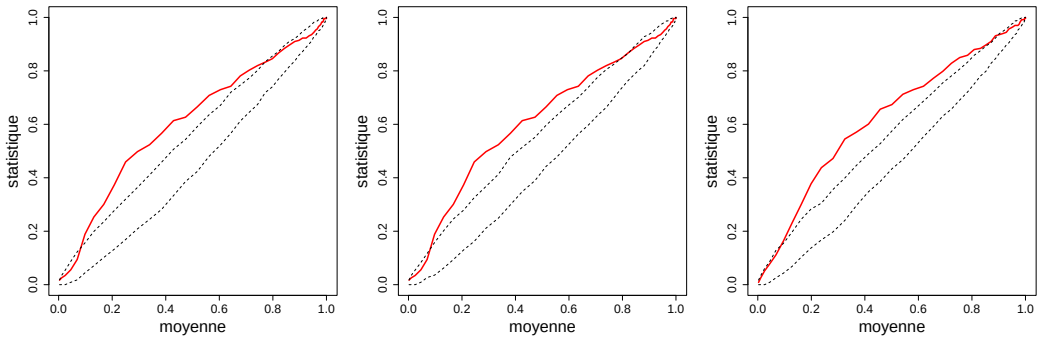
(a) (b)

Figure 31: *Distributions de longueur et d'angle des volis et chablis des années 1967 et 1990. Codes de couleur donnés en figure 30. La courbe noire de la figure (a) correspond à la distribution de longueur toutes années confondues. En figure (b), la distance du cercle central à la courbe le long d'un rayon est proportionnelle à la densité estimée.*

12.1 Exploration préliminaire

Les différents évènements sont inégalement répartis entre volis et chablis, avec 95 chablis en 1967, 98 chablis en 1990, 13 volis en 1967 et 27 volis en 1990. Les estimations non-paramétriques de densité (Bosc et Lecoutre, 1987) faites sur volis sont donc à interpréter avec prudence vu le nombre d'individus présents. La distribution de longueur des troncs, donnée en figure 31(a), illustre la différence entre années. On assiste à une chute préférentielle des plus grands arbres en 1967, arbres offrant plus de prise au vent, avec un pic de distribution de hauteurs aux alentours de 31m. Ceux chutant en 1990 ont une hauteur plus faible, le pic se situant autour de 25m. La distribution d'angles est donnée en figure 31(b). Si on retrouve globalement

une orientation préférentielle des troncs ($-\pi/4$), chaque évènement semble avoir ses propres caractéristiques distributionnelles. On peut remarquer que les distributions de longueurs et d'angles des volis (en bleu) sont chaque année assez proches de la distribution de longueur et d'angle des chablis correspondant. Le graphique des longueurs des troncs en fonction des angles (non donné ici), ne montre aucune orientation préférentielle des arbres en fonction de leur taille.



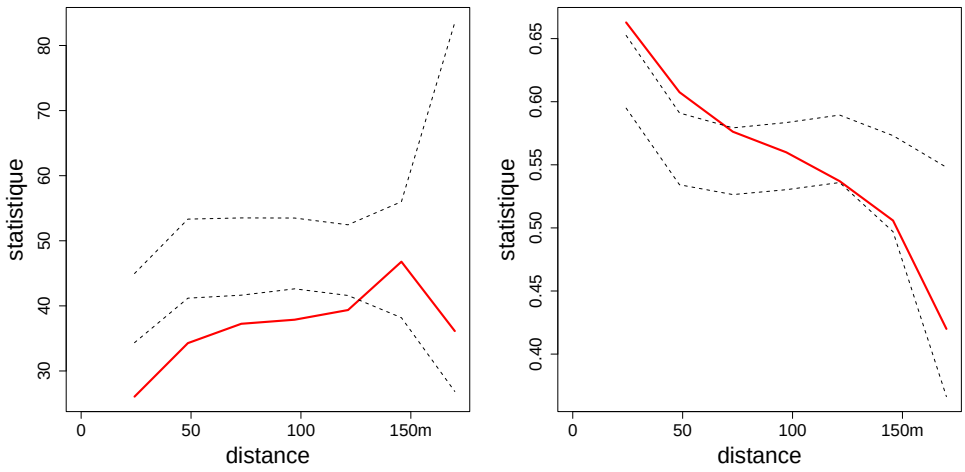
(a) (b) (c)

Figure 32: *(a et b)Tests d'hypothèse booléenne. Distance au plus proche voisin entre fibres. distance maximale 100m. Abscisse : distribution moyenne sous l'indépendance, ordonnée : distribution observée (en gras) et bande de confiance à 95% sous l'indépendance. (a) : hypothèse booléenne isotrope, (b) hypothèse booléenne anisotrope. (c) Tests d'hypothèse CSR sur le semis de points de souches. Distribution des distances au plus proche voisin. Abscisse: moyenne de la statistique sous l'indépendance, Ordonnée : statistique observée (en gras) et bandes de confiance sous l'indépendance.*

