REVUE DE STATISTIQUE APPLIQUÉE

N. PEYRARD
A. CALONNEC
F. BONNOT
J. CHADŒUF

Explorer un jeu de données sur grille par tests de permutation

Revue de statistique appliquée, tome 53, n° 1 (2005), p. 59-78
<http://www.numdam.org/item?id=RSA_2005__53_1_59_0>

Rev. Statistique Appliquée, 2005, **LIII** (1), 59-78

EXPLORER UN JEU DE DONNÉES SUR GRILLE PAR TESTS DE PERMUTATION

N. PEYRARD[1], A. CALONNEC[2], F. BONNOT[3], J. CHADŒUF[1]

[1] *INRA, Biométrie Domaine St Paul, Site Agroparc, 84914 Avignon Cedex 9*
[2] *UMR INRA-ENITAB, Santé végétale, 71 avenue Edouard Bourlaux, BP 81 33883 Villenave-d'Ornon Cedex*
[3] *Cirad-Cp, TA 80/03, 34398 Montpellier Cedex 5*

RÉSUMÉ

Les tests de permutation font référence à des méthodes pour l'exploration d'un jeu de données sans hypothèse sur leur distribution. Ces méthodes, simples à mettre en œuvre, permettent de mettre en évidence des caractéristiques sous-jacentes aux données, que l'on pourra vouloir traduire dans un modèle dans une seconde étape d'analyse plus fine. Ce type de tests est largement utilisé dans le cas d'observations indépendantes. Dans le cadre de données observées sur grille, la mise en évidence de la structure spatiale des données conduit à un ensemble de tests spécifiques. Nous présentons ainsi dans cet article comment aborder la vérification d'un certain nombre d'hypothèses classiques au cas spatial dans le cadre des tests de permutation.

Mots-clés : *tests de permutation, données spatiales, recherche de structures.*

ABSTRACT

Permutation tests are methods for data exploration which do not require assumptions on their distribution. These methods are a simple way to extract data characteristics, that can be integrated in a model in a second finer stage of the analysis. They have been widely used in the context of independent observations. When considering data observed on a grid, analysing the spatial structure of the data leads to specific tests. We present in this article how to deal with the investigation of classical hypothesis in spatial analysis through the use of permutation tests.

Keywords : *permutation tests, spatial data, structures investigation.*

1. Introduction

Les tests de permutation sont des méthodes faciles à mettre en œuvre. Ils compensent l'absence de résultats analytiques sur les seuils des tests par une souplesse d'utilisation associée au calcul numérique de seuils exacts, ou aussi exacts qu'on le souhaite dès que l'on prend un nombre de permutations suffisant, sur toute statistique désirée (Manly 1997, Mielke & Berry 2001).

En effet, le principe des tests de permutation est de travailler à partir des seules données observées, de façon à éviter de formuler des hypothèses sur la distribution de la variable mesurée. Ayant une statistique de test, on la calcule sur les données observées. On peut ensuite redistribuer les données observées au hasard (les permuter) et recalculer la statistique pour chacune des permutations. Le point clé de la méthode est que sous l'hypothèse nulle toute permutation des données observées est équiprobable dans l'ensemble des permutations correspondant à l'hypothèse testée. On confronte ensuite la statistique calculée sur les données observées à la distribution empirique des statistiques calculées sur les permutations.

Les tests de permutation ne permettent pas de répondre finement aux questions que l'on se pose sur un jeu de données, comme le ferait un modèle particulier. Ils vont plutôt permettre de trier les grandes hypothèses que l'on souhaite vérifier avant de bâtir un modèle spécifique. Ainsi que mentionné par Efron & Tibshirani (1993), leur cadre d'application est plus restreint que le bootstrap (puisqu'il n'y a pas toujours quelque chose à permuter pour vérifier une hypothèse), mais ils donnent des réponses très satisfaisantes dans ce cadre, tout en évitant d'avoir à faire des hypothèses paramétriques, ou de devoir se placer dans le cadre asymptotique du bootstrap.

Ainsi par exemple, supposons que l'on dispose d'un plan d'expérience à deux facteurs A et B. Dans A ont été mesurées les valeurs suivantes : (1,2,2,1,2,1,3). Dans B ont été mesurées : (1,3,3,2,2,1,3). Supposons que toutes les mesures sont indépendantes. On souhaite tester l'hypothèse nulle «les deux échantillons ont la même moyenne». On prend comme statistique la différence des moyennes dans chaque facteur, celle classiquement utilisée en analyse de variance. On obtient alors par exemple sur trois permutations :

	Mesures dans A	Mesures dans B	Statistique
Échantillon observé	(1,2,2,1,2,1,3)	(1,3,3,2,2,1,3)	$-0,4285714$
Permutation 1	(1,3,3,2,2,2,1)	(2,1,1,2,3,3,1)	0,1428571
Permutation 2	(1,2,3,1,2,1,3)	(1,3,2,1,3,2,2)	$-0,1428571$
Permutation 3	(2,1,1,2,1,2,1)	(3,2,3,3,1,3,2)	-1

On considérera que les deux échantillons sont de moyennes différentes si la différence des moyennes calculée sur l'échantillon observé est soit très positive, soit très négative par rapport aux valeurs calculées sur les permutations. On fera donc un test bilatéral.

Sur 100 permutations, on obtient l'histogramme de la figure 1. Il y a 5 valeurs en dessous de la valeur observée (de $-0,4285714$). On est donc dans l'intervalle de confiance à 95 % et on conclut que les deux groupes ne sont pas significativement différents.

Ce type de test est largement utilisé dans le cas d'échantillons où les observations sont indépendantes. Le cadre spatial, dans la mesure où il apporte une dimension supplémentaire, permet de développer tout un ensemble de tests spécifiques. Ce sont ces derniers que nous nous proposons de présenter ici, sans souci d'exhaustivité.

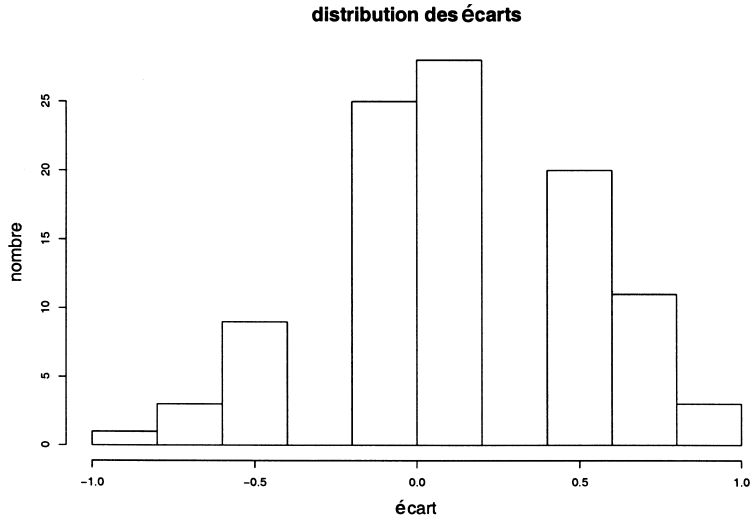

FIGURE 1
Exemple du plan d'expérience à deux facteurs, histogramme de la statistique, calculé sur 100 permutations

2. Cadre de travail

On considèrera par la suite une grille régulière de I lignes et J colonnes, pas obligatoirement à maille carrée, qui peut par exemple représenter un verger. Les données sont recueillies aux nœuds de la grille. On note i l'indice de ligne, j l'indice de colonne, $X_{(i,j)}$ la valeur observée au point (i, j).

