Université Montpellier II

Ecole Doctorale Information, Structures, Systèmes

Quelques contributions aux statistiques spatiales ; applications en environnement et en écologie

Denis Allard

Habilitation à Diriger les Recherches

Soutenue le 10 décembre 2007 à Montpellier devant le jury composé de :

Jean-Marc Azaïs, Université de Toulouse
Jean-Noël Bacro, Université Montpellier II
Gilles Celeux, INRIA, rapporteur
Marc Genton, Universités de Genève et Texas A&M, rapporteur
Xavier Guyon, Université Panthéon-Sorbonne
Christian Lavergne, Univsersité Montpellier III, rapporteur

En mémoire de Vincent, compagnon de promenades et de discussions passionnées.

Nous avions le projet de soutenir notre HdR la même année.

Table des matières

1 Introduction **11**

2 Géostatistique pour la télédétection **13**
2.1 Modélisation par le variogramme 13
2.2 Modéliser l'hétérogénéité pour le changement d'échelle 14
2.3 Modélisation par le variogramme d'ordre 1 17
2.4 Modéliser l'évolution temporelle des scènes 18
2.5 Conclusion 19

3 Détection de structures dans les champs aléatoires **21**
3.1 Détection de zones de changement abrupt 21
3.2 Classification non supervisée pour données géostatistiques 26
3.3 Arbres de régression 29
3.4 Interpolation et classification 32

4 Champs aléatoires gaussiens dissymétriques **35**

5 Processus de points dans le plan **41**
5.1 Introduction 41
5.2 Classification non supervisée d'un mélange de processus de points 42
5.3 Tests locaux d'indépendance 44
5.4 Déformation de l'espace 48

6 Recherches en cours et futures **51**

7 Annexes : étudiants encadrés, publications et autres références **55**

Résumé

Ce mémoire est organisé en trois parties de longueurs inégales. La première, composée des sections 2 à 4 sont consacrées aux données de type géostatistique, en suivant une progression allant des applications les plus directes de la géostatistique pour la télédétection (section 2) à la détection de motifs (i.e. zones de changement abrupt, classification non supervisée, arbre de régression) dans les champs aléatoires gaussiens (section 3), puis à l'exploration d'une nouvelle classe de champs aléatoires dissymétriques (section 4).

La seconde partie, constituée d'une section unique porte sur les recherches effectuées pour les processus de points spatiaux, en abordant successivement (i) la classification non supervisée d'un mélange de processus de points, (ii) les tests locaux d'indépendance entre processus de points et (iii) l'utilisation de la déformation de l'espace pour rendre stationnaire un processus non stationnaire.

La dernière section brosse à grands traits mon programme de recherche pour les prochaines années, qui se feront essentiellement dans le prolongement des recherches passées : détection de structures (zones de changement abrupt pour données non gaussiennes, détection d'agrégats, classification), champs aléaoires gaussiens-dissymétriques, création d'un générateur climatique stochastique.

Abstract

This document is organized in three parts of unequal length. The first part deals with geostatistical data, from rather direct applications of geostatisical tools for remote sensing data in Section 2, to the detection of patterns in spatial data (zones of abrupt change, clustering, regression trees) in Section 3, to spatial skew-normal random field models in Section 4.

The second part has a unique section and presents the developpements carried out for the analysis of spatial point processes : (i) unsupervised clustering in spatial point processes, (ii) testing local independance between point processes and (iii) using a space deformation in order to stationarize a non stationary point process.

The last section is devoted to a brief description of my research program for the next years : pattern detection (zones of abrupt change for non gaussian data, cluster detection, spatial classification) skew-normal random fields, and random climate generator.

Remerciements

La recherche est un métier fait de curiosité, de passion, de tenacité, de doutes, de grandes joies et de petites frustrations. Mais la recherche, c'est avant tout un métier fait de rencontres. Certaines sont déterminantes et conditionnent les grandes orientations que l'on se donne. Je voudrais ici remercier toutes les personnes qui ont rendu le travail que je présente ici possible, et qui à des titres divers ont positivement influencé le cours de ma carrière de chercheur.

En tout premier lieu, bien sûr, un hommage à mon *alma mater*, le Centre de Géostatique (CG) de l'Ecole des Mines de Paris, à Fontainebleau. Je remercie chaleureusement H. Wackernagel et P. Chauvet pour m'avoir accepté et encadré en DEA, A. Galli pour m'avoir proposé un sujet de thèse dont le titre cryptique (*"Connexité des ensembles aléatoires : application à la modélisation des réservoirs pétroliers hétérogènes"*), avait stimulé ma curiosité au point de rester à Fontainebleau trois ans de plus, et enfin G. Matheron pour m'avoir acceuilli au CG. Je tiens surtout à rendre hommage à mon directeur de thèse, Christian Lantuéjoul, qui reste pour moi un modèle de rigueur dans le raisonnement et d'épure dans l'exposition. Pierre Chauvet avait l'habitude de dire de Christian qu'il était le Flaubert de la géostatistique. Merci Christian d'avoir exprimé ton côté flaubertien dans *Geostatistical Simulation : Models and algorithms*, ce livre remarquable qui jette un pont entre morphologie mathématique, processus aléatoires spatiaux et algorithmes de simulation. C'est devenu l'un de mes livres favoris.

Au département de stastistiques de l'Université du Washington à Seattle, grâce à Paul Sampson et Peter Guttorp, je découvris un domaine d'application nouveau (pour moi, j'entends) : les statistiques pour l'environnement. Je fis également la rencontre, déterminante, d'Adrian Raftery et de son groupe de travail sur les modèles statistiques en classification, auquel Adrian m'a proposé de participer. J'y ai appris les statistiques à l'américaine : efficaces, soucieuses de leur applicabilité, partant des problèmes effectivement posés et faisant usage de la révolution informatique. J'y ai également appris que les bonnes statistiques appliquées sont aussi difficiles à réaliser que les bonnes statistiques mathématiques, et que sous d'autres cieux elles avaient toute leur place dans les cursus universitaires. Grâce à Adrian, ce séjour d'un an m'a convaincu de poursuivre une carrière de chercheur en statistiques appliquées. Je le remercie de tout cœur pour le rôle déterminant qu'il a joué à cette époque et pour son soutien permanent depuis.

Une opportunité m'a été offerte dès l'année suivante lorsque, de retour en Europe, j'appris par Paul Sampson qu'un poste était ouvert à Avignon en statisiques spatiales pour les applications en environnement et en écologie. C'est ainsi que j'ai rejoint, en 1996, l'unité d'Avignon, qui s'appelait alors Unité de Biométrie[1]. Cette unité a été le cadre des recherches que je présente dans ce mémoire. De près ou de loin, tous les collègues de cette unité sont associés à mon travail.

[1] On sait ce que terme désigne aujourd'hui. Afin de ne pas apparaître comme une officine du ministère de l'intérieur, l'unité a changé de nom en 2006 pour s'intituler Biostatistique et Processus Spatiaux.

Il règne au sein de l'unité d'Avignon un je-ne-sais-quoi d'unique, fait d'esprit de collaboration (plutôt que de compétition), de chaleur et de fraternité. Sans une base arrière solide, le risque existe d'être en difficulté lors des petits accidents de la vie. L'unité sut jouer le rôle d'amortisseur lorsque cela fut nécessaire, et j'en remercie collectivement l'ensemble de mes collègues, avec une mention spéciale pour mes voisins Montfavetains.

La recherche se nourrit de rencontres et d'échanges, donc de collaborations hors de son unité d'origine. Je dois à Jean-Noël Bacro de m'avoir intéressé au *Wombling*, un bien beau sujet qui nous a donné du fil à retordre. Je me souviens d'une démonstration que nous avions tenté de terminer dans un chalet d'altitude à Rochebrune, entre un blind-test des vins de Savoie organisé par Eric Parent et un atelier sur la sous-estimation des incertitudes chez les experts. Il va sans dire que la démonstration ne fut trouvée qu'une fois de retour dans l'unité.

Philippe Naveau est à l'origine de mon investissement sur les skew-normal distribution. Les premières équations furent jetées sur la table d'un café de Barcelone mais nos principales avancées furent réalisées lors de deux séjours à Boulder, Colorado.

Il est parfois utile de créer des lieux de rencontres et d'échanges scientifiques nouveaux. C'est ainsi que, constatant qu'il n'existait pas de forum pour parler de statistiques pour l'environnement, nous avons créé en 2001 avec Eric Parent, Liliane Bel, Philippe Naveau et d'autres comparses le groupe "Environnement" de la Société Française de Statistique. Ce groupe organise régulièrement des rencontres autour de thèmes ciblés.

Je tiens également à mentionner et à remercier Edith Garbiel, Sébastien Garrigues et plus récemment Cédric Flecher, les étudiants que j'ai eu la chance et le bonheur d'encadrer, sans qui une grande partie des travaux exposés dans ce mémoire n'existeraient pas.

Marc Genton, des universités de Genève et Texas A&M, travaille depuis de nombreuses années sur des thèmes proches des miens, en particulier sur les skew-normal distribution. Il m'a fait l'honneur de consacrer du temps et de l'énergie pour ce rituel de passage typiquement français, et je lui en suis très reconnaissant.

Gilles Celeux, de l'INRIA, côtoie depuis de nombreuses années le département MIA et en particulier l'unité d'Avignon. J'ai toujours apprécié la pertinence de ses points de vue. Je le remercie chaleureusement de continuer à accompagner les statistiques à l'INRA en rapportant mes travaux.

Christian Lavergne, de l'université de Montpellier III, est le rapporteur le plus proche géographiquement. A travers les enseignements en Master et l'acceuil des doctorants d'I2S, l'unité appartient bien à l'aire bio-statistique de Montpellier. Je suis très heureux qu'il ait accepté de rapporter mes travaux.

C'est un privilège pour moi de pouvoir soutenir mon Habilitation devant Xavier Guyon, spécialiste reconnu des statistiques spatiales, qui a de tout temps accordé intérêt et soutien aux

travaux issus de l'unité d'Avignon. Le fait qu'il ait été membre du jury qui m'a recruté à l'INRA me touche particulièrement.

Jean-Marc Azaïs est un ancien collègue de l'INRA ; il connaît bien la maison. Il avait déjà accepté d'être membre du comité de pilotage puis rapporteur de la thèse de mon ancienne étudiante, Edith Gabriel, rôle qu'il avait accompli avec beaucoup de pertinence et de gentillesse. Je me suis permis de le solliciter une fois de plus, et je le remercie chaleureusement d'avoir accepté.

Jean-Noël Bacro est un vieux complice resté jeune, habitué des trajets Paris-Avignon, puis Montpellier-Avignon (et retour). Depuis longtemps il m'encourageait à soutenir mon Habilitation. Aujourd'hui il me fait l'amitié de participer à ce jury et je l'en remercie tout particulièrement. J'espère qu'à l'avenir nous aurons de nouvelles occasions de collaborer ensemble.

1 Introduction

Les recherches présentées dans ce mémoire ont été réalisées dans l'unité de recherche Biostatistique et Processus Spatiaux (anciennement Biométrie) du département Mathématiques et Informatique Appliquées de l'INRA. Ces recherches sont centrées autour des statistiques spatiales. L'INRA est un organisme de recherche finalisé dont les objets de recherche appartiennent au triptyque agriculture-alimentation-environnement ; la recherche en statistique y est donc au service de ces objets finalisés. Elle s'organise le plus souvent comme une série d'allers-retours entre des préoccupations précises, voire opérationnelles, posées par les biologistes (terme entendu au sens large), et des recherches plus génériques qui participent aux avancées méthodologiques en statistiques spatiales. Ces recherches méthodologiques sont tantôt issues des questions posées par les applications traitées, tantôt issues de préoccupations endogènes au champ de la statistique. Elles ont vocation à soutenir les recherches dans les domaines finalisés de l'INRA, mais trouvent parfois, lorsque les développements statistiques sont suffisamment génériques, des applications dans des domaines non prévus au départ. Lors de la phase de retour vers le biologiste, ces méthodes lui permettent de réinterroger ses données, ses hypothèses et parfois même sa façon de formaliser ses questions.

La formation à et par la recherche figure explicitement dans les missions de l'INRA. À ce titre les chercheurs de l'INRA participent pleinement à l'encadrement de thèses. Dans un département de Mathématiques et Informatique Appliquées appartenant à un institut comme l'INRA, on y distingue généralement deux types de doctorants, et donc deux types de thèses. Il y a d'une part les étudiants ayant une reçu une formation en mathématiques qui se spécialisent dans le domaine de la statistique, et d'autre part ceux ayant reçu une formation en biologie ou en agronomie, qui viennent acquérir, dans leur discipline d'origine, une formation complémentaire en statistique, faisant d'eux des scientifiques maîtrisant l'utilisation des méthodes statistiques les plus récentes. J'ai eu la chance d'encadrer une thèse dans chacune de ces catégories.