12.2 Tests d'hypothèse

Le test de l'hypothèse de distribution booléenne de l'ensemble du semis de fibres a été effectué en utilisant comme critère la distribution de distance au plus proche voisin entre fibres. La distance maximale explorée entre fibres est de 100m. L'hypothèse booléenne isotrope est rejetée (figure 32(a), test présenté en 11.1), de même que l'hypothèse booléenne anisotrope (figure 32(b) test présenté en 11.2).

Dans une deuxième étape, nous avons testé si les points d'origine, ici les positions des souches, sont distribués aléatoirement uniformément dans la zone ou non. L'ensemble des souches ayant conduit à chablis ou volis étant répertorié sur la parcelle, nous avons appliqué directement un test d'hypothèse CSR sur le semis de points d'origine, en utilisant comme critère la distance au plus proche voisin entre points d'origine (test présenté au paragraphe 5). Le résultat obtenu est donné en figure 32(c). La courbe observée est nettement en dessus de la bande de confiance. On rejette l'hypothèse d'indépendance totale, ce qui confirme l'agrégation déjà observée par Goreaud (2000) pour les arbres morts comme pour les arbres vivants.

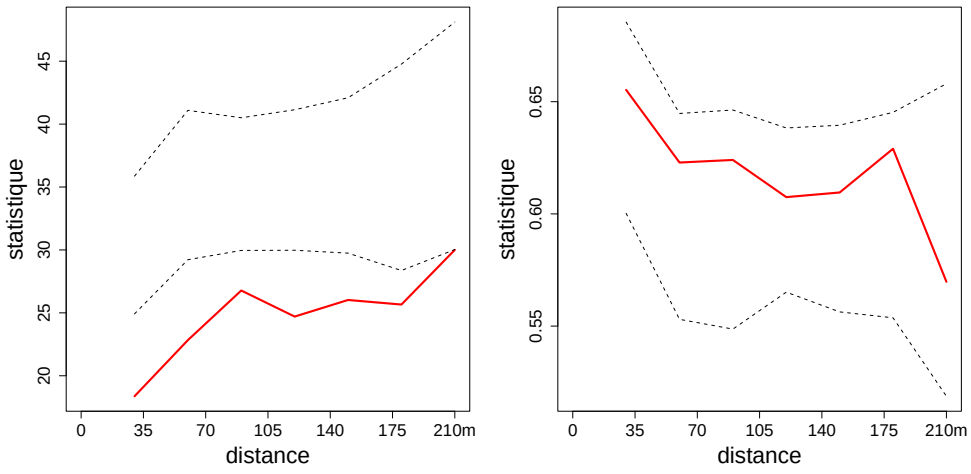


(a) (b)

Figure 33: *variogramme des longueurs (a) et moyenne du cosinus carré de l'angle (b) en fonction de la distance entre fibres et intervalles de confiance sous hypothèse de répartition aléatoire des troncs sur souches.*

Le semis de points d'origine n'est pas poissonien (Fig 32), et les angles et les longueurs des troncs ne sont pas distribués aléatoirement uniformément (dans un segment de bornes données pour les longueurs, entre 0 et 2π pour les angles) comme le montre clairement la figure 31. On souhaite alors savoir si, conditionnellement à la position des points d'origine, les caractéristiques (taille, angle de chute) des troncs sont indépendantes entre elles. Pour cela, on se propose de regarder des statistiques dépendant des moments d'ordre 2

(le variogramme par exemple) du semis de fibres (cf paragraphe 11.5). Dans la mesure où on dispose de l'ensemble des caractéristiques des troncs dont la souche est dans la zone étudiée, plutôt que d'utiliser l'une des procédures proposées en 11.5, on va plutôt permuter les fibres au hasard sur les souches de façon à utiliser le maximum de l'information disponible. Ceci nous évitera d'avoir à gérer les intersections avec le bord de la fenêtre. On se retrouve alors dans le cadre d'analyse d'un processus ponctuel marqué, et on prendra comme critère le variogramme pour la longueur (pour un ensemble de classes de distances, on sélectionne les couples de fibres dont la distance est dans une classe donnée, on calcule la moyenne du carré de la différence de longueurs de tels couples) et le cosinus carré pour les angles (pour les mêmes classes de distances, les mêmes couples, on calcule la moyenne du carré du cosinus de l'angle entre les deux fibres de chaque couple).