On supposera que les valeurs $X_{(i,j)}$ sont quantitatives, ou qualitatives ordonnées, comme ce peut être le cas quand on note la sévérité de symptômes d'une maladie.

On trouvera en figure 2 un exemple d'un tel type de données. Une parcelle de 11 lignes espacées de 80 plants a été plantée en lavande vraie. Cette parcelle est atteinte de dépérissement, une maladie à mycoplasme transmise par une cicadelle. Deux plants consécutifs sur une ligne sont séparés de 20 cm. On a noté de 0 à 5 l'état sanitaire de chaque plant.

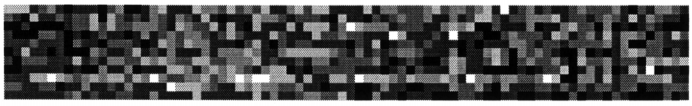

FIGURE 2
Répartition spatiale des notes de dépérissement de la lavande dans une parcelle du plateau de Sault

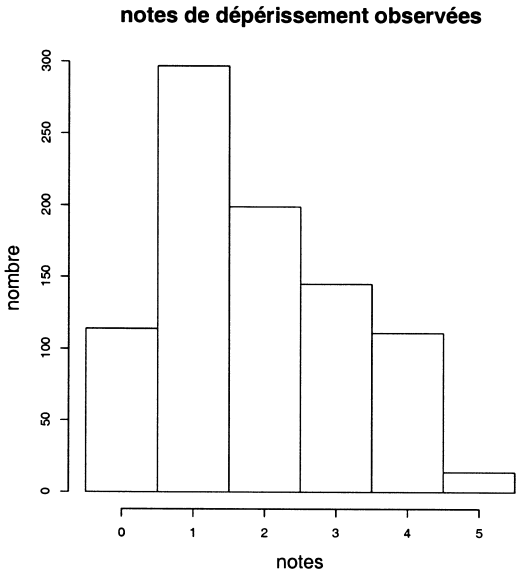

FIGURE 3
Nombre de plants ayant une note de dépérissement donnée dans l'exemple de la lavande

L'histogramme de la figure 3 résume l'état sanitaire de la parcelle, avec peu de plants sains (note 0, pixels noirs dans l'image) ou très fortement atteints (note 5, pixels blancs), beaucoup de plants avec peu de symptômes (note 1, pixels gris foncés dans l'image).

3. Analyse d'une seule image : tests d'indépendance spatiale

Dans ce chapitre, on dispose d'un jeu de données tel que celui présenté plus haut. On suppose que la distance entre lignes est plus grande que la distance entre points le long de la ligne. Dans ce cadre, on imagine que, si propagation il y a, elle se manifestera d'abord le long des lignes.

3.1. Test d'indépendance totale

La première question que l'on se pose face à une telle image est de savoir si il y a ou non dépendance spatiale. L'hypothèse testée (dite hypothèse i.i.d.) est :

« la probabilité d'obtenir une valeur donnée en un point donné est la même en tout point et les valeurs obtenues sur un ensemble de points sont indépendantes ».

Si ces hypothèses sont vraies, les images obtenues par permutation des valeurs initiales sont équiprobables. On peut donc prendre un critère comme le variogramme (Wackernagel 1995, Cressie 1993), le long de la ligne si on privilégie cette direction

de transmission (soit l'hypothèse alternative : « les observations le long d'un ligne ne sont pas indépendantes »), et regarder empiriquement sa distribution par permutation.

Le variogramme au pas d le long de la ligne s'écrit :

$$C(d) = \frac{1}{I(J-d)} \sum_{1 \leqslant i \leqslant I} \sum_{1 \leqslant j \leqslant J-d} (X_{(i,j)} - X_{(i,j+d)})^2$$

Si ϕ est une permutation aléatoire de l'ensemble des indices, on obtient alors une nouvelle image $X^{(\phi)}$ telle que $X^{(\phi)}_{(i,j)} = X_{\phi^{-1}(i,j)}$, à laquelle on associe le variogramme :

$$C_\phi(d) = \frac{1}{I(J-d)} \sum_{1 \leqslant i \leqslant I} \sum_{1 \leqslant j \leqslant J-d} (X^{(\phi)}_{(i,j)} - X^{(\phi)}_{(i,j+d)})^2$$

Le test de permutation peut alors être fait pour toute distance d. On peut aussi de façon équivalente construire une bande de confiance individuelle pour chaque distance d en prenant les quantiles correspondant au niveau souhaité.

Ce test a été appliqué sur les données de dépérissement de la lavande avec un seuil $\alpha = 0,05$ et $k = 200$ permutations. Les résultats sont reportés sur la figure 4 : la courbe en trait plein correspond aux valeurs observées et les courbes en pointillés correspondent aux limites de confiance à 2,5 % et 97,5 % (Ce codage restera le même pour les figures suivantes). On observe un très fort écart à l'indépendance à $d = 20$ cm, écart qui se réduit progressivement, mais reste significatif jusqu'à 3 m.

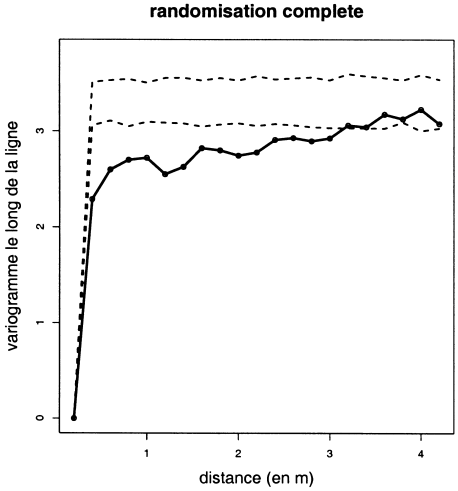

FIGURE 4

Test d'indépendance totale par permutation sur l'exemple de la lavande. La courbe en trait plein correspond aux valeurs observées et les courbes en pointillés correspondent aux limites de confiance à 2,5 % et 97,5 %