Les chercheurs en statistique à l'INRA sont donc soumis à la double exigence de produire de la recherche en statistique à la fois reconnue dans leur discipline propre, et donc publiée dans les journaux de statistique, et à la fois utile à la recherche scientifique dans les domaines de l'INRA, et donc publiée dans les revues des différents domaines d'application. Ma liste de publications reflète ce positionnement. Son analyse détaillée fait apparaître une absence de publications en revues de 2001 à 2005, suivie d'une accumulation en 2006 et 2007. Cela est dû à une conjonction de facteurs qui ont agi de façon simultanée : choix de supports différents (chapitres de livres, conférences avec sélection et proceedings), des délais de publication parfois exagérés, mais aussi et surtout la mise en route de nouvelles questions de recherches liées au coencadrement des deux thèses qui ont démarré en 2001.

J'ai choisi de présenter mes travaux en les organisant par grands thèmes de recherche afin d'en dégager les grandes lignes. Cette organisation ne respecte par contre ni l'ordre chronologique ni

leur importance, tant au plan de la recherche en statistique, que pour les applications considérées.

On classe habituellement les statistiques spatiales en trois grands domaines, qui correspondent à des types de données distincts. Lorsque les données peuvent être mesurées en tout point d'un domaine continu (p. ex. les teneurs en matière organique dans une parcelle agricole) on utilise le formalisme des champs aléatoires continus ; c'est le domaine habituel de la **géostatistique**. Lorsque, par leur nature même, les données sont liées à un réseau (pour des images ou des données récoltées sur des entités administratives, etc.), il est certes possible d'utiliser le formalisme de la géostatistique, mais celui des **champs de Markov** est souvent plus adapté. Enfin, lorsque ce sont les coordonnées qui portent l'information principale (positions des arbres dans une forêt, des points d'impact la foudre, etc.), on parle de **processus de points**.

Bien entendu, les frontières entre ces trois domaines ne sont pas étanches. Ainsi, par exemple, je montre dans la section 2 que le formalisme géostatistique apporte des outils pertinents pour l'analyse et la modélisation de l'hétérogénéité spatiale sur des images issues de la télédétection. À la section 3.2 nous verrons comment les formalismes des champs markoviens et de la géostatistique se comparent pour la classification non supervisée de données géostatistiques (c'est-à-dire non directement liées à un lattice).

Le plan général de ce mémoire s'appuie sur cette typologie. Les sections 2 à 4 sont consacrées aux données de type géostatistique, en suivant une progression allant des applications les plus directes de la géostatistique pour la télédétection (section 2) à l'exploration d'une nouvelle classe de champs aléatoires dissymétriques (section 4) en passant par la détection de motifs (zones de changement abrupt, classification non supervisée, arbre de régression) dans les champs aléatoires gaussiens (section 3). La section 5 porte ensuite sur les recherches effectuées pour les processus de points spatiaux. Enfin, la dernière section brosse à grands traits mon programme de recherche pour les prochaines années.

Concernant les citations et les références, j'ai adopté la convention suivante. Les articles dont je suis co-auteur sont référencés par un numéro d'ordre [X] qui renvoie à ma liste de publications, classées des plus récentes aux plus anciennes. Les références à d'autres auteurs sont indiquées par les noms des auteurs suivis de l'année de publication. Par exemple, Proust (1927).

2 Géostatistique pour la télédétection

Les travaux présentés dans ce chapitre sont pour l'essentiel des applications assez directes de concepts géostatistiques connus, tels qu'ils sont exposés dans Chilès & Delfiner (1999) par exemple. En télédétection, la géostatistique est essentiellement utilisée à des fins descriptives mais reste très peu utilisée dans ces aspects de modélisation. Les travaux présentés ci-dessous constituent un apport de la modélisation par les champs aléatoires au domaine de l'imagerie satellite. Ils sont tous issus de la thèse de Sébastien Garrigues [T3], co-encadrée avec F. Baret (INRA, CSE, Avignon), qui portait sur l'étude de l'hétérogénéité intra-pixel pour les capteurs à faible résolution spatiale (pixels de 500 à 1000 m de côté).

Pour ces résolutions spatiales se posent plusieurs questions, dont par exemple : quelles sont les relations entre les caractéristiques statistiques des réflectances mesurées par de tels capteurs et celles mesurées par des capteurs haute résolution, type SPOT ? Quelle est l'influence de l'hétérogénéité spatiale sur le calcul des paramètres biophysiques habituels (NDVI, LAI,...), en terme de biais notamment ? Comment utiliser la haute résolution temporelle fournie par ce type de capteurs ?

En posant le cadre méthodologique approprié, il s'agissait de déboucher sur des concepts utilisables en imagerie satellite pour traiter certains aspects des questions posées ci-dessus. Sans aller jusqu'à la mise au point d'outils opérationnels, l'exigence était d'en fournir les principes méthodologiques. Du point de vue du statisticien (ou plutôt du géostatisticien, dans ce cas précis), il s'agissait d'organiser le transfert d'outils méthodologiques en privilégiant robustesse et opérationalité. Les publications [8], [7], [3], [2] et [17] sont issues de cette thèse.

2.1 Modélisation par le variogramme

Une étude bibliographique [8] a montré que la modélisation par les champs aléatoires du second ordre que propose la géostatistique est l'outil le plus pertinent pour étudier les scènes stationnaires, pour les objectifs fixés dans le cadre de la thèse. On rappelle brièvement quelques notions concernant les champs aléatoires. Un champ aléatoire, $Z(\cdot)$, défini sur un domaine D, est dit stationnaire au second ordre si les deux premiers moments (espérance et covariance) existent et sont invariants par translation :

$$E[Z(s)] = \mu, \quad E[\{Z(s) - \mu\}\{Z(s+h) - \mu\}] = C(h)$$

pour tout s et tout $s + h$ appartenant à D. Le variogramme associé à Z est la fonction

$$\gamma(h) = \frac{1}{2} E\left[\{Z(s) - Z(s+h)\}^2\right], \quad h \in D.$$

Sous les hypothèses stationnaires d'ordre 2, on a la relation $\gamma(h) = \sigma^2 - C(h)$, avec $\sigma^2 = C(0) = \lim_{|h|\to\infty} \gamma(h)$.

Bien souvent les réflectances, mesurées dans le rouge (R) et dans le proche infra-rouge (PIR), sont résumées par un indice univarié, le NDVI, défini par NDVI=(PIR-R)/(PIR+R). Le NDVI est directement relié à la capacité photsynthétique de la plante (Myneni et al., 1995). Dans ce cadre univarié on a montré, sur la banque d'images SPOT de 3 × 3 km du projet VALERI (`www.avignon.inra.fr/valeri`), que la modélisation par le variogramme permet de proposer une typologie des scènes étudiées en utilisant deux paramètres seulement [8] : la variance, σ^2, et un paramètre d'échelle, $D_c = \sqrt{A}$, avec

$$A = \frac{1}{\sigma^2} \int_{\mathbf{R}^2} C(h)\, dh.$$

On peut montrer par un théorème ergodique sur les champs aléatoires (Lantuéjoul, 2002) que la variance de la moyenne spatiale $\bar{Z}_V$ d'une variable Z sur un domaine V de surface $|V|$ vaut approximativement

$$\mathrm{Var}(\bar{Z}_V) = \mathrm{Var}\left(\frac{1}{|V|}\int_V Z(x)\, dx\right) \simeq \frac{\sigma^2 A}{|V|}$$

lorsque $|V| >> A$. Le paramètre A est homogène en unité à $|V|$, et représente la surface d'influence d'une donnée indépendante dans V. On pose que la stationnarité peut être admise sur une scène I si la variance de $\bar{Z}_I$ est faible par rapport à σ^2, c'est-à-dire si la portée intégrale est inférieure à 5% de la surface de I. Sur une image SPOT de 3 × 3 km, cela correspond à poser $D_c \leq 670$m. La plupart des sites agricoles sont caractérisés par une variance élevée, et un facteur d'échelle D_c inférieur au seuil en raison du parcellaire ; une exception notable dans la base VALERI est le site de Fundulea (Roumanie), dont les parcelles sont de très grandes dimensions (cf. section 2.4). Les sites forestiers sont plutôt caractérisés par des variances faibles. Les facteurs d'échelle peuvent être très variables, mais dans le cas où celui-ci est très élevé, cela s'accompagne généralement d'une variance très faible.

Une approche similaire a été proposée pour un contexte multivarié [2] pour lequel les variogrammes des variables PIR et R sont modélisés simultanément en utilisant le modèle de corégionalisation linéaire (Wackernagel, 2003).

2.2 Modéliser l'hétérogénéité pour le changement d'échelle

Modèle univarié

La modélisation par le variogramme permet également de quantifier la perte de variabilité associée à une perte de résolution des capteurs. Cette perte de variabilité entraîne un biais lorsqu'on applique une transformation non linéaire à la variable d'étude. Dans [7], on quantifie ce biais pour les champs aléatoires stationnaires d'ordre 2, ce qui permet de proposer une méthode pour corriger les produits biophysiques construits à partir des capteurs à moyenne résolution. Soit z_i une variable régionalisée mesurée en tout pixel s_i d'une image I. On note $z_v = 1/n \sum_{i \in v} z_i$ la

valeur moyenne de z sur un domaine v de I. Ici, v sera l'équivalent d'un pixel moyenne résolution. On s'intéresse à la transformée $y = f(z)$, p. ex.

$$y = LAI = \frac{-1}{K_{NDVI}} \ln\left(\frac{z - z_\infty}{z_0 - z_\infty}\right),$$

où z représente le NDVI, et z_∞, z_0 et K_{NDVI} sont des constantes calibrées. Sur un domaine v, la valeur exacte $y_v = 1/n \sum_{i \in v} f(z_i)$ est approchée par $\tilde{y}_v = f(z_v)$. En supposant $z_i - z_v$ petit, l'erreur commise peut être approchée au second ordre :

$$e_v = f(z_v) - y_v = \frac{1}{n} \sum_{i \in v} f'(z_v)(z_i - z_v) - \frac{1}{n} \sum_{i \in v} \frac{f''(z_v)}{2}(z_i - z_v)^2 + R.$$

Le premier terme s'annule, et le second fait intervenir la variance de dispersion dans le volume v :

$$s^2(\cdot \mid v) = \frac{1}{n} \sum_{i \in v} (z_i - z_v)^2.$$

En négligeant le reste R, l'erreur devient donc

$$e_v \simeq -\frac{f''(z_v)}{2} s^2(\cdot \mid v).$$

Dans le formalisme des fonctions aléatoires, la variance de dispersion et l'erreur sont des variables aléatoires. L'espérance de l'erreur est le biais

$$b_v = E[e_v] = -\frac{f''(z_v)}{2} E\left[s^2(\cdot \mid v)\right] = -\frac{f''(z_v)}{2} \bar{\gamma}(v, v)$$

où

$$\bar{\gamma}(v, v) = \frac{1}{|v|^2} \int_v \int_v \gamma(h - h')\, dh\, dh'$$

$\gamma(h)$ étant le variogramme de $Z(s)$.

Au second ordre, le biais ne dépend que de deux quantités qui interviennent de façon multiplicative :

1. $f''(z_v)$, qui représente la non linéarité de la fonction de transfert f au point z_v ; cette quantité est facile à déterminer, soit analytiquement, soit numériquement.
2. La variabilité de la fonction aléatoire dans le domaine v, mesurée par la variance de dispersion théorique $\bar{\gamma}(v, v)$. Celle-ci ne dépend que du variogramme, supposé connu.

Modèle bivarié

En réalité, le NDVI n'est lui-même qu'un indice synthétique obtenu en combinant les réflectances mesurées dans deux longueurs d'onde, le rouge, noté $r(s)$, et le proche infra-rouge, noté

$p(s) : z(s) = \{p(s) - r(s)\}/\{p(s) + r(s)\}$. En remplaçant cette expression dans l'équation liant le LAI au NDVI on obtient :

$$y = g(p(s), r(s)) = \frac{-1}{K_{NDVI}} \ln \left(\frac{\frac{p(s)-r(s)}{p(s)+r(s)} - z_\infty}{z_0 - z_\infty} \right).$$

On suppose bien sûr que les valeurs des constantes z_∞, z_0 et K_{NDVI} sont telles que l'expression ci-dessus existe toujours. Avec un raisonnement similaire à celui du paragraphe précédent, l'erreur commise est $e_v = g(p_v, r_v) - \sum_{i \in v} g(p_i, r_i)$ qui, grâce à un développement au second ordre, s'approche par

$$e_v = -\frac{1}{2} \mathrm{tr} \left\{ \mathbf{H}_g(p_v, r_v) \mathbf{C}(\cdot \mid v) \right\}$$

où $\mathbf{H}_g(p_v, r_v)$ est la matrice hessienne de g calculée en (p_v, r_v) et $\mathbf{C}(\cdot \mid v)$ est la matrice de variance-covariance de dispersion calculée de façon similaire à $s^2(\cdot \mid v)$ dans le cas univarié.