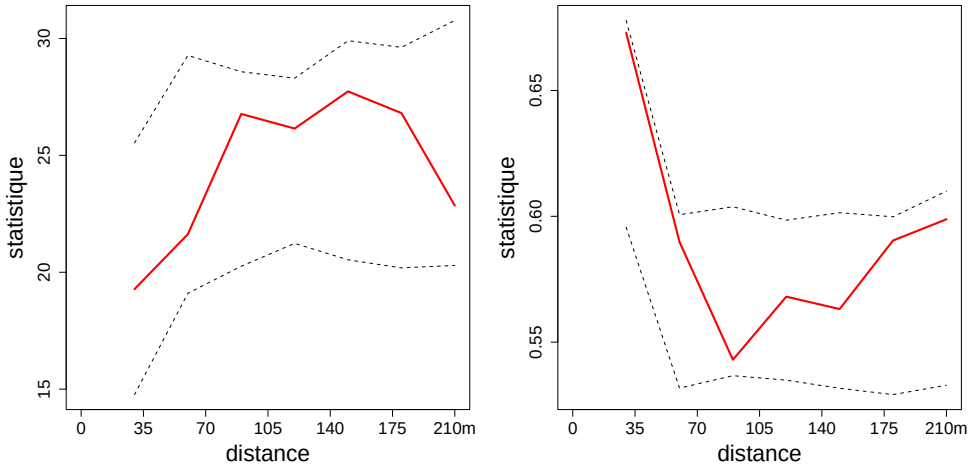


(a) (b)

Figure 34: *variogramme des longueurs (a) et moyenne du cosinus carré de l'angle (b) en fonction de la distance entre fibres et intervalles de confiance sous hypothèse de répartition aléatoire des troncs sur souches restreint aux chutes de 1967.*

Le résultat du test est donné en figure 33. On note un écart significatif à l'indépendance entre longueurs jusqu'à une distance de 125m, alors que l'indépendance entre les angles est rejetée jusqu'à 50m seulement. La structure de dépendance de 120m sur les longueurs, correspondant à une moyenne

du carré des différences entre couples de segments moins grande que sous le hasard, peut être due à une structuration des hauteurs de la parcelle. Si on revient à la figure 31(a), on a vu que les longueurs les plus grandes sont dues aux chablis et volis de 1967, et, si on revient à la figure 30, ces derniers sont préférentiellement répartis sur deux taches. La dépendance spatiale des longueurs peut alors correspondre à cette structure en tache des arbres que traduit les différences 1967/1990.
La structure de dépendance à moins de 50m des angles des segments correspond à un cosinus carré moyen plus grand sur le terrain que sous indépendance, soit une différence d'angle plus petite que sous le hasard. Elle peut traduire l'existence de chutes successives d'arbres, soit sous l'effet d'un même coup de vent, soit que le premier arbre pousse le deuxième... L'écart entre les deux échelles traduit alors l'existence de deux phénomènes : la structuration de la parcelle pour la longueur, les caractéristiques des chutes pour les angles.



(a) (b)

Figure 35: *variogramme des longueurs (a) et moyenne du cosinus carré de l'angle (b) en fonction de la distance entre fibres et intervalles de confiance sous hypothèse de répartition aléatoire des troncs sur souches restreint aux chutes de 1990.*

Comme les deux années sont relativement différentes, tant en répartition spatiale de dégâts (figure 30) qu'en caractéristiques des arbres concernés (fig-

ure 31), nous proposons de regarder si le rejet des hypothèses d'indépendance entre segments conditionnellement à la position des points d'origine est également rejetée par année. Nous avons donc appliqué la même procédure aux sous-semis de fibres de 1967 (figure 34) et 1990 (figure 35). Dans les deux cas (1967 et 1990), on note que l'hypothèse d'indépendance angulaire entre segments n'est pas rejetée. L'hypothèse d'indépendance entre longueurs n'est pas rejetée en 1990, mais l'est en 1967 et pour des distances plus grandes.

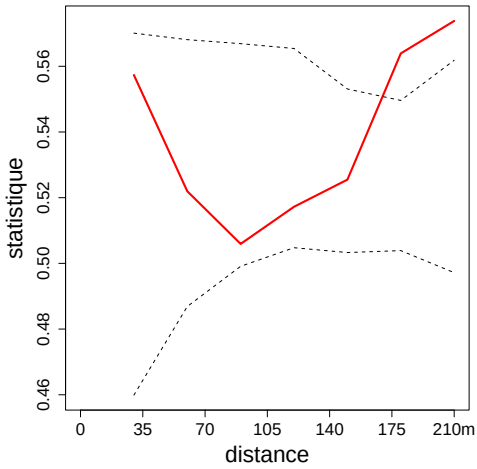


Figure 36: *moyenne du cosinus carré de l'angle entre un tronc tombé en 1967 et un tronc tombé en 1990 en fonction de la distance entre fibres et intervalles de confiance sous hypothèse de répartition aléatoire des troncs sur souches conditionnellement à l'année de chute.*