Si l'on privilégie l'hypothèse alternative «les lignes n'ont pas la même moyenne», on peut faire le même test en prenant comme critère la moyenne par ligne :

$$C(i) = \frac{1}{J} \sum_{1 \leqslant j \leqslant J} X_{(i,j)}$$

3.2. Test d'indépendance entre lignes

Si l'indépendance totale est rejetée, on peut se poser la question de la dépendance entre lignes. Cette question se pose par exemple dans le cas de maladies à faible distance de propagation, la distance interligne elle même, le travail du sol ou le microclimat entre lignes pouvant constituer un frein à la propagation. Il importe alors de tester cette hypothèse en conservant la structure dans les lignes. Le principe est alors de restreindre l'espace des permutations, en ne tirant au hasard que parmi celles qui laissent la structure de chaque ligne inchangée. Trois solutions sont alors possibles :

- Permuter globalement les lignes entre elles. On considère une permutation ϕ de l'ensemble$[1, I]$ et $X^{(\phi)}_{(i,j)} = X_{(\phi^{-1}(i),j)}$. Le problème posé est celui du faible nombre de permutations disponibles si I est petit. On dispose en effet de $I!$ permutations pour un ensemble de taille I. Un minimum d'une dizaine de lignes est souhaitable.
- Refermer chaque ligne sur elle même (le point (i, J) est alors suivi du point $(i, 1)$), et prendre une rotation de pas aléatoire et indépendant d'une ligne à l'autre. On note $p(i)$ le pas pour la ligne i. Alors, $X^{(\phi)}_{(i,j)} = X_{(i,j-p(i))}$. Cette procédure marchera bien dès que le nombre J d'individus par ligne est assez grand pour négliger l'effet de bord.
- Se restreindre aux rotations de pas compris dans $[-l, l]$ pour éviter l'effet de bord (si les lignes sont assez grandes). On restreint alors le calcul de l'indice dans la zone centrale (*i.e.* la ligne moins les l plants en début et les l plants en fin). L'efficacité sera alors un compromis mesuré par l. S'il est trop petit, peu de permutations seront disponibles et le test sera peu puissant. S'il est trop grand, la zone de calcul de l'indice sera petite et le test sera encore peu puissant.

Le critère à partir duquel le test sera calculé doit mesurer la cohérence entre lignes. On peut par exemple utiliser le variogramme entre lignes :

$$C(d) = \frac{1}{J(I-d)} \sum_{1 \leqslant j \leqslant J} \sum_{1 \leqslant i \leqslant I-d} (X_{(i,j)} - X_{(i+d,j)})^2$$

La procédure est ensuite la même que dans le paragraphe 3.1. On notera que chacun des deux grands types de procédure (rotation aléatoire dans la ligne contre permutation aléatoire des positions des lignes) répond à un type d'hypothèse alternative différent. Ainsi par exemple, si une non-stationnarité apparaît dans le

sens des lignes, le test de permutation des positions des lignes peut ne rien voir (on retrouvera toujours face à face des points statistiquement de même valeur). Il faudra donc préférer le test par rotation qui va casser cet effet (voir illustration sur figure 5). Inversement, si la non-stationnarité apparaît dans le sens des colonnes, il faut préférer le test par permutation des lignes.

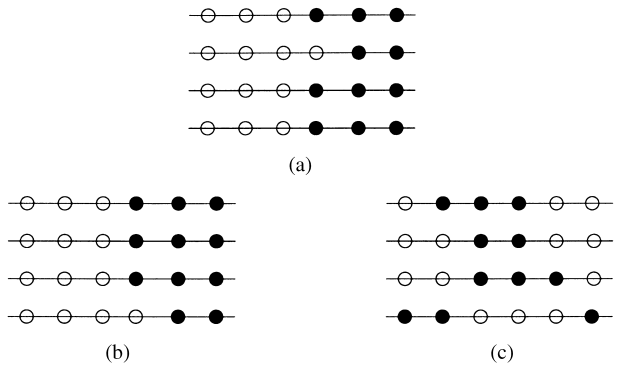

FIGURE 5
Tests d'indépendance entre lignes : (a) image observée, (b) image après permutation des lignes entre elles, (c) image après rotations indépendantes dans les lignes

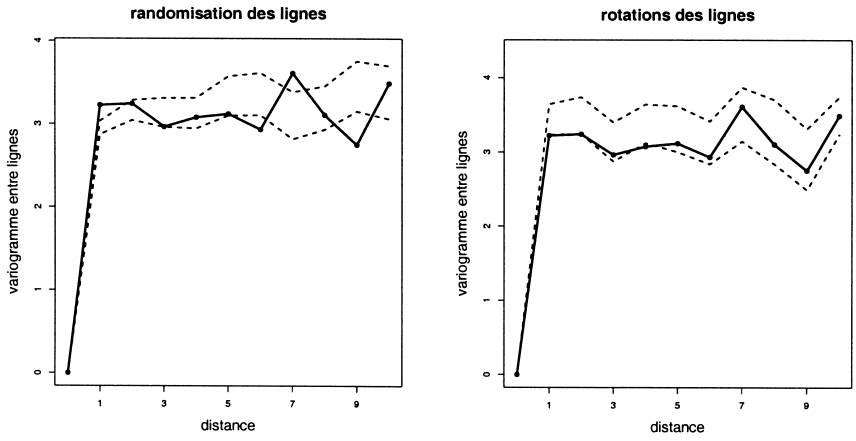

FIGURE 6
Tests d'indépendance entre lignes, par permutations des lignes entre elles à gauche, et par rotation dans les lignes à droite, pour l'exemple de la lavande

Permutations aléatoires des lignes et rotations aléatoires dans les lignes ont été appliquées dans l'exemple de la lavande (figure 6). Dans le premier cas, le test est rejeté au premier pas ($d = 1$) et limite au second ($d = 2$), les écarts entre notes observées étant plus grands que sous le hasard. Le second test conclut plutôt

l'inverse, même s'il est très limite. Cet effet *a priori* contradictoire s'explique par la présence d'une grosse tache plus claire au niveau du premier tiers de la parcelle (en passant vers la gauche), suivie d'une grosse tache sombre. Dans le cas de rotations aléatoires, on casse les concordances entre lignes consécutives. Dans le cas de permutations aléatoires des lignes, les taches sont globalement conservées, on analyse la concordance conditionnellement à cette tache.

3.3. Présence d'un état neutre

Dans le cas par exemple d'une attaque par une maladie, on va trouver des notes de sévérité d'attaque, et une note indiquant que la maladie n'est pas présente (ou n'a pas été détectée), supposée égale à 0 ici. C'est ce dernier état que l'on appelle l'état neutre.