Lorsqu'on considère que le vecteur de variables $(p(s), r(s))$ est la réalisation d'un champ aléatoire bivarié et que l'on considère l'espérance de l'erreur, on obtient le biais de l'erreur :

$$\begin{aligned} B_v = E[e_v] &= -\frac{1}{2} \mathrm{tr} \left\{ \mathbf{H}_g(p_v, r_v) E[\mathbf{C}(\cdot \mid v)] \right\} \\ &= -\frac{1}{2} \mathrm{tr} \left\{ \mathbf{H}_g(p_v, r_v) \bar{\mathbf{\Gamma}}(v, v) \right\} \end{aligned}$$

où

$$\bar{\mathbf{\Gamma}}(v, v) = \begin{pmatrix} \bar{\gamma}_{p,p}(v, v) & \bar{\gamma}_{p,r}(v, v) \\ \bar{\gamma}_{r,p}(v, v) & \bar{\gamma}_{r,r}(v, v) \end{pmatrix}$$

est la matrice de variance-covariance de dispersion calculée à partir de l'ajustement des variogrammes simples $\gamma_{p,p}(\cdot)$ et $\gamma_{r,r}(\cdot)$ et du variogramme croisé,

$$\gamma_{p,r}(h) = \frac{1}{2} E[\{P(s) - P(s+h)\}\{R(s) - R(s+h)\}],$$

réalisé par exemple par le modèle de corégionalisation linéaire (Wackernagel, 2003).

L'expression du biais est plus compliquée dans le cadre bivarié que dans le cas univarié ; notamment son signe dépend du produit des matrices $\mathbf{H}_g(p_v, r_v)$ et $\bar{\mathbf{\Gamma}}(v, v)$.

Dans [7], on montre que corriger le biais (univarié) de $f(z_v)$ permet de diminuer l'erreur quadratique commise sur $f(z_v)$ de 20% à 80% sur 11 sites étudiés sur 12. Le site sur lequel la correction dégrade l'estimation est un site très particulier composé d'un petit nombre de très grandes parcelles, dont les tailles sont du même ordre de grandeur que les pixels moyennes résolutions. Utiliser un modèle stationnaire d'ordre 2 dans ce cas n'a que peu de sens. Corriger le biais bivarié de $y_v = g(p_v, r_v)$ a moins de succès. En effet, d'une part l'estimation des variogrammes bivariés est plus difficile, et d'autre part l'erreur due à la non-linéarité entre (p_v, r_v) et $NDVI = z_v$ compense en partie l'erreur due à la non-linéarité entre z_v et $LAI = y_v = f(z_v)$.

Une difficulté subsiste pour pouvoir corriger le biais de façon opérationnelle selon cette méthode : il est nécessaire de connaître la valeur de $\bar{\gamma}(v,v)$ ou de $\bar{\mathbf{\Gamma}}(v,v)$. Pour l'instant cela nécessite d'avoir une image haute résolution pour estimer un modèle de variogramme puis calculer la variance de dispersion, mais dans ce cas on peut s'interroger sur l'utilité des images basses résolutions... Nous présenterons à la section 2.4 une modélisation spatio-temporelle du NDVI, qui permet une estimation de la quantité $\bar{\gamma}(v,v)$ à une date non échantillonnée par image haute résolution.

2.3 Modélisation par le variogramme d'ordre 1

Plusieurs modèles de champs aléatoires peuvent partager la même fonction de covariance et le même histogramme théorique (voir p. ex. Chilès & Delfiner, 1999). En particulier, la mosaïque définie par un réseau de droites poissoniennes d'intensité a_m, dont les valeurs dans les cellules sont des variables aléatoires i.i.d., définit un champ aléatoire $Z_m(s;a_m)$ qui possède une covariance exponentielle $C_m(h) = \sigma^2 \exp(-|h|/a_m)$, où σ^2 est la variance des variables aléatoires. Mais il est évidemment aussi possible de construire un champ gaussien, $Z_g(s;a_g)$, possédant cette fonction de covariance, et pour lequel on peut choisir $a_g = a_m$. Le variogramme habituel (variogramme d'ordre 2) ne permet donc pas de discriminer entre ces deux classes de champs aléatoires.

Le variogramme d'ordre 1, défini par

$$\gamma_1(h) = \frac{1}{2} E\left[|Z(s) - Z(s+h)|\right],$$

est un outil complémentaire au variogramme habituel d'ordre 2. En effet, lorsque la v.a. remplissant les cellules de la mosaïque est une gaussienne centrée et réduite, on peut montrer que le variogramme d'ordre 1 de $Z_m(s;a_m)$ est $\gamma_{1,m}(h;a_m) = \gamma_m(h;a_m)/\sqrt{\pi}$. Pour un champ aléatoire gaussien centré normé, on a en revanche $\gamma_{1,g}(h;a_g) = \sqrt{\gamma_{2,g}(h;a_g)}/\sqrt{\pi}$. En d'autres termes, la relation qui lie les variogrammes d'ordre 1 et 2 est linéaire pour le champ aléatoire mosaïque tandis qu'elle est quadratique pour un champ aléatoire gaussien. Ainsi, on voit que considérer les variogrammes d'ordre 1 et 2 simultanément permet de discriminer entre ces deux classes de champs aléatoires.

Pour les hypothèses ci-dessus, on considère dans [3] que la variable mesurée (le NDVI, par exemple) est issue d'un champ aléatoire qui est la somme pondérée des deux champs définis ci-dessus :

$$Z_\omega(s) = \mu + \sigma^2 \left\{ \omega Z_g(s;a_g) + (1-\omega^2)^{1/2} Z_m(s;a_m) \right\}, \quad 0 \le \omega \le 1. \tag{1}$$

Le paramètre ω^2 est le poids relatif des variations modélisées par un champ gaussien, tandis que $1-\omega^2$ est le poids relatif des variations modélisées par un champ mosaïque.

$Z_\omega(s)$ est un champ aléatoire dont la loi marginale est une $N(\mu,\sigma^2)$ et dont le variogramme d'ordre 2 est

$$\gamma_\omega(h) = \sigma^2\omega^2[1-\exp(-|h|/a_g)] + \sigma^2(1-\omega^2)[1-\exp(-|h|/a_m)].$$

Notons que si $a_g = a_m = a$, on a $\gamma_\omega(h) = \sigma^2[1 - \exp(-|h|/a)]$ pour tout $\omega \in [0, 1]$. On montre que pour les champs Z_ω ci-dessus, le variogramme d'ordre 1 s'écrit

$$\gamma_1(h) = \frac{\sigma}{\sqrt{\pi}} \left\{ \omega(1 - \gamma_m(h; a_m))\sqrt{\gamma_g(h; a_g)} + \gamma_m(h; a_m)\sqrt{\omega^2\gamma_g(h; a_g) + (1 - \omega^2)} \right\} \qquad (2)$$

Dans [3], ces modèles théoriques ont été utilisés sur des images simulées selon le modèle (1) et sur des scènes de la base VALERI. Les paramètres structuraux (ω, a_g, a_m) sont estimés par une méthode de moindres carrés. Sur les images simulées selon le modèle, la réestimation des paramètres est bonne ou très bonne. Pour appliquer cette approche sur les images de la base VALERI, on fait l'hypothèse que l'indice de végétation peut se décomposer en deux variables, dont l'une varie continûment sur le domaine selon un champ aléatoire gaussien, et dont l'autre est distribuée spatialement comme une mosaïque poissonienne. Ces hypothèses reviennent à supposer que le NDVI est distribué selon un modèle aléatoire, avec un effet aléatoire correspondant aux parcelles et un résidu aléatoire auto-corrélé. En outre, on fait l'hypothèse (assez hasardeuse) que les parcelles sont réparties selon une mosaïque aléatoire.

Pour tous les sites de culture on observe que $\hat{\omega}^2 \leq 0.5$ tandis que pour les sites de végétation plus continue (forêts, prairies naturelles, etc...) $\hat{\omega}^2$ est toujours très proche de zéro. Ces résultats montrent que la composante mosaïque propre aux parcellaires est malgré tout correctement détectée.

Considérer des modèles de paysages plus réalistes et pour lesquels on peut calculer la fonction de covariance apporterait probablement une amélioration notable.

2.4 Modéliser l'évolution temporelle des scènes

Les capteurs hautes résolutions spatiales (famille SPOT p. ex.) ont un temps de retour au nadir[2] d'une scène de plusieurs semaines, ce qui est trop peu fréquent pour les suivis de végétation en zones de cultures. A l'inverse, les capteurs basses ou moyennes résolutions ont un temps de retour de l'ordre de la journée.

Il faut donc combiner la haute résolution spatiale des uns avec la haute fréquence temporelle des autres. Une façon de parvenir à ce résultat est de proposer une modélisation spatio-temporelle des scènes. Dans [17] on propose une modélisation spatio-temporelle parcimonieuse pour une scène acquise à 21 dates différentes dans une zone de grande culture en Roumanie (Fundulea). On fait les hypothèses suivantes : i) le parcellaire est décrit selon le processus de droites poissoniennes décrit dans la section 2.3 ; ii) à chaque parcelle correspond une culture d'hiver (probabilité p) ou une culture de printemps (probabilité $q = 1 - p$) ; iii) le NDVI moyen pour ces deux types de cultures est connu et représenté par les courbes $W(t)$ et $S(t)$.

[2]Point ou ensemble de points de la surface du globe, situés directement sous un capteur à mesure que celui-ci se déplace le long de son orbite.

On modélise le NDVI observé par

$$Z(s,t) = Y(s) + \begin{cases} W(t) & \text{si } s \text{ appartient à une culture d'hiver} \\ S(t) & \text{si } s \text{ appartient à une culture de printemps,} \end{cases}$$

où $Y(s)$ est un champ aléatoire stationnaire d'ordre 2, d'espérance nulle et de variogramme $\gamma_Y(h)$. On en déduit le variogramme de $Z(s,t)$ à la date t :

$$\gamma(h;t) = \gamma_Y(h) + pq\{W(t) - S(t)\}^2\gamma_m(h),$$

$\gamma_m(h)$ est le variogramme de la mosaïque poissonnienne introduit dans la section 2.3.

Au temps t, la variance de $Z(s,t)$, notée $\sigma_Z^2(t)$, est donc liée à la variance de $Y(s)$, notée σ_Y^2, par $\sigma_Z^2 = \sigma_Y^2 + pq(W(t) - S(t))^2$. Cette équation permet une estimation par moindres carrés de pq et de σ_Y^2, et par suite, une estimation des variogrammes $\gamma_m(h)$ et $\gamma_Y(h)$ à partir des variogrammes spatiaux calculés sur les différentes images aux temps $t_1, \ldots, t_n$. Des essais réalisés sur la succession d'images de Funduela ont montré que 5 images seulement suffisent pour pouvoir estimer les variances de dispersion de $Z(\cdot)$ à une date t, notée $\bar{\gamma}(v,v;t)$, pour les 16 autres dates échantillonnées avec une erreur inférieure à 20%. On dispose ainsi des valeurs $\bar{\gamma}(v,v;t)$ nécessaires à la correction du biais présentée à la section 2.2 à des dates non échantillonnées.

2.5 Conclusion

D'une façon générale, ces travaux ont montré l'utilité de l'outil géostatistique en télédétection pour décrire, modéliser et prédire les grandeurs biophysiques d'intérêt. Ils ouvrent également deux pistes de recherche ?

1. La construction de modèles stochastiques de paysages plus réalistes que la mosaïque poissoniennes, pour lequel l'estimation des paramètres soit possible ; un tel modèle doit présenter une organisation hiérarchique de la tesselation (les limites de parcelles aboutissent à des chemins, qui eux-mêmes aboutissent à des routes, etc.) ainsi que la possibilité de définir des orientations privilégiées.
2. L'estimation pour des modèles statistiques spatio-temporels de l'évolution du NDVI dans ces paysages. Ces modèles doivent être capables de prendre en compte de façon plus complète les écarts phénologiques entre parcelles, p. ex. les écarts liés à des variétés ou dates de semis différentes.

3 Détection de structures dans les champs aléatoires

Une des hypothèses les plus simples pour les champs aléatoires présentant des dépendances spatiales est l'hypothèse de stationnarité d'ordre 2, qui peut aisément être enrichie en ajoutant un modèle de régression spatiale à coefficients inconnus sur des covariables spatialisées. Dans ce cas l'hypothèse de stationnarité porte sur les résidus du modèle de régression. Ce modèle enrichi est non stationnaire, et toute la non-stationnarité est portée par le terme d'espérance sous la forme d'un modèle de régression.

Mais la non-stationnarité est bien souvent plus complexe que celle décrite ci-dessus. Sous le terme générique de *structures* nous appelons ici une non-stationnarité du champ aléatoire ou du champ de résidu qui ne se résume pas à un modèle de régression à filtrer. Nous nous intéresserons plus particulièrement à deux types de structures : les changements abrupts et la classification spatialisée.

3.1 Détection de zones de changement abrupt

Ce travail a été le sujet de la thèse d'Edith Gabriel [T2] soutenue en décembre 2004[3]. Les publications [4], [29], [26], [21] et [1] sont issues de sa thèse.