On observe ainsi une structuration spatiale des longueurs des arbres tombés en 1967, arbres globalement plus grands qu'en 1990, relativement forte, alors qu'aucune structure spatiale n'apparaît dans ceux de 1990. La courbe du variogramme des longueurs en 1967 (Figure 34a) sort de la bande individuelle de confiance jusqu'à une distance de 210m correspondant au diamètre des deux zones de chute observées en Figure 37. Cet effet peut alors être du à un effet d'hétérogénéité de la parcelle, dont on sait par ailleurs qu'elle présente une différence de sol et donc de fertilité entre la partie gauche et la partie droite. De ce fait le test global (figure 33(a)) réalise un moyen terme entre ces deux extrêmes.
La dépendance angulaire observée en figure 33(b) n'est plus significative

quand on regarde les deux années séparément. Pour expliquer que cet effet apparaisse sur l'analyse globale, mais pas sur les analyses séparées, on peut envisager soit simplement l'existence d'une différence de distribution entre les deux années (déjà notée en figure 31(b)), soit l'existence en plus d'un effet local des trouées de 1967 sur la direction de chute des arbres de 1990. Pour le vérifier, nous avons testé l'indépendance entre les directions de couples de fibres issues l'une du semis de fibres de 1967 et l'autre de celui de 1990, conditionnellement à une indépendance des orientations des fibres de chaque processus. On s'attend à ne pas avoir de test significatif si les deux processus sont indépendants l'un de l'autre, chacun avec ses propres caractéristiques distributionnelles, et à avoir un test significatif aux petites distances s'il existe un effet local des trouées.

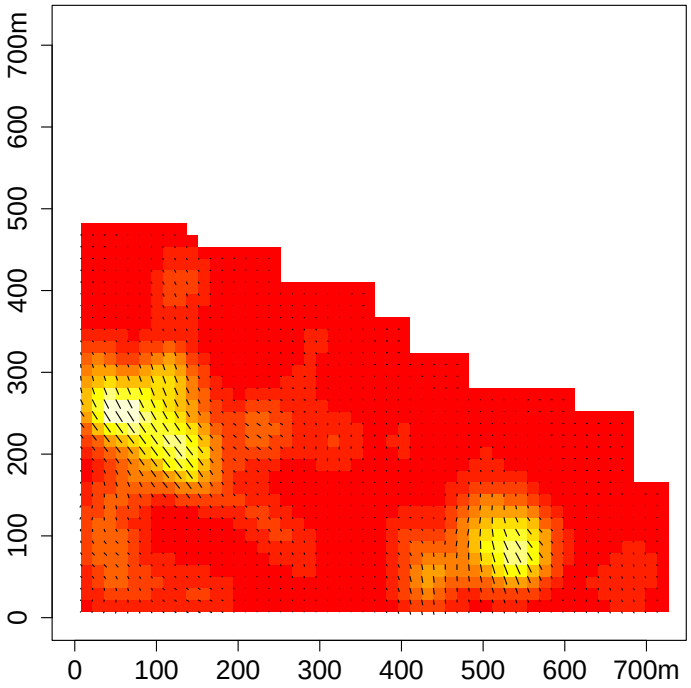


Figure 37: *densités locales de longueur et d'orientation de chablis et volis en 1967. Plus la zone est claire et plus la densité de longueur est forte. En chaque point, l'orientation du segment noir représente l'orientation moyenne locale de fibres en ce point, sa longueur la densité de longueur de fibres au même point.*

Le résultat du test est donné en figure 36. Le test est non significatif pour toutes les distances inférieures à 150m, conduisant à ne pas rejeter l'hypothèse d'une absence d'influence de la direction de chute des arbres de 1967 sur 1990. Au delà de 150m, il devient significatif, et peut correspondre à une structuration à longue distance, mais qu'il convient de prendre avec prudence vu la taille de la parcelle.