Les deux tests précédents permettent de discuter si les notes observées sont corrélées ou non suivant une direction privilégiée et s'appliquent pour des observations de type présence/absence ou de type sévérite/absence. Dans le second cas, où les individus malades sont codés parmi plus d'une note, on peut pousser l'analyse plus loin et s'intéresser à savoir si, une fois les points attaqués identifiés, les notes de ceux-ci sont structurées entre elles ou non. Il s'agira de déterminer par exemple s'il y a structuration à l'intérieur de taches.

3.3.1. Y a-t-il une structure dans les taches ?

On va simplement effectuer le test vu en 3.1 conditionnellement à la disposition des sites sains, c'est-à-dire permuter entre elles uniquement les valeurs non nulles de façon à conserver la position et la taille des groupes de points de valeur non nulle. Ainsi, les valeurs de ces points sont indépendantes les unes des autres. Autrement dit, on va faire un test d'indépendance totale par permutation mais avec des permutations restreintes aux attaqués.

On peut ainsi souhaiter regarder d'abord si il y a un effet moyen de la distance au bord des taches, et prendre comme critère :

$$C(d) = \frac{1}{J(d)} \sum_{(i,j)\in\mathcal{J}(d)} X_{(i,j)}, \quad \text{critère (1)}$$

où $\mathcal{J}(d)$ est l'ensemble des points attaqués à distance d du bord des taches, $J(d)$ leur nombre. L'hypothèse alternative associée à ce critère remet ainsi en cause le fait que les valeurs des notes à des distances différentes du bord de la tache soit de même distribution. On pourra aussi prendre un critère de type variogramme si l'on souhaite regarder les effets de dépendance entre notes :

$$C(d) = \frac{1}{I(J-d)} \sum_{1\leqslant i\leqslant I} \sum_{1\leqslant j\leqslant J-d} (X_{(i,j)} - X_{(i,j+d)})^2, \quad \text{critère (2)}$$

L'hypothèse alternative associée à ce critère remet en cause l'indépendance entre notes d'une même ligne. Notons que la formule proposée intègre les notes

nulles. Elle prend donc en compte en particulier (pour $d = 1$) l'écart entre le bord extérieur ($X_{(i,j)} = 0$) et intérieur ($X_{(i,j)} > 0$) de la tache. Si on désire exclure cet écart, il suffit de restreindre la somme aux couples dans les taches.

On constate sur la figure 7 que les lavandes attaquées ne se répartissent pas indépendamment entre elles, mais que des lavandes proches ont globalement des notes plus semblables que sous le hasard, avec un fort écart à l'indépendance.

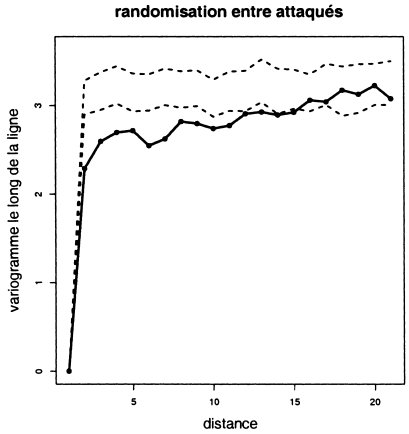

FIGURE 7
Test d'indépendance entre notes de dépérissement des lavandes, critère (2)

3.3.2 Y a-t-il une structure autre que moyenne dans les taches ?

Supposons une maladie, comme une carence, qui se développe si un élément n'est pas présent dans le sol au dessus d'un seuil. Si la zone d'étude présente des sous-zones très contrastées où l'on est soit au dessus, soit au dessous du seuil, le test précédent conviendra car il permet de tester l'hypothèse nulle « les observations dans les taches sont indépendantes et de même distribution ».

Si on suppose que l'on a une évolution régulière du taux de présence de cet élément, on s'attend à ce que les individus dans les taches soient plus ou moins attaqués selon leur distance au bord de la tache, sans qu'il y ait nécessairement transmission de maladie. Dans ce cas, le test précédent détectera une structure, liée à la présence d'un changement de moyenne (*i.e.* il verra que les individus au bord de la tache sont moins attaqués que ceux au centre).

Par contre, si on désire tester l'existence d'une transmission dans ce cadre, il est nécessaire de prendre en compte cet effet de moyenne. Si transmission il y a, on s'attend à ce que, si un point est plus contaminé que les autre points à même distance, les points voisins le soient aussi, soit parce que cela traduit une virulence plus forte, soit parce que cela traduit une avance dans l'attaque. On va alors permuter entre elles les notes des individus atteints étant à la même distance du bord de la tache. Ceci permet de tester l'hypothèse d'indépendance des notes entre points attaqués conditionnellement à leur distance au bord de la tache, et donc conditionnellement à

une non-stationnarité d'ordre 1 créée par la distance au bord de la tache. On pourra utiliser ensuite un critère comme le variogramme.

Appliqué au cas de la lavande, ce test détecte effectivement une dépendance entre plants voisins sur la ligne (figure 8). On perd beaucoup en puissance par rapport au test précédent, ce qui n'est pas étonnant car, dans une maladie par transmission, on s'attend aussi à trouver les individus les plus atteints au centre des taches. Le test précédent (présenté au paragraphe 3.3.1 et avec le critère 2) cumule donc un effet d'ordre 1 (moyennes variant continûment avec la distance au bord des taches) et d'ordre 2 (deux individus proches auront plus souvent que sous le hasard des écarts de note à leur moyenne respective semblables).

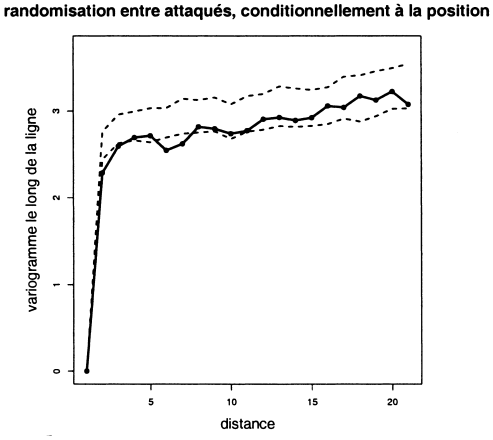

FIGURE 8
Test d'indépendance entre notes de dépérissement des lavandes, conditionnellement à leur position dans les taches, critère (2)

Si on soupçonne une différence de moyennes entre taches, on peut effectuer le même test, mais en conditionnant les permutations simultanément à l'appartenance à la tache et à la distance au bord.

4. Analyse de deux images

Par la suite on notera X et Y les deux images correspondant à deux mesures dans le même site aux mêmes points. On commence généralement par une analyse séparée des deux images. Ensuite, la première question qui se pose est en général assez large.

Considérons par exemple la parcelle expérimentale de vigne située près de Bordeaux et représentée en figure 9. Elle est formée de 5 lignes espacées de 2m. Sur chacune d'elles sont plantés 66 pieds espacés de 1m.