Problématique générale

Pour de nombreuses études en sciences de l'environnement ou en écologie il est intéressant de pouvoir détecter les zones où les variables présentent des changements abrupts. L'exemple historique est celui traité dans Womble (1951) qui s'intéresse aux changements abrupts pour des fréquences d'allèles afin de les mettre en relation avec des frontières environnementales entre populations. Cette approche a ensuite été généralisée dans Barbujani *et al.* (1989) et Bocquet-Appel & Bacro (1994). Ces méthodes sont purement numériques. En particulier les algorithmes de Wombling détectent toujours des frontières, sans pouvoir évaluer leur significativité. Dans [31], un test basé sur une technique par permutation a été proposé. Cependant celui-ci est d'un intérêt limité car il repose sur une hypothèse nulle d'absence de dépendance spatiale.

Banerjee, Gelfand & Sirmans (2003) propose de détecter si le gradient en un point du domaine est significativement différent de 0. Banerjee & Gelfand (2006) généralise cette approche à la moyenne du gradient le long d'une courbe et propose un algorithme pour construire une courbe le long de laquelle le gradient est significativement non nul. Cette approche présente deux inconvénients principaux. D'une part, l'algorithme nécessite un point de départ, fourni par l'utilisateur, pour lequel le gradient doit être significativement non nul. D'autre part, leur méthode consiste à réaliser des tests locaux en tous points d'une grille, pour un seuil qui ne tient pas compte du problème des tests multiples (le seuil choisi est $\alpha = 0.05$ dans leurs travaux), problème rendu très complexes par la présence des corrélations spatiales. Cela conduit à détecter

[3] Edith Gabriel a obtenu le prix Marie-Jeanne Laurent-Duhamel 2006 de la SFdS pour son travail de thèse.

des points ou des courbes significatives pour des champs simulés sous l'hypothèse stationnaire d'ordre 2.

L'apport de notre travail a consisté : a) à poser une formalisation du problème qui permet de faire la distinction entre les variations compatibles avec une hypothèse stationnaire d'ordre 2 et les variations brusques, qui rejettent cette hypothèse ; b) à proposer une procédure de test permettant de détecter ces dernières variations.

La méthode a été validée par une étude systématique sur des données simulées en absence et en présence de zones de changement abrupt puis par l'analyse de données de référence. Les données traitées spécifiquement dans le cadre de la thèse d'E. Gabriel correspondaient à des variables de sol mesurées sur une parcelle agricole : teneur en sable, argile, calcaire (variables permanentes) ; teneur en azote minéral et en humidité (variables non permanentes). Sur ces données, on a pu montrer que les zones de changement abrupt détectées correspondent aux transitions entre type de sol différents [1], [21].

Principe général de la méthode

Le problème est formalisé de la façon suivante. La variable d'étude, $Z(\cdot)$, échantillonnée aux points $(s_1, \ldots, s_n)$ d'un champ d'étude D est modélisée sous l'hypothèse nulle d'absence de zones de changement abrupt comme un champ gaussien stationnaire à l'ordre 2. L'hypothèse alternative est que l'espérance de $Z(\cdot)$ présente quelque part sur D des changements abrupts. Il s'agit donc de tester, à l'échelle du domaine étudié, l'existence de changements abrupts de l'espérance du champ $Z(\cdot)$, et dans le cas où l'hypothèse nulle est rejetée de cartographier ces lieux, que nous appellerons les Zones de Changement Abrupts (ZCAs).

La méthode est décrite de façon complète dans [T2], Gabriel (2007) et Allard, Gabriel & Bacro (soumis à publication). On fait les hypothèses suivantes : $\mathcal{H}_1$: le champ $Z(\cdot)$ est gaussien ; $\mathcal{H}_2$: la fonction de covariance $C(s, s')$ de $Z(\cdot)$ est connue ; $\mathcal{H}_3$: la fonction de covariance est stationnaire, $C(s, s') = C(s - s'),\ s, s' \in D$; et $\mathcal{H}_4$: $C(h)$ est indéfiniment différentiable pour tout h tel que $\|h\| > 0$. La première et troisième hypothèses sont usuelles ; la quatrième est vérifiée par un grand nombre de fonctions de covariance, comme la fonction exponentielle, ou plus généralement par la famille de Matern. Sous ces hypothèses, le prédicteur linéaire optimal (krigeage ordinaire) construit à partir de l'échantillon $(Z(s_1), \ldots, Z(s_n))$, est

$$Z^*(s) = C'(s)\mathbf{C}^{-1}Z + (1 - C'(s)\mathbf{C}^{-1}\mathbf{1})\frac{\mathbf{1}'\mathbf{C}^{-1}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}}Z, \tag{3}$$

où $\mathbf{1}$ est le vecteur de longueur n d'élément 1, $C(s)$ est le vecteur de covariance entre s et les points de données et $\mathbf{C}$ est la matrice de covariance entre les données. Notons que le champ $Z^*(\cdot)$ est non stationnaire. Par ailleurs, $Z^*(\cdot)$ est indéfiniment différentiable presque sûrement et en moyenne quadratique pour tout $s \in D$, sauf éventuellement aux points d'échantillonnage. Le

gradient $W(\cdot)$ de $Z^*(\cdot)$ existe donc : $W(s) = (W_1(s), W_2(s))' = (\partial_1 Z^*(s), \partial_2 Z^*(s))'$, avec

$$\partial_i Z^*(s) = \partial_i C(s)' \left(\mathbf{C}^{-1} + \frac{\mathbf{C}^{-1}\mathbf{1}\mathbf{1}'\mathbf{C}^{-1}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}} \right) Z = \partial_i C(s)' \mathbf{K}^{-1} Z,$$

∂_i désignant la dérivée partielle le long de la $i$ème coordonnée, $i \in \{1, 2\}$.

Notons $\mathbf{\Sigma}(s)$ la matrice de covariance du champ interpolé $W(s)$, $\mathbf{\Sigma}(s) = E\left[W(s)W'(s)\right] = \partial C(s)' \mathbf{K}^{-1} \partial C(s)$, où $\partial C(s)$ est la $n \times 2$ matrice $(\partial_1 C(s), \partial_2 C(s))$. Comme $Z(s)$ est un champ gaussien, $W(s)$ est un vecteur gaussien et $\mathbf{\Sigma}(s)^{-1}$ est sa matrice de précision. Nous définissons $T(s)$ par :

$$T(s) = W(s)' \mathbf{\Sigma}(s)^{-1} W(s).$$

Sous l'hypothèse de stationnarité d'ordre 2, $T(s)$ a pour distribution une loi du χ^2 à deux degrés de liberté. En revanche, le champ $T(\cdot)$ n'est pas un champ de χ^2 standard, car il n'est ni stationnaire, ni la somme du carré de champs gaussiens indépendants.

Nous pouvons construire un test local, au point s, pour tester la présence d'un changement abrupt en s. Les hypothèses locales sont $H_0(s)$: $E\left[Z(s')\right] = m$, pour tout s' dans un voisinage de s et $H_1(s)$: le champ $E[Z(\cdot)]$ varie brusquement en s. L'hypothèse locale nulle sera rejetée lorsque $T(s)$ sera supérieur au quantile d'ordre $1-\alpha$ de la distribution d'une loi du χ^2 à deux degrés de liberté, que nous noterons $t_{1-\alpha}$. Nous définissons les ZCAs potentielles comme étant l'ensemble d'excursion de la statistique de test local au-dessus du niveau $t_{1-\alpha}$: $\mathcal{A}_{t_{1-\alpha}} = \{s : T(s) > t_{1-\alpha}\}$.

Si l'on souhaite réaliser non plus un test local en s, mais un test global sur l'ensemble du domaine D, il faut agréger les tests locaux. Ceux-ci sont basés sur une statistique de test $T(s)$ qui est un champ présentant de fortes dépendances spatiales. On ne peut donc pas utiliser les techniques de tests multiples, développées dans le cadre de tests indépendants. Il est plus adapté de développer une approche comparable à celle développée en Imagerie par Résonance Magnétique fonctionnelle (Worsley, 2001). En effet, si H_0 est rejetée, les points s où $H_0(s)$ est rejetée seront organisés en ensembles connexes de grande surface, en tout cas de surface plus grande que sous H_0. Dès lors, on peut baser un test global sur les surfaces des composantes connexes de $\mathcal{A}_{t_{1-\alpha}}$. Notons $\mathcal{C}_{t_{1-\alpha}}$ l'une d'elles et $S_{t_{1-\alpha}}$ sa surface.

Dans le cadre de la théorie des ensembles d'excursion de champs aléatoires (Adler, 2000 ; Cao, 1999), nous avons établi le théorème suivant sur la distribution asymptotique de $S_{t_{1-\alpha}}$ pour un seuil élevé.

Théorème 1 *On se place sous les hypoyhèses $\mathcal{H}_1$ à $\mathcal{H}_4$. Alors, conditionnellement à l'évènement «$T(s_0)$ est un maximum en s_0 de hauteur $T_0 > t_{1-\alpha}$», on a :*

$$t_{1-\alpha} S_{t_{1-\alpha}} \xrightarrow{\mathcal{L}} \pi \det\left(\mathbf{\Lambda}(s_0)\right)^{-1/2} E(2), \text{ quand } t_{1-\alpha} \to \infty, \tag{4}$$

où $E(2)$ est une variable aléatoire exponentielle d'espérance 2 indépendante de $T(\cdot)$ et $\mathbf{\Lambda}(s_0)$ est une matrice associée à la courbure du champ $T(\cdot)$ autour de son maximum dans $\mathcal{C}_{t_{1-\alpha}}$.

Sous H_0, lorsque le niveau $1-\alpha \to 1$, la probabilité pour qu'il y ait plus d'une composante connexe tend très vite vers zéro. Dès lors, un test global au niveau $1-\eta$ revient à tester au même niveau chaque composante connexe. Le test global consiste donc à déterminer la significativité de chaque composante connexe de $\mathcal{A}_{t_{1-\alpha}}$, indépendamment les unes des autres, en utilisant la loi (4). La démonstration de ce théorème est longue et technique. Elle figure dans [T2] et dans Allard, Gabriel & Bacro (soumis).

Mise en œuvre et résultats

Le théorème 1 est vrai sur une partie D du plan $\mathbf{R}^2$. En pratique, toutefois, la méthode doit s'appliquer sur une grille. Le passage de la version continue à une version discrète soulève un problème : il faut trouver d'une part le degré de discrétisation nécessaire et le niveau $1-\alpha$ adéquat correspondant à ce niveau de discrétisation pour le schéma d'échantillonnage considéré. Ce dernier est déterminé par bootstrap paramétrique d'échantillons gaussiens simulés pour le même schéma d'échantillonnage et pour la fonction de covariance estimée. On retient le niveau $1-\alpha$ qui est tel qu'une fraction η des simulations présentent des ZCA significatives.

L'hypothèses $\mathcal{H}_2$ du théorème 1 suppose la fonction de covariance connue, alors qu'en pratique celle-ci doit être estimée. Dans sa thèse, Edith Gabriel a montré sur des simulations que remplacer la fonction covariance par son estimation n'altérait pas le niveau du test sous H_0. De façon classique donc, on estime les paramètres de la fonction de covariance en passant par le variogramme et on remplace dans les équations la covariance par son estimation paramétrique. Sous l'alternative cependant, le variogramme expérimental est biaisé par rapport au variogramme théorique, en raison des variations de $E[Z(\cdot)]$. La difficulté est contournée par une procédure itérative dans laquelle nous estimons à chaque itération alternativement les paramètres du variogramme et les ZCAs. Dans un premier temps nous estimons le variogramme global en utilisant tous les couples de points. En présence de ZCAs, le variogramme est ré-estimé en éliminant tous les couples de données $\{Z(s_\alpha), Z(s_\beta)\}$ tels que le segment $[s_\alpha, s_\beta]$ intersecte une ZCA. Si les paramètres du variogramme changent, le niveau $1-\alpha$ doit être recalculé. Les ZCAs sont alors réestimées à l'aide des nouveaux paramètres. La procédure est réitérée jusqu'à ce que les ZCAs soient identiques pour deux itérations successives. Malgré l'absence de preuve de la convergence de cette procédure, son application aussi bien sur des simulations que sur des données réelles a montré que la convergence était toujours atteinte en moins de 5 itérations. Cela s'explique par le fait que la plupart des couples de points écartés lors de la réestimation du variogramme sont éliminés dès la première itération. Les paramètres de la fonction de covariance varient ensuite assez peu lors des itérations suivantes.

L'analyse de données de sol [1], [21] a montré que cette méthode parvient à détecter les transitions entre types de sols qui correspondent à des variations rapides des variables étudiées.