12.3 Cartographie

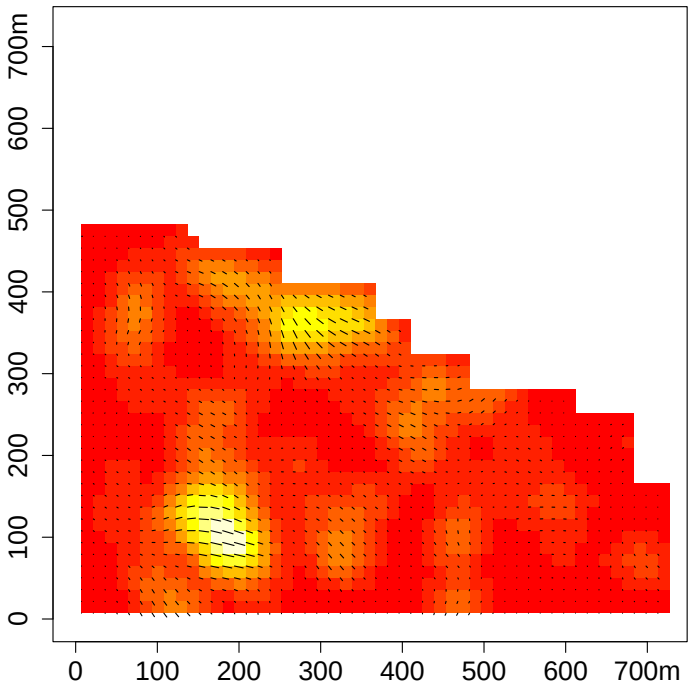


Figure 38: *densités locales de longueur et d'orientation de chablis et volis en 1990. Plus la zone est claire et plus la densité de longueur est forte. En chaque point, l'orientation du segment noir représente l'orientation moyenne locale de fibre en ce point, sa longueur la densité de longueur de fibre au même point.*

Pour obtenir une vue globale de l'orientation et de la densité de longueur d'arbres abattus, nous avons calculé la densité locale de longueur de fibres par une méthode du noyau (Silverman 1986). Partant d'un noyau g, c'est à dire d'une fonction positive vérifiant $\int_{x \in I\!R^2} g(x)dx = 1$, on estime en chaque point M de l'espace la densité de longueur de fibres en M par

$$\hat{\lambda}(M) = \int_{x \in \Phi} g(x - M)dx^3$$

qui va estimer l'intensité en M (c'est à dire la longueur moyenne par unité de surface autour du point M) du processus de fibres défini par $\lambda(M) = \lim_{B \to M} E(\Phi(B))/\nu(B)$. Comme pour tous ces estimateurs, la largeur de bande (que l'on peut visualiser comme l'écart-type de la fonction g vue comme une densité si on prend g symétrique autour de 0) va influencer les propriétés de l'estimateur. Trop petite, l'estimateur va avoir une grande variance. Trop grande, l'estimateur va être biaisé. Le choix d'une largeur de bande "optimale" est généralement faite par minimisation du l'erreur quadratique moyenne définie comme la moyenne sur l'ensemble des points de la zone de la somme de la variance de l'estimateur et du carré de son biais.

Les images représentées figures 37 et 38 donnent ces intensités pour l'union des chablis et volis de 1967 et 1990. Nous avons choisi un noyau gaussien d'écart-type 25m dans les deux cas. On notera simplement qu'il y a exclusion mutuelle des zones à fortes densités de longueur tombées entre 1967 et 1990, représentées en jaune et en blanc sur l'image, plutôt qu'une extension en 1990 des zones de chablis et volis de 1967.

Nous avons représenté sur ces images le champ de vecteurs $\hat{V}$ qui estime l'orientation moyenne locale autour de chaque point M, calculée en prenant en compte la longueur des fibres [4]. Plus précisément, si on attache à chaque point x sur une fibre du processus sa tangente $\Phi'(x)$, on définit un processus

[3] En pratique, on peut visualiser $g(x)$ comme une surface 3D ayant la forme d'une petite cloche centrée sur 0. Pour estimer la densité locale d'un semis de points Φ, on place une telle petite cloche au dessus de chaque point du processus. Pour chaque point M du plan, on somme les hauteurs de toutes les cloches pour obtenir $\hat{\lambda}(M)$. Plus la densité de points du processus est forte autour de M et plus il y aura de petites cloches proches de M et plus $\hat{\lambda}(M)$ sera important. Pour un semis de fibres, on discrétise les fibres de façon à revenir ň processus ponctuel.

[4] De même que précédemment on aurait pu calculer le nombre de fibres par unité de surface, on pourrait aussi s'intéresser ici à l'orientation moyenne d'une fibre par unité de surface, en donnant à chaque fibre le même poids.

de fibres marqué par sa tangente $(\Phi, \Phi') = \cup_{x\in\Phi}(x, \Phi'(x))$ et $\hat{V}(M)$ estime

$$V(M) = \lim_{B\to M} E(\Phi'(B))/\nu(B)$$

En pratique, il est calculé comme précédemment par noyau, la projection de $\hat{V}(M)$ dans la direction θ étant donnée par :

$$\hat{V}_\theta(M) = \int_{x\in\Phi} v_\theta(x) g(x - M) dx$$

où $v_\theta(x)$ est la projection dans la direction θ de la tangente à la fibre passant par x.