L'oïdium est une maladie qui se développe rapidement sur feuilles à partir d'un nombre relativement faible de foyers, pour envahir quasi complètement la parcelle en

FIGURE 9
Pourcentage de feuilles par cep atteintes par l'oïdium mesuré le 02 juin (haut), et pourcentage de grappes par cep attaquées à plus de 75 % par l'oïdium mesuré le 28 septembre (bas) sur une parcelle expérimentale. Plus le cep (la feuille) est attaqué(e), plus le pixel correspondant est clair.

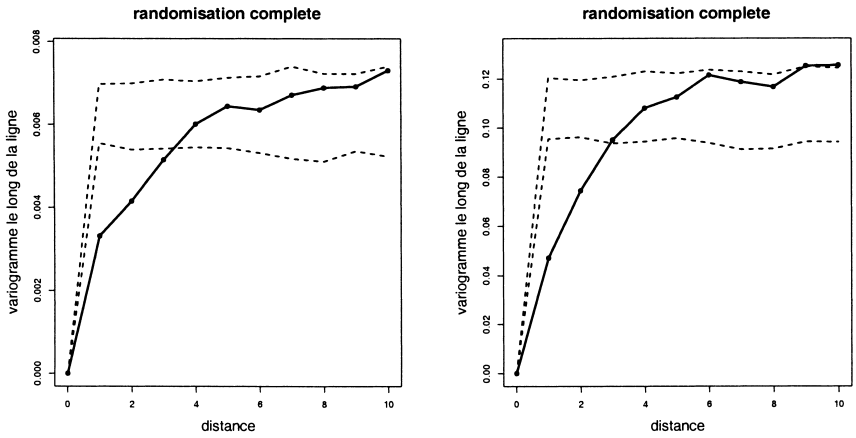

FIGURE 10
Tests d'indépendance des répartitions d'attaques sur feuilles (à gauche) et sur grappes (à droite) par permutation totale, dans l'exemple de l'oïdium de la vigne

automne si aucun traitement n'est prodigué. Elle attaque les grappes et peut causer une dépréciation du vin si le pourcentage de grappes fortement attaquées est important. Il est alors souhaitable de pouvoir repérer le plus tôt possible les zones susceptibles de porter de fortes proportions de telles grappes, de façon à pouvoir soit mettre en œuvre des traitements localisés, soit mettre en place une récolte sélective.

Il est intéressant de baser une technique de repérage sur l'analyse des dégâts sur feuilles, plus précise que sur l'attaque sur grappes. La première étape est donc d'analyser les dépendances entre pourcentage de feuilles attaquées début juin (figure 9 haut) et le pourcentage de grappes attaquées à plus de 75 % (figure 9 bas) observé en septembre.

Sur la figure 9, on note que les deux taches de forte attaque des feuilles coïncident avec celles sur grappes, mais de nombreuses autres taches sont présentes. L'analyse des répartitions spatiales des deux images par permutation totale (figure 10) montre

des variogrammes très similaires et un rejet de l'hypothèse d'indépendance dans les deux cas. Analyser le niveau d'attaque sur grappes en fonction de celui sur feuilles, par régression par exemple, n'est pas licite car l'indépendance entre erreurs n'est plus respectée. Il faut donc passer par des méthodes appropriées, capables de prendre en compte cette dépendance.

4.1. Tester l'indépendance entre images

L'hypothèse d'indépendance entre images, que l'on supposera issues d'un processus stationnaire, repose sur l'idée que, même s'il existe une structure spatiale dans chaque image, ces structures sont disposées au hasard l'une par rapport à l'autre. Ainsi par exemple, si on translate l'une des deux images, on doit observer statistiquement la même chose au niveau du groupe de deux images. S'il y a une structure spatiale dans les images, on va comme précédemment chercher à la conserver dans le test, sans quoi ce dernier mélangera test d'indépendance entre images et test d'indépendance entre deux points de la même image. Deux options se présentent encore, selon le type de structure rencontrée :

- si l'une des images, par exemple Y, n'est structurée que le long des lignes (lignes indépendantes et iid), on pourra permuter les lignes de Y entre elles et regarder la concordance entre X et $Y^{(\phi)}$ le long des lignes.
- si les deux images développent des structures dans les deux directions, on va translater aléatoirement l'une des deux structures, par exemple Y, avant de regarder la concordance entre X et $Y^{(\phi)}$ le long des lignes. En pratique, on utilise la convention du tore, *i.e.* on referme la grille sur elle-même de la même façon que pour les lignes au paragraphe 3.2. Partant d'un vecteur (v_x, v_y) choisi aléatoirement uniformément dans le tore ainsi formé, on va comparer X et $Y^{(\phi)}$ où $Y^{(\phi)}_{(i,j)} = Y_{(i+v_x, j+v_y)}$.

Le critère de comparaison doit mesurer l'accord entre images. On peut alors s'intéresser aux effets de moyenne. On pourra prendre par exemple le covariogramme défini par :

$$C(d) = \frac{1}{I(J-d)} \sum_{1 \leqslant i \leqslant I} \sum_{1 \leqslant j \leqslant J-d} (X_{(i,j)} - X_{(i,j+d)})(Y_{(i,j)} - Y_{(i,j+d)})$$

Ce test a été mis en œuvre dans l'exemple précédent (figure 11). L'ensemble des translations possibles est de 5*66 translations. On note un rejet de l'hypothèse d'indépendance, le covariogramme étant plus grand que sous le hasard. Les deux variables évoluent donc en moyenne dans le même sens quand on se déplace d'un point donné à un autre point.

4.2. Tester l'indépendance conditionnelle d'une image par rapport à l'autre

Dans le cadre par exemple de suivi temporel d'images, la question n'est pas vraiment de savoir si deux images successives sont indépendantes entre elles. La

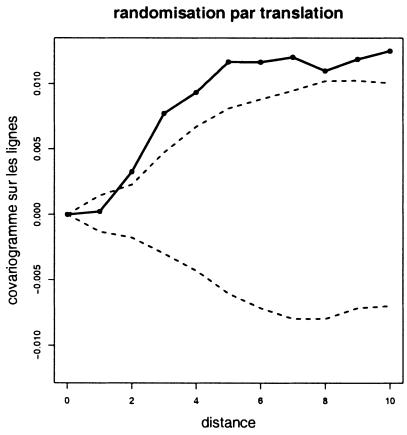

FIGURE 11
Test d'indépendance entre attaques sur feuilles et sur grappes, dans l'exemple de l'oïdium de la vigne

permanence des structures spatiales va en effet assurer une dépendance entre images. Il s'agit plutôt de savoir si les valeurs $Y_{(i,j)}$ observées en un point à la deuxième date ne dépendent que de la valeur $X_{(i,j)}$ observée en ce même point à la première date, ou s'il reste une dépendance malgré la prise en compte des valeurs à la première date.