Puissance locale

La puissance du test local au point s_0 est la probabilité de rejeter l'hypothèse nulle $H_0(s_0)$ lorsque l'hypothèse alternative est vérifiée en ce point. Pour calculer cette puissance il faut spécifier un modèle pour l'hypothèse alternative. Les variations brusques de $E[Z(\cdot)]$ sont modélisées par des discontinuités sur un ensemble de courbes Γ. Localement, l'hypothèse alternative devient $H_1(s_0) : s_0 \in \Gamma$. Plus spécifiquement, on utilise une fonction notée $f(s)$ qui a les propriétés suivantes : elle est lisse, sauf le long d'une discontinuité, avec $\int_{\mathbf{R^2}} f(s)\, ds = 0$ et $f(s) \to 0$ lorsque l'on s'éloigne de la discontinuité. On utilise une fonction gaussienne signée pour modéliser la fonction perturbation :

$$f(s) = a\sqrt{\frac{\pi}{2}}\frac{L}{4}\text{sign}\left\{\sin\left(\phi(s) - \theta\right)\right\} g(4||s - s_0||/L), \tag{5}$$

où L est le paramètre d'échelle de la discontinuité, g est la densité gaussienne, $\phi(s)$ est l'angle de la droite (s, s_0) et sign est la fonction signe. Cette fonction présente une discontinuité d'amplitude a perpendiculairement à une droite Δ_{s_0} d'angle θ passant par s_0. L'hypothèse alternative locale est alors définie par l'existence d'une fonction $f(\cdot; s_0, a, \theta, L)$. On peut ainsi approcher une large classe de courbes Γ par une succession de fonctions de perturbation de ce type.

Une fois un modèle d'alternative spécifié, on peut calculer la puissance locale. En effet, notons $1 - \beta(s_0)$ la puissance du test local en s_0. En l'absence d'information sur l'orientation de la discontinuité, on pose que la puissance en s_0 est l'intégrale sur toutes les orientations θ possibles de la discontinuité, $1 - \beta(s_0) = \frac{1}{\pi}\int_0^\pi \left\{1 - \beta(s_0; \theta)\right\} d\theta$, où $1 - \beta(s_0; \theta)$ est la puissance locale lorsque l'orientation de la discontinuité est θ. En pratique cette intégrale est discrétisée en une somme sur quelques orientations.

Sous $H_1(s_0; \theta)$ le gradient estimé en s s'écrit :

$$W(s; \theta) = \partial C'(s)\mathbf{K}^{-1}Z + \partial C'(s)\mathbf{K}^{-1}A(s_0; \theta) = W_{H_0}(s) + k_a(s; \theta), \tag{6}$$

où $A(s_0; \theta)$ est un vecteur de longueur n, d'éléments $f(s_i; s_0, a, \theta, L)$, $i = 1, \ldots, n$. Le premier terme correspond au gradient sous H_0 et le second terme correspond à la contribution au gradient local de la fonction de perturbation ajoutée en s_0. Il est alors possible d'évaluer la puissance locale, pour une fonction de perturbation fixée. Les détails de ces calculs ont été publiés dans [4]. En calculant la puissance pour tous les points d'une grille, on peut cartographier la puissance locale. Une telle carte permet notamment de visualiser aisément les zones où la densité d'échantillonnage n'est pas suffisante pour détecter les zones de changement abrupt.

Perspectives

Nous avons proposé une formalisation de la notion de changement abrupt dans le plan, qui est la généralisation aux dimensions supérieures ou égales à 2 de la notion de rupture pour les séries temporelles. Dans ce cadre, nous avons proposé une méthode pour la détection des zones de

changement abrupt et montré qu'il est possible de calculer une carte de la puissance locale lorsque l'hypothèse alternative est décrite par des courbes qui portent une discontinuité. Ces résultats ont été établis dans le cadre des hypothèses $\mathcal{H}_1$ à $\mathcal{H}_4$. Nous avons montré que les hypothèses $\mathcal{H}_3$ et $\mathcal{H}_4$ peuvent être affaiblies : la stationnarité intrinsèque est suffisante et il suffit que la fonction de convariance soit de classe $\mathcal{C}_3$ lorsque $||h|| > 0$. Par ailleurs, remplacer la fonction de covariance supposée connue (hypothèse $\mathcal{H}_2$) par son estimation ne nuit pas trop aux performances de la méthode [19]. En revanche, nous n'avons jamais remis l'hypothèse gaussienne ($\mathcal{H}_1$) en question. Il est nécessaire maintenant de travailler sur le prolongement de cette approche dans un cadre non gaussien : variables continues qui se ramènent au cas gaussien par transformation (p. ex. données dissymétriques), variables discrètes (p. ex. données génétiques) et variables qualitatives (p. ex. présence/absence).

3.2 Classification non supervisée pour données géostatistiques

La problématique

La classification en contexte spatialisé a essentiellement été étudiée pour les modèles hiérarchiques dans lesquels i) un champ caché, notée $\mathbf{Z}$, est un champ de Markov à nombre d'états finis, K, défini sur un graphe construit à partir des lieux des échantillons et ii) un champ observé, notée $\mathbf{Y}$, est une dégradation du champ caché tel que, conditionnellement à $\mathbf{Z}$, les valeurs Y_i sont i.i.d. C'est le modèle utilisé dans Besag (1986, 1991), Ambroise & Govaert (1998) et d'autres encore.

L'hypothèse d'absence de dépendances spatiales à l'intérieur d'une classe $k \in \{1, \ldots, K\}$ donnée n'est cependant pas adaptée aux données environnementales. En effet, dans de nombreuses applications environnementales et/ou écologiques, les données, géoréférencées par des coordonnées $(s_1, \ldots, s_n)$ dans un domaine $\mathcal{D}$, proviennent le plus souvent de variables présentant des corrélations spatiales. Il faut donc considérer qu'à l'intérieur de chaque classe k, le vecteur $\mathbf{Y}$ présente des dépendances spatiales.

Le problème de la classification pour ce type de modèle a été abordé dans [14] et [32] en proposant un critère de variance prenant en compte les dépendances spatiales. On y montre qu'introduire les corrélations spatiales dans les estimateurs de la moyenne et de la variance des groupes mène à des classifications parfois fort différentes de celles obtenues en ne modélisant pas ces dépendances. On y propose un algorithme itératif pour estimer les corrélations spatiales en même temps que la classification. Dans le sujet de Master [M2] co-encadré avec N. Peyrard, on a proposé une autre approche, dont la résolution relève de l'algorithme EM.

Le modèle est le suivant. On considère qu'il existe une partition du domaine D en K sous-domaines inconnus D_k, qui correspond à K classes pour $\mathbf{Z}$: $Z_{ik} = 1$ lorsque $s_i \in D_k$ et $Z_{ik} = 0$ sinon, avec $P(Z_{ik} = 1) = \pi_k,\ \forall i = 1, \ldots, n$. Pour des sites s_i et s_j dans D_k, on note $Y(s_i) = Y_i$

et

$$E[Y_i] = \mu_k, \quad \mathrm{Cov}(Y_i, Y_j) = \sigma_k^2 \rho_k(s_i - s_j; b_k),$$

où $\rho_k(h; b_k)$ est une fonction de corrélation dont l'ensemble des paramètres est noté b_k. Lorsque s_i et s_j sont dans deux domaines différents, on considère que $\mathrm{Cov}(Y_i, Y_j) = 0$. Les modèles de champs de Markov cachés (HMRF) supposent en général que $f(\mathbf{Y} \mid \mathbf{Z}) = \prod_i f(Y_i \mid Z_i)$. Ici, nous avons

$$f(\mathbf{Y} \mid \mathbf{Z}) = \prod_k g_{n_k}(\mathbf{Y}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

où $\mathbf{Y}_k$ est la restriction de $\mathbf{Y}$ aux n_k sites situés D_k et g_m est la densité gaussienne de dimension m. En outre $\boldsymbol{\mu}_k$ est le vecteur $\boldsymbol{\mu}_k = (\mu_k, \ldots, \mu_k)$, de longueur n_k, et $\boldsymbol{\Sigma}_k$ est la matrice de covariance de dimension $n_k \times n_k$, dont les éléments $\boldsymbol{\Sigma}_k[i, j] = \sigma_k^2 \rho_k(s_i - s_j; b_k)$.

En ce qui concerne $\mathbf{Z}$, on pose le modèle très simple d'indépendance entre sites : $P(\mathbf{Z}) = \prod_i P(Z_i)$, avec $Z_i = (Z_{i1}, \ldots, Z_{iK})$. Il s'agit donc d'un modèle de mélange gaussien spatialisé dont la dépendance spatiale est modélisée sur $\mathbf{Y}$ et non sur $\mathbf{Z}$, comme cela est le cas dans les modèles HMRF.

L'intérêt de considérer ce modèle simplifié est double : il est plus simple à mettre en œuvre qu'un modèle complet dans lequel $\mathbf{Z}$ serait par exemple modélisé par un modèle auto-régressif conditionnel ; d'autre part, il permet une comparaison entre les deux hypothèses d'indépendance : celle de $\mathbf{Y} \mid \mathbf{Z}$ faite dans les HMRF, et celle sur $\mathbf{Z}$ que nous considérons ici.

L'algorithme initial

Connaissant K, le problème de classification consiste à assigner à chaque point s_i une des valeurs $k = 1, \ldots, K$, connaissant les observations $\mathbf{Y}$ et en tenant compte des corrélations spatiales. Pour ce faire, on note $t_{ik} = P(Z_{ik} = 1 \mid \mathbf{Y}, \mathbf{Z}_{-i})$ la probabilité que le site s_i soit dans D_k sachant $\mathbf{Y}$ et les valeurs de $\mathbf{Z}$ sauf au site s_i, et $\mathbf{t}$ la matrice de dimension $n \times K$ correspondante.

Pour ce modèle, l'algorithme EM ne peut pas être mis en œuvre directement car contrairement au mélange gaussien non spatialisé, l'étape E n'aboutit pas à des expressions analytiques simples. Pour résoudre cette difficulté, l'idée est d'utiliser les équations de mises à jour utilisées dans le cas des mélanges gaussiens indépendants, et de les adapter au cas spatialisé. Notons a^q la valeur d'un paramètre a à l'itération q. Pour le calcul de t_{ik}^{q+1}, on remplace la densité gaussienne marginale $f(y_i \mid \mu_k^q, \sigma_k^{2,q})$ par la densité gaussienne conditionnelle $g(y_i \mid Z_{ik}, \mathbf{Z}_{-i} = \mathbf{t}_{-i}^q; \theta^q)$, où θ^q désigne l'ensemble des paramètres estimés à l'itération q lors de l'étape M :

$$t_{ik}^{q+1} = \frac{\pi_k^q g(y_i \mid y_{-i}, Z_{ik} = 1, \mathbf{Z}_{-i} = \mathbf{t}_{-i}^q; \theta^q)}{\sum_{l=1}^K \pi_l^q g(y_i \mid y_{-i}, Z_{il} = 1, \mathbf{Z}_{-i} = \mathbf{t}_{-i}^q; \theta^q)}.$$

Dans l'étape M, les paramètres sont estimés selon un principe mixte ; les paramètres b_k^{q+1} de la fonction de corrélation sont estimés par une méthode des moindres carrés, réputée plus robuste que le maximum de vraisemblance (Cressie, 1993 ; Chilès & Delfiner, 1999), tandis que

les paramètres μ_k^{q+1}, $\sigma_k^{2,q+1}$ et π_k^{q+1} sont estimés par maximum de vraisemblance. Enfin $\pi_k^{q+1} = \sum_{i=1}^n t_{ik}^{q+1}/n$.

Améliorations

La mise à jour de t_{ik}^q est très coûteuse car il faut inverser $K \times n$ matrices de taille $n \times n$. Une modification mineure de l'algorithme permet de ne réaliser qu'une seule inversion en remplaçant $Z_{ik} = 1$ par son estimation t_{ik}^q dans le conditionnement des densités. De la sorte, on ne considère qu'une seule matrice $\mathbf{t}^q$, et en utilisant l'écriture Gibbsienne, il en résulte que l'ensemble des $n \times K$ lois conditionnelles s'obtiennent en une seule inversion. Par ailleurs, pour éviter la sur-estimation des variances, on estime $\sigma_k^{2,q}$ par

$$\sigma_k^{2,q} = \frac{\sum_{i=1}^n t_{ik}^q (y_i - \eta_i^q)^2}{\sum_{i=1}^n t_{ik}^q} \quad \text{avec} \quad \eta_i^q = \sum_{k=1}^K t_{ik}^q \mu_k^q.$$

Nous appellerons cet algorithme Geo-EM.

Résultats

Les tests sur simulations rapportés dans [M2] montrent que l'algorithme Geo-EM présente de bonnes performances comparé à d'autres algorithmes. Il présente des taux de mauvaise classification plus faible que l'ICM de Besag et que la méthode par champ moyen simulé (algorithme SF-EM, Celeux, Peyrard & Forbes, 2003) pour le modèle de mélange de gaussiennes spatialisées considéré ci-dessus. De façon plus inattendue peut-être, il est également meilleur pour le modèle complet dans lequel $\mathbf{Z}$ est un modèle de Potts à K couleurs et $\mathbf{Y} \mid \mathbf{Z}$ présente des corrélations spatiales. Ainsi par exemple pour un modèle de Potts à 2 couleurs, on observe un taux de mauvaise classification de 7.1% à 18.4% (selon les paramètres choisis) pour Geo-EM, et un taux de 10.4% à 26.2 % pour SF-EM.