Nous avons pris le même noyau pour estimer ce champ de vecteur les deux années. Il est représenté par superposition sur la figure 37. En dehors des zones à forte densité de fibres, la norme de $\hat{V}_\theta(M)$ est petite puisque $\| v_\theta(x) \| = 1$. L'orientation n'est de plus pas la même dans toutes les zones, ce qui peut correspondre à des effets locaux du vent. La direction similaire moyenne des arbres tombés dans la tache située en (100,200) en 1967 et de ceux tombés dans la tache située en (300-350) en 1990 pourrait être la cause du test significatif au delà de 150m en figure 36.

13 Conclusion

Les premières questions que l'on se pose devant un jeu de données formé d'un semis de points ou de fibres sont généralement des questions de tests visant à dégrossir le sujet pour lequel ce jeu de données a été acquis, le souci d'estimation ou de quantification des effets significatifs relevant plutôt d'une deuxième étape de modélisation, qu'elle soit statistique ou non. Les tests de permutation, de par leur nature non-paramétrique et leur facilité d'utilisation liée à leur coté intuitif, sont un moyen de répondre rapidement et efficacement à ces premières questions.

Si quelques procédures standardisées existent, en particulier dans le cas des processus ponctuels, leur intérêt réside aussi dans leur grande souplesse d'application. Dès qu'une question peut se formuler en terme d'égalité et d'indépendance, un tel type de test peut être envisagé. De ce point de vue, l'information apportée par des covariables est fondamentale, car elle va enrichir les possibilités de permutations.

- Concernant le premier point, à savoir la souplesse d'utilisation, nous sommes restés ici dans le cadre classique de l'exploration par test exact d'un processus de fibres pour lequel peu de résultats préliminaires existent. Dans la pratique, les tests de permutations sont très souples, et les procédures de test présentées doivent être considérées comme des exemples à partir desquels on peut bâtir des procédures de test adaptées aux cas rencontrés. Ainsi par exemple, si on connaît la distribution des fibres, on peut l'intégrer dans les tests et obtenir ainsi des tests plus puissants que ceux présentés ici, qui travaillent conditionnellement aux fibres observées.

- En ce qui concerne la formulation des questions, on notera que les notions d'indépendance comme celles d'agrégation, si elles semblent intuitives en première approche, méritent souvent d'être redéfinies très précisément dès que le processus est un peu complexe. Ainsi, dans le cas de processus ponctuels marqués, nous avons rencontré deux types de tests correspondant classiquement à deux définitions de la notion d'indépendance. De même, nous avons vu pour les processus de fibres que l'on peut avoir deux conclusions divergentes quand à l'agrégation d'un même processus de fibres, selon que l'on regarde l'entité fibre ou sa longueur.

Enfin, si le jeu de données observé est important, des procédures de type bootstrap peuvent être envisagées, ce qui permettra d'élargir les possibilités de test, au prix d'un possible biais des tests, d'autant plus minime qu'il se fera à partir d'un grand échantillonnage.

Références

Bénédeau M. 2003. Evolution des différentes states dans le peuplement ligneux de la réserve biologique de la Tillaie en forêt de Fontainebleau entre 1968 et 2000. *Rev. For. Fr.*, LV, 323-332.

Bosq D. & Lecoutre J.P. 1987. Théorie de l'estimation fonctionnelle. Economica, Paris, 342 p.

Burghardt AF. 1959. The location of river towns in the central lowland of the United States. *Annals of the Association of American Geographers.*, 49, 305-323.

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

Efron B. & Tibshirani R.T. 1993. *An introduction to the bootstrap.* Chapman and Hall, New York, 436 p.

Elliot P., Wakefield J., Best N. & Briggs D. 2001. *Spatial Epidemiology. Methods and Applications.* Oxford University Press, Oxford, 475p.

Erschbamer B., Grabherr G. & Reisigl H. 1983. Spatial pattern in dry grassland communities of the Central Alps and its ecophysiological significance. *Vegetatio.* 54, 143-154.

Goreand F. 2000. Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude et la modélisation des peuplements complexes. ENGREF (Thèse Docteur) 361p.

Goreand F. & Pélissier R. 2003. Avoiding misinterpretation of biotic interactions with the intertype K_{12}-function: population independance vs random labelling hypotheses. Journal of Vegetation Science. 14, 681-692.

Hope A.C.A. 1968. A simplified Monte-Carlo significance test procedure. *Jl. R. statist. Soc.* B, 30, 582-598.