Cette hypothèse peut se schématiser comme en figure 12 (a) où deux points sont reliés par une arête s'ils sont dépendants conditionnellement à la connaissance du reste. Ainsi, sur ce schéma, les points Y_1 et Y_2 sont indépendants sachant X_1 et X_2. L'hypothèse alternative se schématisera par la figure 12 (b) : il reste une dépendance entre Y_1 et Y_2 une fois prises en comptes les dépendances entre Y_1 et X_1, Y_2 et X_2, X_1 et X_2.

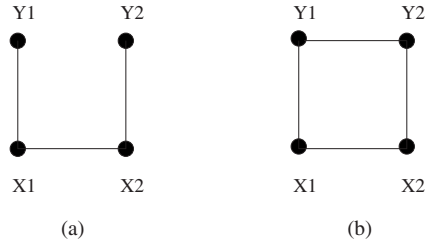

FIGURE 12
Schémas de dépendance.
(a) Y_1 et Y_2 sont indépendants conditionnellement à X_1 et X_2
b) Y_1 et Y_2 sont dépendants conditionnellement à X_1 et X_2

Pour tester cela, on va se restreindre aux permutations qui redistribuent au hasard entre elles les valeurs $Y_{(i,j)}$ ayant même valeur $X_{(i,j)}$ (*i.e.* recenser tous points

ayant une valeur donnée à la première date, puis permuter les valeurs $Y_{(i,j)}$ entre ces points, et recommencer pour toutes les valeurs possibles de la première image X). Cela suppose que les observations correspondant à l'image X sont discrètes.

La dépendance spatiale sera ensuite mesurée en utilisant par exemple comme critère le variogramme des pourcentages d'attaque de la deuxième image ou en gardant le covariogramme. On notera que le cas étudié au paragraphe 3.3.2 est un cas particulier de ce dernier.

Si on étudie ainsi la répartition des attaques de l'oïdium sur les grappes en fonction des attaques sur feuilles, on constate un rejet très fort de l'indépendance conditionnelle (figure 13). La bande de confiance ne contient plus 0 dès que la distance dépasse 1 mètre, traduisant l'existence d'un effet de moyenne : les moyennes des notes sur grappes à notes sur feuilles fixées sont différentes. En permutant entre elles ces notes sur grappes à notes sur feuilles fixées on garde ces moyennes et le covariogramme sous indépendance conditionnelle mesure cette différence de moyennes. La courbe calculée sur les données observées est en dessous de la bande de confiance, ce qui traduit une relation résiduelle entre notes sur grappes conditionnellement aux notes sur feuilles.

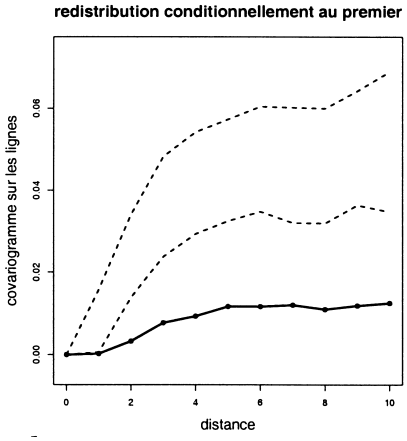

FIGURE 13
Test d'indépendance entre attaques sur grappes conditionnellement aux attaques sur feuilles, dans l'exemple de l'oïdium de la vigne

Il se peut que la liaison entre les deux images ne soit pas aussi simpliste, mais que les valeurs entre deux points de la deuxième image soient indépendantes conditionnellement à des voisinages de taille plus importante. Par exemple, il se peut que la valeur $Y_{(i,j)}$ dépende aussi des deux voisins le long de la ligne : $\big(X_{(i,j-1)}, X_{(i,j)}, X_{(i,j+1)}\big)$. On peut alors explorer cette hypothèse de la même façon, en élargissant la zone par rapport à laquelle on conditionne.

4.3. Tester l'égalité des structures de deux sous-populations dans une image

Le paragraphe précédent s'intéressait à l'existence d'une structure spatiale dans la deuxième image, une fois prise en compte l'information donnée par l'image à la première date. De façon générale, on s'est intéressé jusqu'à présent à tester la nullité de la dépendance.

On se propose ici de tester une égalité : l'égalité, sur la seconde image, des structures spatiales présentes dans deux sous-ensembles définis par la structure spatiale observée sur la première image. Si par exemple la première image X mesure l'existence de deux types de sols, la deuxième Y une note de maladie, on voudrait savoir si les deux types de sols ont une influence sur la répartition de la maladie et son intensité. Supposons que ces deux zones soient définies par $X_{(i,j)} = 0$ pour la première, $X_{(i,j)} = 1$ pour la seconde.

On va comme en 4.1 utiliser les translations aléatoires de la deuxième image dans le cas général, des permutations ou des rotations aléatoires des lignes de la deuxième image si seul un effet le long des lignes est présent. On va par contre jouer sur le critère. Deux possibilités sont offertes :

- chercher s'il y a un effet moyen, donc prendre comme critère la moyenne dans chacun des niveaux identifiés sur X, soit $C_0 = \frac{\sum_{ij}(1-X_{(i,j)})Y_{(i,j)}}{\sum_{ij}(1-X_{(i,j)})}$ et $C_1 = \frac{\sum_{ij} X_{(i,j)}Y_{(i,j)}}{\sum_{ij} X_{(i,j)}}$, ou la différence des moyennes $C_0 - C_1$. On peut raffiner ce critère en définissant des classes de distances (*i.e.* les points dans un niveau donné de X à une distance donnée du point le plus proche de l'autre niveau) et en regardant la moyenne par niveau et par classe.
- chercher un effet de structure, et regarder un critère comme le variogramme dans chaque classe, ou leur différence.

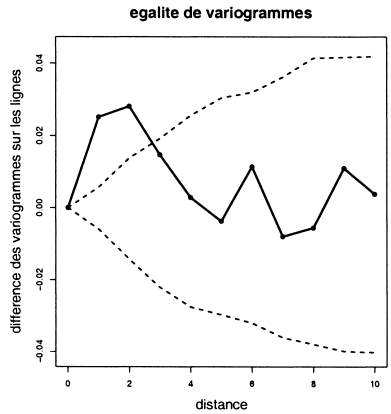

FIGURE 14

Test d'égalité des structures des notes sur grappes de la partie gauche et droite de la parcelle, basé sur le calcul d'une différence de variogrammes, dans l'exemple de l'oïdium de la vigne

À titre d'exemple, nous avons distingué deux zones (droite et gauche) dans la parcelle et regardé si les structures spatiales des notes sur grappes, observées à l'aide du variogramme, y étaient les mêmes. En effet, la vigueur des plants ne semble pas uniforme sur la parcelle, et cette variable peut avoir une influence sur le développement de la maladie. Les résultats obtenus sont illustrés en figure 14. Les structures sont différentes dans les deux sous-zones, les différences apparaissant surtout pour des distances de 1 et 2 mètres. Au-delà, aucune différence significative n'est mise en évidence.