L'algorithme a également été testé sur des données de teneur en métaux lourds échantillonnées dans une région du Jura Suisse (Atteia, Dubois & Webster, 1994). Sur ce jeu de données, le sous-sol est connu et classé en 5 types. Des analyses antérieures ont montré que le facteur sous-sol est significatif. L'algorithme Geo-EM parvient à retrouver la distinction principale entre la couche argovienne et les autres couches. L'absence de dépendance dans la couche cachée entraîne que l'algorithme Geo-EM fournit des cartes aux frontières moins régulières que l'algorithme SF-EM.

Perspectives

Ce travail a montré l'intérêt qu'il y avait à renverser la logique habituelle concernant la modélisation des dépendances spatiales, notamment pour le modèle complet dans lequel $\mathbf{Z}$ est distribué selon un modèle de Potts. En effet, nous avons montré sur des simulations que *pour ce modèle* et dans un objectif de restauration du champ caché, il est plus efficace de modéliser les dépendances spatiales sur $\mathbf{Y} \mid \mathbf{Z}$ plutôt que sur $\mathbf{Z}$. Il reste maintenant à poursuivre ce travail

d'une part en cherchant des situations pour lesquelles l'inverse se produirait et d'autre part en proposant un algorithme pour le modèle complet, qui tienne compte des corrélations à la fois sur $\mathbf{Z}$ et sur $\mathbf{Y} \mid \mathbf{Z}$.

3.3 Arbres de régression

Présentation du problème

Ce travail fait l'objet d'une collaboration avec Liliane Bel (Université d'Orsay, puis INA-PG) et d'Avner Bar-Hen (INAPG), sur des données de paléo-écologie de l'Institut des Sciences de l'Evolution (CNRS, Montpellier) décrites dans Laurent et. al. (2004). On notera $X^1, \ldots, X^p$ les variables explicatives et Y la variable dépendante. Un arbre de régression ou de classification (en anglais, Classification And Regression Tree, CART (Breiman et al., 1984)) est un modèle, noté T, qui se construit à partir de données $(Y, X^1, \ldots, X^p)$ de la façon suivante : un sous-ensemble des données, noté t, que l'on appellera une feuille de l'arbre T est divisé en deux sous-ensembles (deux sous-feuilles) qui minimisent un critère d'hétérogénéité calculé sur les sous-feuilles. Chaque division est basée sur une seule variable et un seul seuil ; ainsi par exemple le seuil x_t^i sépare la feuille t en deux sous-feuilles, l'une correspondant à $X^i \leq x_t^i$ et l'autre à l'ensemble complémentaire. Nous ne considérons ici que le cas où Y est une variable catégorielle ou ordinale. Dans ce cas, notons $p_t(j)$ la proportion de la classe j de la variable Y dans la feuille t. Les deux indices d'hétérogénéité les plus fréquemment utilisés sont l'indice de Gini, $G_t = \sum_j \sum_{k \neq j} p_t(j) p_t(k) = 1 - \sum_j p_t^2(j)$ et l'indice d'entropie, $E_t = \sum_j p_t(j) \ln p_t(j)$, (avec $x \ln x = 0$ lorsque $x = 0$). Les deux indices sont égaux à 0 lorsqu'il n'y a qu'une seule classe dans une feuille t et sont maximums lorsque toutes les classes sont présentes en proportions égales. L'indice de Gini pouvant s'interpréter en termes de variance, il sera préféré à l'indice d'entropie qui ne sera plus considéré par la suite. Le principe de la division d'une feuille t en deux sous-feuilles t_- et t_+ est de rechercher le seuil x_t^i de la variable X^i tel que $G_t - (n_{t_-} G_{t_-} + n_{t_+} G_{t_+})$ est positif et maximal. La procédure est ensuite répétée jusqu'à ce qu'on ne puisse plus diviser de feuilles. Chaque feuille t est ensuite affectée à la classe modale de Y dans t. En règle générale, cette première étape mène à un arbre surparamétré dont l'erreur de prédiction $R(T) = P\{T(X^1, \ldots, X^p) \neq Y\}$ est élevée. L'arbre est donc dans un deuxième temps élagué pour aboutir à un sous-arbre T' dont l'erreur de prédiction est inférieure à celle de T. L'algorithme d'élagage ((Breiman et al., 1984) fait appel à la validation croisée pour rechercher cet arbre T'. Tout au long de la procédure, il faut donc estimer les proportions, $p^t(j)$, l'indice de Gini, G_t, et l'erreur de prédiction $R(T)$.

Les données sont habituellement considérées comme indépendantes dans l'algorithme CART. Dans ce cas, les grandeurs ci-dessus sont estimées par les fréquences empiriques : $\hat{p}_t(j) = n_{tj}/n_t$, où n_t est le nombre de données dans la feuille t et n_{tj} le nombre de celles-ci dans la classe j ; le

critère à minimiser est $\hat{G}_t - (n_{t_-}\hat{G}_{t_-} + n_{t_+}\hat{G}_{t_+})$ avec $\hat{G}_t = 1 - \sum_j \hat{p}_t^2(j)$ et le risque empirique est

$$\hat{R}(T) = \frac{1}{n}\sum_{\alpha=1}^{n} \mathbf{I}\{T(X_\alpha^1, \ldots, X_\alpha^p) \neq Y_\alpha\}.$$

Cependant, de même qu'à la section précédente, les données environnementales et/ou écologiques — géoréférencées en des points $(s_1, \ldots, s_n)$ d'un domaine D — présentent le plus souvent des corrélations spatiales. Il faut donc considérer que les données sont issues de champs aléatoires dont les dépendances doivent être modélisées et proposer un algorithme qui prenne en compte de façon implicite ou explicite l'aspect spatialisé des données. Nous verrons également à la section 3.4 comment un arbre de régression issu de CART peut servir à améliorer l'interpolation par krigeage.

Méthodes avec pondération

Cette méthode consiste à affecter à chaque donnée $(X_\alpha^1, \ldots, X_\alpha^p)$ localisée en s_α un poids w_α, $\alpha = 1, \ldots, n$, de manière à ce que les données spatialement groupées aient des poids plus faibles que les données isolées [24]. L'idée générale est en effet que des données spatialement proches sont partiellement redondantes en raison des corrélations spatiales. Au lieu des estimateurs empiriques habituels ci-dessus, on construira donc les estimateurs pondérés

$$\widehat{p}(i \mid t) = \sum_{\alpha : s_\alpha \in t} w_\alpha \mathbf{I}\{Y(s_\alpha) = i\}/w_t \quad \text{et} \quad \widehat{R}(T) = \sum_{\alpha=1}^{n} w_\alpha \mathbf{I}[T\{X^1(s_\alpha), \ldots, X^p(s_\alpha)\} \neq Y(s_\alpha)]/w_t, \tag{7}$$

avec $w_t = \sum_{\alpha : s_\alpha \in t} w_\alpha$ et la condition $\sum_\alpha w_\alpha = 1$. Trois pondérations ont été envisagées :

1. La première consiste à utiliser la tesselation de Voronoï définie par les lieux des échantillons. On rappelle que la cellule de Voronoï autour du point s_α est l'ensemble des points de $\mathcal{D}$ plus proche de s_α que de tout autre point d'échantillonnage. A des données fortement groupées seront associées de petites cellules, tandis que des données isolées auront des cellules associées de plus grande taille. On pose que les poids w_α sont proportionnels à la surface de la cellule de Voronoï associée. Cette approche est intuitivement simple, facile à implémenter, mais présente l'inconvénient d'avoir des effets de bords difficiles à gérer, les cellules extérieures ayant des surfaces dépendant très fortement de la façon dont sont définies les limites du domaine. Comme nous le verrons à la section 5.2, l'approche par cellule de Voronoï a également été utilisée dans [15] puis poursuivie dans Magnussen, [9] pour faire de la classification non supervisée de processus ponctuels.
2. Pour la seconde pondération, une grille est superposée sur le domaine d'étude et les poids w_α sont inversement proportionnels au nombre de données dans la même maille de la grille que w_α. Ici aussi, une densité localement élevée de points d'échantillonnage entraîne des poids faibles.

3. À l'inverse des deux premières pondérations qui sont totalement non paramétriques, la troisième se propose d'utiliser les corrélations spatiales existant dans les données afin de déterminer des poids « optimaux ». Ceux-ci sont les poids w_α du krigeage de la moyenne régionale de Y (lorsque Y est une variable ordinale) sur le domaine D. L'intérêt de cette approche est double. D'une part les poids ne dépendent pas seulement de la géométrie de l'échantillonnage mais également de la fonction de covariance de la variable considérée. D'autre part, les poids w_α ont un caractère d'optimalité pour un critère (l'erreur d'estimation de la moyenne de Y sur D) qui est assez proche des critères utilisés dans CART. Nous verrons au paragraphe suivant comment la logique de cette approche peut être généralisée pour proposer un algorithme CART spatialisé. Dans certaines circonstances, les poids de krigeage peuvent être négatifs. Or, on ne peut pas utiliser de poids négatifs dans l'algorithme CART, ce qui nous a amené à utiliser les poids du système de krigeage contraint :

$$\min_W \operatorname{var}(W^T Y - Y_D), \quad \text{avec} \quad \mathbf{1}^T W = 1 \quad \text{et} \quad w_\alpha \geq 0,$$

où $Y = (Y_1, \ldots, Y_n)^T$.

Estimation spatialisée

La seconde approche, présentée dans [24], consiste à généraliser la logique de l'estimation spatiale pour toutes les grandeurs intervenant dans l'algorithme. Pour cela il est nécessaire de préciser un modèle pour les dépendances spatiales dans le cadre d'une approche populationelle des arbres de régression (Ripley, 1996, chap. 7). Le modèle général adopté est qu'il existe une fonction R, catégorielle ou ordinale, telle que

$$Y(s) = R(\beta^1(s), \ldots, \beta^p(s)).$$

R est le modèle de régression que l'on cherche à retrouver. On observe les fonctions $X^i(s) = \beta^i(s) + \epsilon^i(s)$. Les fonctions $\beta^i(s)$ sont des fonctions déterministes et les $\epsilon^i(s)$ sont des résidus gaussiens corrélés. Dans ce cadre, un arbre T est une estimation de R.

Soit t une feuille. La proportion théorique $p_t(j)$ est estimée par krigeage ordinaire de la moyenne de l'indicatrice $I_{tj}(s) = \mathbf{I}\{Y(s_\alpha) = j \mid s_\alpha \in t\}$ dans le domaine $\mathcal{D}_t$ correspondant à la feuille t. On a donc

$$\hat{p}(j \mid t) = \sum_{\alpha \in t} \lambda_\alpha I_{tj}(s_\alpha)$$

où (λ_α) est la solution du système à n_t équations pour les échantillons α dans la feuille t :

$$\sum_{\beta \in t} \lambda_\beta C_j(s_\alpha, s_\beta) = \frac{1}{|D_t|} \int_{D_t} C_j(s_\alpha, s)\, ds, \tag{8}$$

sous la contrainte $\sum_\alpha \lambda_\alpha = 1$. Dans les équations ci-dessus, $C_j(s, s')$ est la fonction de covariance de l'indicatrice $I_{tj}(s)$, et D_t est le domaine de D correspondant à t. L'estimateur de l'indice de Gini est ensuite calculé par :

$$\hat{G}_t = 1 - \sum_i \hat{p}(i \mid t)^2. \tag{9}$$

Dans ce cadre, on peut montrer que sous des conditions assez classiques, $\hat{G} \to G$ lorsque $D \to \mathbf{R}^2$ (si la densité d'échantillonnage reste finie lorsque $D \to \mathbf{R}^2$). On estime de la même façon le critère d'hétérogénéité et le risque empirique.

Résultats

Les différentes approches ont été testées par simulation. Il en ressort que lorsque les données sont spatialement corrélées et que les échantillons sont très irrégulièrement placés (en particulier, en présence de clusters d'échantillons), il est toujours préférable de chercher à tenir compte des corrélations spatiales que de les ignorer. En effet, en l'absence de corrélation spatiale ou en présence d'un échantillonnage très régulier on ne dégrade que très peu les performances de CART en utilisant une des méthodes présentées ci-dessus ; par contre, on améliore toujours les performances dans le cas contraire, et parfois de façon très importante. Les méthodes avec pondération sont plus rapides et plus robustes (il n'est pas nécessaire de réestimer les variogrammes et les poids dans chaque feuille). Les méthodes par cellule de Voronoï et par krigeage de la moyenne donnent systématiquement des poids plus élevés pour les échantillons situés au bord du domaine que la méthode par grille régulière. Celle-ci est par contre assez sensible à la taille des mailles et à l'emplacement de l'origine. Aussi il est judicieux de faire varier ces deux paramètres puis de considérer la moyenne des poids obtenus. Lors de l'exploration des données, il peut se révéler intéressant d'appliquer CART avec les différentes méthodes de spatialisation et de comparer les résultats obtenus. Une compréhension supplémentaire des données peut naître de cette comparaison.