Mahfoud I. 2003. Dynamique des écosystèmes forestiers méditerranéens : structuration spatiale de semis de sapin concolor (Abies concolor) issus

de régénération naturelle dans la montagne de Lure. Mémoire (DEA), option Biosciences de l'environnement, Chimie et Santé, Université d'Aix-Marseille 3, Faculté des Sciences et Techniques de St Jérôme, Marseille & INRA Centre d'Avignon, Unité de Recherches Forestières Méditerranéennes, 34p.

Martinez V.J. & Saar E. 2002. *Statistics of the galaxy distribution.* Chapmann & Hall, London, 450p.

Matheron G. 1975. *Random sets and integral geometry.* John Wiley & Sons; Chichester. 261p.

Peyrard N., Calonnec A., Bonnot F. & Chadœuf J. 2004. Explorer un jeu de données sur grille par tests de permutation. A paraître dans Revue de Statistique Appliquée.

Pontailler J.Y., Faille A. & Lemée G. 1998. Storms drive successional dynamics in natural forests: a case study in Fontainebleau forest (France). *Forest Ecology and Management.* 1-15.

Ripley B.D. 1976. The second-order analysis of stationary point processes. *J. Appl. Prob.*, 13, 255-266.

Stoyan D., Kendall W.S. & Mecke J. 1995. *Stochastic geometry and its applications.* John Wiley & Sons; Chichester, 436p.

Silverman, B.W. 1986 *Density estimation for statistics and data analysis* Chapman & Hall; London, 175 p.

Upton G.J.G. & Fingleton, B. 1988. *Spatial data analysis by example. Volume1 : Point pattern and quantitative data.* Wiley; Chichester , 409p.

Van Lieshout M.N.M. & Baddeley A.J. 1997. A non-parametric measure of spatial interaction in point patterns. *Statist. Neerland.*, 50,344-361.

Zähle M. 1982. Random processes of Hausdorff rectifiable closed sets. *Math. Nach.*, 108, 49-72.

Zähle M. 1982. Random set processes in homogeneous Riemannian spaces. *Math. Nach.*, 110, 179-193.

List of Figures

Juin 2005

Compte rendu 220517001

INRA

J. CHADOEUF, F. GOREAUD, J.Y. PONTAILLER et S. SOUBEYRAND

Contribution à l'élaboration de méthodes de statistique spatiale dans le traitement de données agricoles permettant de prendre en compte le contexte géographique et d'améliorer la précision des références fournies aux organismes de développement agricole

Annexe A1.3 : Tests non paramétriques d'indépendance de la répartition d'objets complètement observables distribués dans le plan