4.4. Problème de la censure

Les tests de permutation reposent sur le principe d'égalité des probabilités des réalisations, qu'elles soient observées ou simulées. Dans le cas d'une censure, c'est-à-dire du masquage d'une partie des données, on ne pourra plus respecter cette hypothèse.

Considérons le cas d'une maladie sans rémission. Supposons que l'on veuille regarder la position relative des individus attaqués à deux dates, en utilisant par exemple la méthode proposée en 4.1, mais en prenant comme critère la distribution de distances entre un point atteint à la deuxième date et le point le plus proche atteint à la première date. L'image X correspondra aux observations à la date 1 ($X_{(i,j)} = 0$ si l'individu est sain, $X_{(i,j)} = 1$ s'il est attaqué à la date 1). L'image Y correspondra aux observations à la date 2 ($Y_{(i,j)} = 0$ si l'individu n'a pas été attaqué à la date 2, $Y_{(i,j)} = -1$ s'il est attaqué à la date 1, et donc si on ne sait pas s'il est ou non attaqué à la date 2, $Y_{(i,j)} = 2$ s'il est attaqué à la date 2). Les données censurées ($Y_{(i,j)} = -1$) correspondront aux positions où $X_{(i,j)} = 1$.

À chaque permutation (que ce soit une rotation si on peut supposer les lignes indépendantes, ou une translation dans le cas général), on va déplacer par exemple les individus attaqués à la date 1, soit l'image X. Trois situations posant *a priori* problème vont alors se rencontrer :

- des points où $X^{(\phi)}_{(i,j)} = 0$ mais $Y_{(i,j)} = -1$ (et non 0 ou 2) car, si les points attaqués à la date 1 sont translatés dans X, ils n'ont pas bougé dans Y.
- des points où $X^{(\phi)}_{(i,j)} = 1$ et $Y_{(i,j)} = 2$, donc considérés comme attaqués à la date 1 une fois le processus translaté.
- des points où $X^{(\phi)}_{(i,j)} = 1$ et $Y_{(i,j)} = 0$, également considérés comme attaqués à la date 1 une fois le processus translaté.

Les deux dernières situations ne devraient pas poser de problème si l'on considère qu'il peut y avoir plus d'une attaque en un point, à des temps différents. Ainsi, le deuxième cas va correspondre dans le processus observé aux points attaqués à la date 1, soumis à une deuxième attaque à la date 2. La troisième situation va correspondre aux points attaqués à la date 1 qui n'ont pas été réattaqués à la date 2. Mais dans la première situation, on ne sait pas si les points attaqués à la date 1, mais considérés comme non-attaqués après translation à cette même date, ont été attaqués ou non à la date 2.

Deux grands cas vont généralement se présenter, la censure par un facteur externe (supposé indépendant), ou la censure due à la nature des données, comme ci-dessus. Dans les deux cas cependant, le type de permutation sera comme précédemment lié à l'hypothèse à tester. Par contre, on va chercher à symétriser les situations observée et simulées, en restreignant Y, pour chaque permutation ϕ, à la zone Δ_ϕ formée de l'intersection des zones non censurées du processus observé et du processus après permutation. On calcule donc le critère d'intérêt sur les données observées et sur les données après permutation en se restreignant aux points dans Δ_ϕ. Puis on calcule la différence entre les deux valeurs du critère ainsi obtenues. Le rang du 0 parmi les différences calculées nous permettra de décider si il y a une interaction ou non. On en trouvera une démonstration dans (Chadœuf 2000) dans le cadre de données sur support continu. Le cas de grilles se démontre de façon similaire, seul l'ensemble de permutation est différent.

Dans les deux cas, on peut représenter la situation comme l'analyse de deux images X et Y, avec une censure, une image B dont les points valent 0 ou 1 selon que les points sont censurés ou non. Dans le cas d'une censure due à un facteur externe, on supposera B indépendant de X et Y pour analyser la dépendance entre X et Y. Dans le cas d'une censure due aux données, $B_{(i,j)}$ sera égal à $1 - X_{(i,j)}$.

Si ϕ est une permutation aléatoire de la grille, $\Delta_\phi = B \cap B^{(\phi)}$ est l'ensemble des points non censurés à l'origine et après permutation. On note alors :

- $C_1^{(\phi)}$ la valeur du critère d'intérêt calculée sur $\left\{X_{(i,j)}\right\}$ et $\left\{Y_{(i,j)}/B_{(i,j)}B_{(i,j)}^{(\phi)} = 1\right\}$, c'est-à-dire sur les valeurs avant permutation de X, et les valeurs de Y restreintes à la zone non censurée avant et après permutation de X,
- $C_2^{(\phi)}$ sa valeur calculée sur $\left\{X_{(i,j)}^{(\phi)}\right\}$ et $\left\{Y_{(i,j)}/B_{(i,j)}B_{(i,j)}^{(\phi)} = 1\right\}$, c'est-à-dire sur les valeurs après permutation de X, et les valeurs de Y restreintes à la zone non censurée avant et après permutation de X,
- $I^{(\phi)} = C_1^{(\phi)} - C_2^{(\phi)}$ leur différence.

On va alors s'intéresser à la distribution de $I^{(\phi)}$ dont la valeur sur les données observées correspond à ϕ égale à l'identité et vaut 0.

Le principe est toujours le même. On sait que sous H_0, $I^{(\phi)}$ est d'espérance nulle. On va donc calculer la distribution de $I^{(\phi)}$ et rejeter si le cas observé ($I = 0$) est trop improbable. Dans le cas d'un test bilatéral, on rejette l'hypothèse que X et Y sont spatialement indépendants si la probabilité d'observer une différence nulle est plus petite que 2,5 % ou plus grande que 97,5 %

En pratique, cela consiste à tirer au hasard k permutations sous l'hypothèse d'indépendance et à regarder le rang de 0 dans l'ensemble $0, I^{(\phi_1)}, .., I^{(\phi_k)}$.

À titre d'exemple, nous avons considéré la parcelle de lavande atteinte de dépérissement et supposé que les gravités d'attaque correspondent à des dates d'attaque (figure 15). Ainsi, la note 5, la plus grave, correspondrait aux premières attaques, la note 4 aux attaques suivantes. Les notes 3, 2 et 1 aux attaques les plus jeunes. En ne regardant que les deux notes les plus graves, on obtient la répartition donnée en figure 15.