3.4 Interpolation et classification

Le krigeage avec dérive externe de classes

La prédiction d'une variable régionalisée en un point non échantillonné (le krigeage) peut être sensiblement améliorée si l'on dispose d'une classification du domaine d'étude en K sous-domaines. Cette approche a fait l'objet du stage de DEA de I. Navarro-Sanchez (M9) et du stage de fin d'études de l'Ecole des Mines de Paris de C. Royer (M7) que j'ai encadrés. Elle est décrite dans [14] pour l'analyse de données de pluviométrie en Suisse, puis dans [33] et [10] pour l'interpolation locale de la température servant de variables d'entrée pour les modèles agronomiques de plantes.

Dans [14], on utilise une classification non supervisée des données, telle que décrite à la section 3.2. Dans [33], on utilise une classification obtenue par un algorithme CART non spatialisé à

partir des fréquences d'occupation des sols décrites dans la base CORINE (IFEN, 1996). Dans [10], on quantifie comment le gain de précision sur l'interpolation locale de la température se répercute sur les variables agronomiques de sortie du modèle de plante considéré.

La moyenne locale $m(s)$ est modélisée par des effets fixes $e_k,\ k = 1, \ldots, K$ correspondant aux K classes. Les valeurs e_k sont inconnues et doivent être estimées. En un point s, on a $m(s) = \sum_k e_k \mathbf{1}_k(s)$, où $\mathbf{1}_k(s)$ est la fonction indicatrice de la classe k en s, valant 1 si s est dans la classe k, et 0 sinon. La variable $Z(s)$ s'écrit

$$Z(s) = \sum_{k=1}^{K} e_k \mathbf{1}_k(s) + \epsilon(s) \tag{10}$$

où $\epsilon(s)$ est un résidu d'espérance nulle. Le krigeage avec dérive externe de classe (KDEC) en s_0 sera alors le prédicteur linéaire, optimal, sans biais, correspondant au modèle (10) : $Z^*(s_0) = \sum_i \lambda_i Z(s_i)$, i étant un indice parcourant les données. On montre facilement que les pondérateurs $(\lambda_i)_i$ sont solutions du systèmes à $(n + K)$ équations et $(n + K)$ inconnues

$$\begin{cases} \sum_{j=1}^{n} \lambda_j C(s_i - s_j) - \sum_{k=1}^{p} \mu_k \mathbf{1}_k(s_i) = C(s_j - s_0) & \text{pour } i = 1, \ldots, n \\ \\ \sum_{i=1}^{n} \lambda_i \mathbf{1}_k(s_i) = \mathbf{1}_k(s_0) & \text{pour } k = 1, \ldots, K \end{cases} \tag{11}$$

Il y a K contraintes de non-biais qui doivent être simultanément vérifiées. Notons que la contrainte globale $\sum_i \lambda_i = 1$ est automatiquement vérifiée. Lorsque les classes sont connues aux points de mesure, ce qui est le cas ici, une estimation non biaisée de la fonction de covariance se fait en n'utilisant que les couples de points appartenant à la même classe. Lorsque les classes sont construites par une classification spatiale des données (voir section 3.2), les fonctions de covariance sont estimées simultanément à la classification.

Sur des données de températures (maximum journalier sur 116 stations météorologiques du Sud-Est de la France), on a pu montrer ([33], [10]) que l'approche par KDEC corrigeait le biais dans les 4 classes (sur 6) pour lesquelles le KO ordinaire était fortement biaisé (de 0.54°C à 2.57°C), sans introduire de biais dans les deux classes restantes ni créer de biais global. En termes agronomiques, un biais systématique de 1°C pour tous les jours de la saison se traduit par un écart de rendement de l'ordre de 1.5 à 2 t/ha et et par une date de moisson décalée de 12 jours environ.

Lorsque les classes ne sont pas connues, mais seulement prédites, on construit un prédicteur au point s_0 intégrant la probabilité $p_k(s)$ d'appartenir à chaque classe k, [14], [22] :

$$Z^*(s_0) = \sum_k Z^{k,*}(s_0) p_k(s_0)$$

où $Z^{k,*}(s_0)$ est l'estimateur de krigeage en s_0 dans la classe k. Notons $\mathbf{0}$ et $\mathbf{1}$ des vecteurs de 0 et de 1 de longueur appropriée, $\boldsymbol{\lambda}_k$ le vecteur des poids des éléments dans la classe k, $\mathbf{C}_{kl}$ la matrice

de covariance construite à partir des points des classes k et l et $\mathbf{C}_{0k}$ le vecteur des covariances entre le point s_0 et les points dans la classe k. Les poids sont solution du système :

$$\begin{pmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1K} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{C}_{K1} & \cdots & \mathbf{C}_{KK} & \mathbf{0} & \cdots & \mathbf{1} \\ \mathbf{1}^t & \cdots & \mathbf{0}^t & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{0}^t & \cdots & \mathbf{1}^t & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_K \\ -\mu_1 \\ \vdots \\ -\mu_K \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{01} \\ \vdots \\ \mathbf{C}_{0K} \\ p_1 \\ \vdots \\ p_K \end{pmatrix}, \tag{12}$$

Le système (11) est une restriction de (12) dans le cas où $p_k \in \{0, 1\}$. Afin de mettre en œuvre (12), il faut une estimation de $p_k(s)$ pour tout s, ce qui peut se faire par plusieurs techniques. Sur un jeu de données de pluviométrie en Suisse, on montre dans [22] par validation croisée qu'une classification en probabilité mène à des prédicteurs avec un écart quadratique moyen plus faible que par une prédiction ne prenant pas en compte la classification.

Pistes de travail

Il y a plusieurs manières de prendre en compte une carte thématique dans le but d'améliorer la prédicition spatiale. Le modèle (10) n'est qu'un modèle possible parmi d'autres. Il y aurait donc un travail de sélection de modèles à réaliser. Des pistes basées sur le variogramme avaient été explorées dans le travail de fin d'étude [M6]. Ce travail n'a pas été repris ensuite par manque de temps, et parce qu'il était moins prioritaire que d'autres travaux.

Cette thématique resurgit cependant dans le cadre des ZCAs. En effet, nous avons vu à la section 3.1 qu'en présence de ZCA il est nécessaire de réaliser l'estimation du variogramme en excluant les couples de points situés de part et d'autres d'une ZCA. La même remarque pourrait s'appliquer pour la prédiction spatiale : lorsqu'on réalise l'interpolation en un point situé à proximité d'une ZCA, il pourrait s'avérer plus performant d'exclure du voisinage considéré pour la prédiction spatiale les points de mesure situé de l'autre côté d'une ZCA. Cette idée reste à tester.

4 Champs aléatoires gaussiens dissymétriques

Problématique générale

L'hypothèse gaussienne pour les champs aléatoires se comprend pour de multiples raisons : la densité multivariée gaussienne est entièrement caractérisée par ses deux premiers moments ; elle est stable pour l'addition, les transformations linéaires et le conditionnement ; l'espérance conditionnelle est linéaire, etc.

Cependant, pour un grand nombre d'applications environnementales les variables étudiées (teneur, cumul de pluie, vitesse du vent...) présentent une distribution dissymétrique. C'est le cas notamment des variables positives mentionnées ci-dessus, nécessairement bornées inférieurement, dont l'histogramme présente des valeurs élevées plus fréquentes que sous l'hypothèse gaussienne. L'approche classique pour modéliser ce type de distribution consiste à effectuer une transformation des variables. Celle-ci peut être paramétrique (transformation de Box-Cox) ou non paramétrique (voir par exemple Wackernagel, 2003, Chapitre 33).

Une autre approche consiste à rechercher des modèles de champs aléatoires dont la distribution est par construction dissymétrique. C'est l'approche suivie en collaboration avec Philippe Naveau (LSCE, CNRS Gif-sur-Yvette). Dans [28] et [5], nous proposons une nouvelle classe de champs gaussiens dissymétriques, les *Skew-Normal Random Fields*, (SNRF).

Cette classe est issue des distributions skew-normal multivariées généralisées (les CSN), qui sont elles-mêmes une extension des distributions normales. Elles sont définies comme le produit d'une densité gaussienne par une fonction de répartition gaussienne. Elles sont donc caractérisées par des paramètres de moyenne et de variance-covariance ainsi que par un paramètre supplémentaire qui porte l'essentiel de la caractérisation de l'asymétrie. Les distributions CSN présentent une asymétrie tout en conservant les propriétés les plus intéressantes des distributions gaussiennes (Azzalini, 2005 ; Genton, 2004) : stabilité pour l'addition, les combinaisons linéaires et le conditionnement.

Les champs aléatoires SNRF

La densité de probabilité d'un n-vecteur CSN $\mathbf{Y}$, notée $\mathrm{CSN}_{n,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Delta})$, est :

$$c_m \; \phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \; \Phi_m(\mathbf{D}^t(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Delta}), \text{ avec } c_m^{-1} = \Phi_m(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Delta} + \mathbf{D}^t \boldsymbol{\Sigma} \mathbf{D}), \tag{13}$$

où $\boldsymbol{\mu} \in R^n$, $\boldsymbol{\nu} \in R^m$, $\boldsymbol{\Sigma} \in R^{n \times n}$ et $\boldsymbol{\Delta} \in R^{m \times m}$ sont deux matrices de covariance, $\mathbf{D} \in R^{n \times m}$, $\phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $\Phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ sont respectivement la densité et la fonction de répartition de la distribution normale d'espérance $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$, et où $\mathbf{D}^t$ est la transposée de la matrice $\mathbf{D}$. Si $\mathbf{D} = \mathbf{0}$, la densité (13) se réduit à la densité normale multivariée. Lorsque $m = 1$, on retrouve la densité Skew-Normal d'Azzalini (2005) : la variable $\mathbf{Y}$ suit une distribution $\mathrm{CSN}_{n,1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, 0, 1)$ où $\boldsymbol{\alpha}$ est un vecteur de longueur n. Cette définition offre une vaste possibilité de modèles, selon les choix que l'on fait pour $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\Delta}$, $\boldsymbol{\Sigma}$ et $\mathbf{D}$.

La distribution $\mathrm{CSN}_{n,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Delta})$ peut se construire de la façon suivante : soit $\mathbf{U}$ un vecteur gaussien de dimension m et soit le vecteur gaussien augmenté $(\mathbf{U}^t, \mathbf{Z}^t)^t$ tel que

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{Z} \end{pmatrix} \stackrel{\mathrm{d}}{=} N_{m+n}\left(\begin{pmatrix} \boldsymbol{\nu} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Delta} + \mathbf{D}^t\boldsymbol{\Sigma}\mathbf{D} & -\mathbf{D}^t\boldsymbol{\Sigma} \\ -\boldsymbol{\Sigma}\mathbf{D} & \boldsymbol{\Sigma} \end{pmatrix}\right), \tag{14}$$

où $\stackrel{\mathrm{d}}{=}$ désigne l'égalité en distribution. Alors, on peut montrer que conditionnellement à $\mathbf{U} \leq \mathbf{0}$, le vecteur $\boldsymbol{\mu} + [\mathbf{Z}|\mathbf{U} \leq \mathbf{0}]$ est distribué selon une $\mathrm{CSN}_{n,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Delta})$, où $\mathbf{U} \leq \mathbf{0}$ signifie que $U_i \leq 0$, pour tout $i = 1, \ldots, m$. Cette propriété entraîne qu'il est aisé de construire un algorithme pour simuler un vecteur CSN.

Nous définissons un champ SNRF $\{Y(s)\}$ par

$$Y(s) \stackrel{\mathrm{d}}{=} \mu + [Z(s) \mid \mathbf{U} \leq \mathbf{0}].$$

Ainsi, pour tout n et tout vecteur $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))^t$ de $Z(\cdot)$, on a

$$\mathbf{Y} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + [\mathbf{Z} \mid \mathbf{U} \leq \mathbf{0}], \tag{15}$$

où $\boldsymbol{\mu}^t = \mu\ (1, \ldots, 1)^t$ avec $\mu \in R$, $\mathbf{U} \stackrel{\mathrm{d}}{=} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Delta} + \mathbf{D}^t\boldsymbol{\Sigma}\mathbf{D})$ et $\mathbf{Z}$ est distribué selon (14). En pratique, on observe uniquement un échantillon $(Y(s_1), ..., Y(s_n))^t$, mais on n'observe ni $\mathbf{U}$ ni $\mathbf{Z}$.