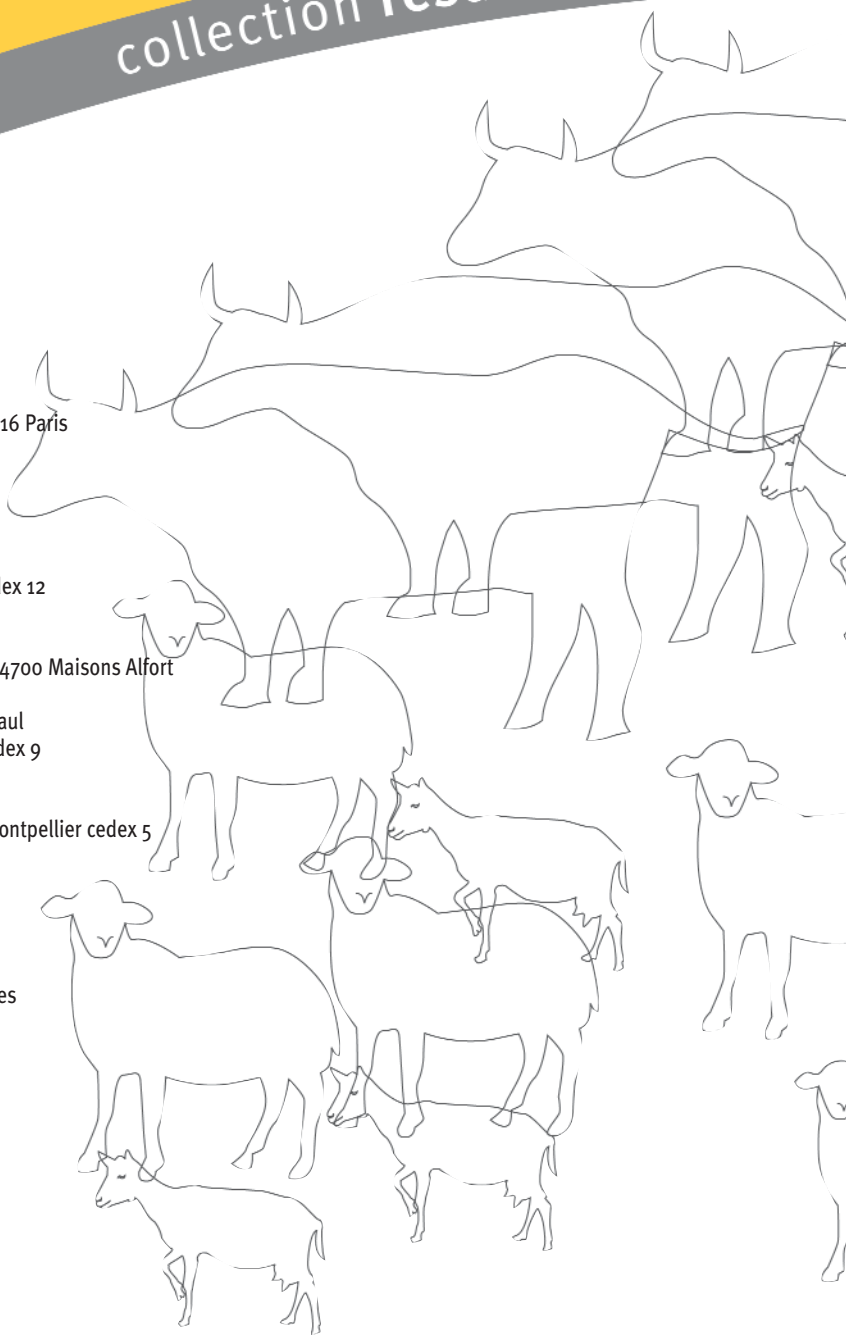
Une action de recherche inter-instituts a été conduite de 2002 à 2004 en collaboration avec la recherche et l'enseignement supérieur afin de dresser un inventaire des méthodes statistiques regroupées sous le terme de 'statistiques spatiales' et de les caractériser par rapport aux thèmes de travail rencontrés en agriculture, selon les contextes (agronomiques, environnemental, zootechnique, épidémiologique, ...), les questions posées (description, prédiction, optimisation (plan d'échantillonnage) et les valorisations envisagées (production de cartes, caractérisation de clusters, modèles de prédiction, ...).

Ce travail mené par l'Institut de l'Elevage est le résultat d'une collaboration avec ARVALIS – Institut du végétal, l'ITP, l'ITAVI ainsi que l'Ecole Vétérinaire d'Alfort (Unité Epidémiologie et Analyse des Risques), l'INRA (Unité de Biométrie – Avignon), le laboratoire de Mathématiques Appliquées (Pôle d'Enseignement Supérieur et de Recherche Agronomique) de Rennes, l'IUT de Vannes, l'Université Montpellier II.

Quatre séminaires ont été organisés : *« Le Krigeage »*, *« Les processus de champs Markoviens / processus Markoviens »*, *« Les systèmes et les traitements d'informations géographiques »*, *« L'approche bayésienne en modélisation statistique spatiale »* qui ont abouti à la rédaction de 3 documents de synthèse : *« Représentation d'un phénomène aléatoire sur un réseau à partir d'une réalisation complète »*, *« Cartographier une variable continue à partir d'un échantillon : l'approche géostatistique »*, *« Tests non paramétriques d'indépendance de la répartition d'objets complètement observables distribués dans le plan »*.

Les résultats, présentés lors d'un séminaire final qui s'est tenu à Paris les 17-18 mars 2005, ont permis de montrer l'intérêt et les limites d'utilisation des méthodes employées dans le contexte des applications étudiées. Une réflexion a également été initiée, concernant le plus long terme, pour la mise à disposition d'outils informatiques de traitement de données spatialisées et la mise en place de formations à destination des ingénieurs et techniciens des ICTA afin d'aider à promouvoir l'utilisation de ces méthodes.

collection résultats



ARVALIS - Institut du Végétal
3 rue Joseph et Marie Hackin - 75116 Paris

ITAVI
4 rue Bienfaisance - 75008 Paris

ITP
149 rue de Bercy - 75595 Paris cedex 12

ENV Alfort
7 avenue du Général de Gaulle - 94700 Maisons Alfort

INRA (Avignon) - Domaine Saint-Paul
Site Agroparc - 84914 Avignon cedex 9

Université de Montpellier II
Place Eugène Bataillon - 34095 Montpellier cedex 5

IUT Vannes
8 rue Montaigne - 56000 Vannes

Agrocampus Rennes
65 rue Saint-Brieuc - 35000 Rennes

Institut de l'Élevage
149, rue de Bercy
75595 Paris CEDEX 12
www.inst-elevage.asso.fr