FIGURE 15
Répartition des notes les plus graves de dépérissement. Codage des pixels : blanc = «attaque à la date 1» (censure), noir = «pas attaqué aux dates 1 et 2», gris = «attaque à la date 2»

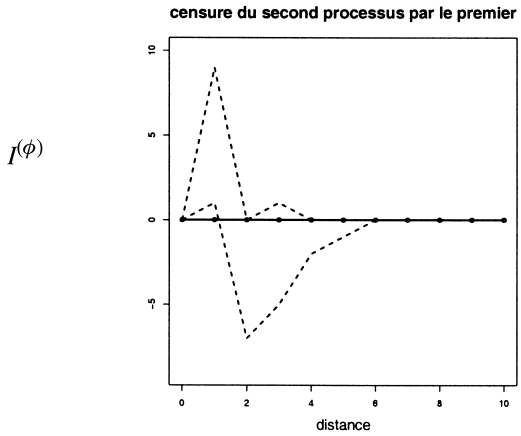

FIGURE 16
Test d'indépendance entre dates correspondant aux deux notes les plus graves, dans l'exemple de la lavande

Si on suppose que les lavandes atteintes à la première date restent atteintes à la seconde, on entre dans le cadre présenté de censure de la deuxième date (notes 4) par la première (note 5). Le test d'indépendance entre dates est fait comme présenté ci-dessus, en utilisant des translations aléatoires associées à une convention du tore. On obtient alors la figure 16 qui montre que l'hypothèse est rejetée à $d = 1$ m.

Il est à noter que cette procédure, si elle permet de prendre en compte l'existence d'une censure, coûte relativement cher en terme de puissance. Elle marchera d'autant mieux que la censure représente une proportion faible de la surface étudiée.

5. Extensions

5.1. Processus dans un espace continu

Les procédures d'exploration d'hypothèses par tests de permutation ne sont pas limitées aux seuls processus sur grille. On peut aussi les utiliser dans le cadre de la statistique géométrique (Stoyan *et al* 1995).

Cette dernière s'intéresse aux répartitions aléatoires d'objets dans un espace. Les plus courants sont des points (processus ponctuels), des structures linéaires (processus de fibres) ou des taches (processus booléens). À ces structures peuvent être attachées des valeurs appelées marques. Ainsi par exemple :

- dans le cas d'un processus de points représentant une répartition d'arbres, on attachera à chaque point une valeur qui peut être sa hauteur, son statut (dominé, sous-dominant, dominant) ou son espèce.
- dans le cas d'un processus de fibres représentant une structure de fissuration du sol, on pourra attacher une notion d'âge, mais aussi d'épaisseur, si on la considère négligeable en elle-même mais pas l'une par rapport à l'autre.

Les types de permutations possibles vont alors varier selon la nature de l'objet et les hypothèses disponibles. Ainsi par exemple, dans le cas d'un processus ponctuel marqué, deux types de procédures sont possibles pour tester l'indépendance entre marques (Diggle 1983, Manly 1997) :

- permuter les marques au hasard parmi les points du processus. Cette procédure permet de tester l'agrégation des marques parmi les points du processus. Elle est utilisée par exemple en épidémiologie. Les deux marques seront alors individu sain ou contaminé et cette méthode permettra rapidement de décider si la maladie apparaît en foyer ou au hasard.
- translater au hasard le processus formé des points ayant l'une des marques et garder le reste fixe.

La première procédure casse la structure marginale du processus formé de l'une des marques. Elle est donc parfaitement adaptée quand la marque n'apparaît qu'après le processus de points. Par contre, si on considère deux espèces d'arbres, on peut s'attendre à ce que chacun d'eux ait une structure (en bouquets par exemple) et la question est alors de savoir si les deux structures sont indépendantes entre elles. La deuxième procédure permettra alors de conserver chaque structure.

Les processus de fibres et de taches posent des problèmes spécifiques de censure. On trouvera dans (Berman 1986) l'utilisation des tests de permutation pour tester l'indépendance entre un processus de fibres et un processus de points. Le test de l'interaction entre fibres d'un processus de fibres et le problème des effets de bord, *i.e.* la connaissance partielle de la longueur des fibres censurées, sont abordés dans (Monestiez *et al* 1993). Dans (Chadœuf *et al* 2000) est abordé le problème de censure par un processus de taches.

5.2. Gérer l'hétérogénéité à grande échelle

Dans les paragraphes précédents, nous avons présenté l'application globale de la procédure à l'ensemble de l'image. Dans le cas de grands jeux de données, il va souvent apparaître des gradients qui vont rendre les tests significatifs alors que ce gradient ne constitue pas le phénomène d'intérêt, mais une nuisance.

On peut alors tester localement, et parcourir l'ensemble de la parcelle. La cartographie de la valeur du critère comme de la significativité du test pourra alors être utilisée pour explorer la variabilité spatiale du phénomène en minimisant l'effet du gradient. On pourra aussi, moyennant quelques précautions, réunir ces tests locaux

dans un test global, voir (Allard *et al* 2001) pour l'interaction entre deux processus de points, (Brix *et al* 2001) pour l'analyse de la répartition d'un processus et (Couteron *et al* 2002) pour son application sur des données forestières.

Références

ALLARD D., BRIX A. & CHADŒUF J. (2001), Testing local dependence between two point processes. *Biometrics*, 57, 508-517.

BERMAN M. (1986), Testing for spatial association between a point process and another stochastic process. *Applied Statistics*. 35, 54-62.

BRIX A., SENOUSSI R., COUTERON P. & CHADŒUF J. (2001), Assessing goodness of fit of spatial inhomogeneous Poisson processes. *Biometrika*, 88, 2, 487-497.

CHADŒUF J., BRIX A., PIERRET A. & ALLARD D. (2000), Testing local dependances on images. *Journal of Microscopy,* 200, 1, 32-41.

COUTERON P., SEGHIERI J. & CHADŒUF J. (2003), A test to investigate spatial relationships between neighbouring plants in plots of varying plant density. *Journal of Vegetation Science*, 14, 163-172.

CRESSIE N. (1993), *Spatial statistics, Second edition*. John Wiley & Sons, Chichester, 436 p.

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, 436p.

MANLY B.F.J. (1997), *Randomization, bootstrap and Monte Carlo methods in biology.* Chapman and Hall; London, 399 p.

MIELKE P. & BERRY K. (2001), *Permutation methods. A distance function approach.* Springer, New York. 252 p.

MONESTIEZ P., KRETZSCHMAR A. & CHADŒUF J. (1993), Modelling natural burrow systems in soil by fibre processes : Monte-Carlo tests on independance of fibre characteristics. *Acta Stereologica* 12, 237-242.

STOYAN D., KENDALL W.S. & MECKE J. (1995), *Stochastic geometry and its applications. Second edition.* John Wiley & Sons, Chichester, 436 p.

WACKERNAGEL H. (1995), *Multivariate geostatistics. An introduction with applications.* Springer, Berlin, 256 p.