Le vecteur $\mathbf{Y}$ peut aussi s'exprimer comme la somme de deux processus indépendants. En effet, soit

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \stackrel{\mathrm{d}}{=} N_{m+n}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Delta} + \mathbf{D}^t\boldsymbol{\Sigma}\mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{pmatrix}\right),$$

où $\mathbf{I}_n$ est la matrice identité de taille n. Alors on a $\mathbf{Z} = -\mathbf{F}\mathbf{U} + \mathbf{G}^{\mathbf{1/2}}\mathbf{V}$ avec $\mathbf{F} = \boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{\Delta} + \mathbf{D}^t\boldsymbol{\Sigma}\mathbf{D})^{-1}$ et $\mathbf{G} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{\Delta} + \mathbf{D}^t\boldsymbol{\Sigma}\mathbf{D})^{-1}\mathbf{D}^t\boldsymbol{\Sigma}$. Le vecteur $\mathbf{U}$ est un vecteur gaussien d'espérance nulle et de matrice de covariance $\boldsymbol{\Sigma}$ et le couple bivariable $(\mathbf{U}^t, \mathbf{Z}^t)^t$ vérifie (14). En conséquence, (15) et l'indépendance de $\mathbf{U}$ et $\mathbf{V}$ nous permettent d'écrire

$$\mathbf{Y} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} - \mathbf{F}[\mathbf{U} \mid \mathbf{U} \leq \mathbf{0}] + \mathbf{G}^{1/2}\mathbf{V}. \tag{16}$$

Cette décomposition, particulièrement utile pour calculer le second moment de $Y(\cdot)$, fait apparaître qu'un champ SNRF est la somme d'un champ gaussien et d'un champ gaussien tronqué indépendant.

Afin de réduire le nombre de paramètres des champs SNRF, on pose que $\mathbf{D} = \delta\mathbf{A}$, où $\delta \in R$ est un paramètre unique contrôlant l'asymétrie et $\mathbf{A}$ est une matrice non nulle dont les éléments sont supposés connus et indépendants de δ. $\mathbf{A}$ modélise le lien entre les coordonnées de $\mathbf{Y}$ et de $(\mathbf{U}, \mathbf{V})$. Lorsque $\delta = 0$, $\mathbf{Y}$ est indépendant de $\mathbf{U}$, i.e. $\mathbf{Y}$ est un vecteur gaussien d'espérance $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$. Lorsque δ^2 tend vers l'infini, la matrice qui multiplie $\mathbf{V}$ tend vers la matrice nulle et $\mathbf{Y}$ tend vers un vecteur gaussien tronqué.

Lorsque $m = 1$, on a montré dans [28] que l'asymétrie introduite par $[U \mid U \leq 0]$ se répartit sur toutes les coordonnées du vecteur $\mathbf{Z}$ pour un poids total égal à 1, les deux situations extrêmes étant que toutes les coordonnées présentent une asymétrie de poids $1/n$ chacune, ou que toute l'asymétrie soit concentrée sur une seule coordonnée.

Le modèle proposé dans [5] consiste à poser $m = n$, $\boldsymbol{\mu} = \mu\mathbf{1}$,$\boldsymbol{\Delta} = \boldsymbol{\Sigma}$ et $\mathbf{A} = \mathbf{I}_n$. La matrice $\boldsymbol{\Sigma}$ est une matrice de covariance construite à partir d'une fonction de covariance $c(h)$ et des points $(s_1, \ldots, s_n)$ de D. Ces choix permettent de réduire les paramètres tout en conservant une structure spatiale forte et un degré d'asymétrie important. Avec ces choix, le modèle devient

$$\mathbf{Y} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \frac{\delta}{\sqrt{1+\delta^2}}[\mathbf{U} \mid \mathbf{U} \geq \mathbf{0}] + \frac{1}{\sqrt{1+\delta^2}}\boldsymbol{\Sigma}^{1/2}\mathbf{V}. \tag{17}$$

Notons que le côté de l'asymétrie est déterminé par le signe de δ.

Estimation de la fonction de covariance

Il est possible de calculer les deux premiers moments de ce modèle :

$$E[Y_i] = \mu + \frac{\delta^*\sigma}{\sqrt{2\pi}} \sum_{k=1}^{n} R_{ik} \frac{\Phi_{n-1}(\mathbf{0}; \mathbf{0}, \mathbf{R}_k)}{\Phi_n(\mathbf{0}; \mathbf{0}, \mathbf{R})}, \tag{18}$$

avec $\delta^* = \delta(1+\delta^2)^{-1/2}$ et où $\mathbf{R}$ est la matrice de corrélation correspondante à $\boldsymbol{\Sigma}$. Les probabilités $\Phi_n(\mathbf{0}; \mathbf{0}, \mathbf{R})$ sont calculées à l'aide du package `mvtnorm` de `R` (Genz, 1992).

La fonction de covariance $c(h)$ est estimée par le variogramme expérimental :

$$\hat{\gamma}_Y(h) = \frac{1}{2N(h)} \sum_{\{(i,j):|s_i - s_j| \approx |h|\}} \{Y(s_i) - Y(s_j)\}^2 \tag{19}$$

où $N(h)$ est le nombre de paires $\{s_i, s_j\}$ telles que la distance $|s_i - s_j|$ est approximativement égale à $|h|$. En notant $\gamma_{ij} = \sigma^2(1 - R_{ij})$, on montre alors que

$$E[\hat{\gamma}_Y(h)] = \left[\frac{1}{N(h)} \sum_{\{(i,j):|s_i - s_j| \approx |h|\}} \gamma_{ij}\right] + \frac{\delta^{*2}\sigma^2}{2\pi}\Gamma(h), \tag{20}$$

avec

$$\Gamma(h) = \frac{1}{2N(h)} \sum_{\{(i,j):|s_i - s_j| \approx |h|\}} (\Psi_{ii} + \Psi_{jj} - 2\Psi_{ij}), \tag{21}$$

et

$$\Psi_{ij} = \sum_{k=1}^{n} R_{ik} \sum_{l \neq k} \frac{R_{jl} - R_{kl}R_{jk}}{\sqrt{1 - R_{kl}^2}} \frac{\Phi_{n-2}(\mathbf{0}; \mathbf{0}, \mathbf{R}_{kl})}{\Phi_n(\mathbf{0}; \mathbf{0}, \mathbf{R})}, \tag{22}$$

où $\mathbf{R}_{kl}$ est la matrice de corrélation conditionnelle lorsque les variables aléatoires sont fixées en s_k et s_l. Ces équations permettent de proposer un algorithme basé sur la méthode des moments pour estimer les paramètres du champ SNG.

1. Calculer le variogramme expérimental $\hat{\gamma}_Y(h)$ à partir des données brutes.
2. Estimer le paramètre de porté, $\hat{a}$, en ajustant un modèle de variogramme paramétrique, p. ex. le modèle exponentiel $b(1-\exp\{-|h|/a\})$, à $\hat{\gamma}_Y(h)$. Il faut noter que bien que le paramètre b soit également estimé, son estimation ne sera pas utilisée dans la suite. On note $\gamma(h;\hat{a}) = 1-\exp\{-|h|/\hat{a}\}$ et $\rho(h;\hat{a}) = \exp\{-|h|/\hat{a}\}$.
3. Calculer la matrice de corrélation $\mathbf{R}$ à l'aide du modèle $\rho(h;\hat{a})$. Calculer ensuite Ψ_{ij} et $\Gamma(h)$ en utilisant (21) et (22).
4. Estimer le couple $(\hat{\sigma}^2, \hat{\delta}^{*2})$ qui ajuste $\sigma^2\{1-\rho(h;\hat{a})+\delta^{*2}\hat{\Gamma}(h)/(2\pi)\}$ au variogramme expérimental $\hat{\gamma}_Y(h)$.
5. Déterminer le signe de $\hat{\delta}^*$ et calculer

$$\hat{\mu} = \bar{Y} - \text{sign}\{\hat{\delta}^*\}\frac{\hat{\sigma}|\hat{\delta}^*|}{\sqrt{2\pi}}\sum_{k=1}^{n} R_{ik}\frac{\Phi_{n-1}(\mathbf{0};\mathbf{0},\mathbf{R}_k)}{\Phi_n(\mathbf{0};\mathbf{0},\mathbf{R})}.$$

Cet algorithme a été testé et validé sur des données synthétiques dans [5], mais ce modèle présente un inconvénient majeur. Les probabilités de type $\Phi_n(\mathbf{0};\mathbf{0},\mathbf{R})$ ne peuvent pas être calculées analytiquement. On doit recourir à des approximations par intégration de Monte-Carlo. Cela entraîne d'une part des calculs assez longs, et d'autre part des résultats approximatifs. C'est pourquoi on privilégiera à l'avenir une modélisation alternative.

Nouvelle modélisation

L'objectif est de trouver une paramétrisation qui soit à la fois suffisamment riche pour modéliser les dépendances spatiales et bien choisie de façon à éviter le calcul de la fonction de répartition d'un vecteur gaussien de grande dimension. Ce travail en cours fait partie de la thèse de Cédric Flécher co-encadrée avec Philippe Naveau [T1]. On a démontré dans [M1] que tout modèle de la forme

$$\mathbf{Y} \overset{\text{d}}{=} \boldsymbol{\mu} + \mathbf{P}[\mathbf{U} \mid \mathbf{U} \geq \mathbf{0}] + \mathbf{Q}\mathbf{V}, \tag{23}$$

est un vecteur CSN. Dans (23), $\mathbf{U}$ est un vecteur gaussien de matrice de covariance $\boldsymbol{\Gamma}$, $\mathbf{V}$ est un vecteur de même dimension que $\mathbf{U}$ de V.A. gaussiennes $N(0,1)$ indépendantes, $\mathbf{P}$ et $\mathbf{Q}$ sont des matrices quelconques.

Nos choix se dirigent vers $\boldsymbol{\Gamma} = \mathbf{I}_n$, $\mathbf{P} = \sigma_P\boldsymbol{\Sigma}^{1/2}$ et $\mathbf{Q} = \sigma_Q\boldsymbol{\Sigma}^{1/2}$, où $\boldsymbol{\Sigma}$ est une matrice de covariance issue d'une fonction de covariance $c(h)$. Choisir $\boldsymbol{\Gamma} = \mathbf{I}_n$ simplifie considérablement les calculs, puisque dans ce cas, $\Phi_n(\mathbf{0};\mathbf{0},\mathbf{I}_n) = 2^{-n}$. Des résultats préliminaires ont montré que ce choix permet à la fois une modélisation pertinente des corrélations spatiales et l'estimation de la fonction de covariance $c(h)$.

Perspectives

Nous avons exhibé une classe de champs aléatoires skew-normal pour lesquels l'inférence statistique des paramètres semble possible. Il reste maintenant à mettre complètement en œuvre ce modèle, c'est-à-dire (i) construire des estimateurs efficaces et établir leurs propriétés ; (ii) établir les équations pour la prédiction spatiale ; (iii) proposer, pour cette classe de modèles, un test permettant de tester $\delta = 0$ contre $\delta \neq 0$.

Sur un plan plus appliqué, l'objectif de la thèse [T1] consiste à utiliser le modèle (23) pour modéliser des séries temporelles de données climatiques afin de proposer des générateurs de séries météorologiques. Plus de détails sur ce projet de thèse figurent en Section 6.

5 Processus de points dans le plan

5.1 Introduction

Présentation des questions de recherche

Alors que dans les premiers chapitres les points de mesure étaient fixés et n'étaient que des supports neutres de l'information, on considère ici que c'est l'ensemble des positions des points qui porte l'essentiel de l'information. On parle de processus de points, ou de processus ponctuels.

Lorsque le scientifique ou le statisticien analyse des données qui se présentent sous la forme d'un processus ponctuel, l'une des premières questions que celui-ci se pose est de tester l'hypothèse d'une répartition obéissant au hasard complet, dans laquelle aucun structure spatiale n'apparaît, contre des alternatives plus agrégées ou au contraire plus régulières. Les outils pour réaliser cette analyse, les fonctions K de Ripley par exemple, existent depuis une trentaine d'année. La procédure de test proprement dite repose sur des techniques de Monte-Carlo, voir p. ex. Diggle (1983) pour un exposé détaillé de ces techniques. Ces analyses se font le plus souvent à l'échelle du domaine d'étude et nécessitent donc une hypothèse de stationnarité, rarement vérifiée. Il convient donc de proposer des outils statistiques permettant ou bien de caractériser cette non-stationnarité, ou bien de s'en affranchir en travaillant localement, c'est-à-dire en faisant une hypothèse plus faible de stationnarité locale.

Les trois paragraphes de cette section s'inscrivent dans cette démarche. La section 5.2 considère une modélisation très particulière de la non-stationnarité. On suppose que le processus de points observé provient de la superposition de deux processus de points, dont l'un se manifeste sur un support contenu dans le premier. Il s'agit alors de proposer une méthode permettant d'estimer ce support.

La section 5.3 propose ensuite de tester l'indépendance entre deux processus de points observés dans un même domaine en filtrant les corrélations à grande échelle provenant de dépendances communes à des covariables externes.

Enfin la section 5.4 étudie les conditions que doit vérifier un processus de points pour pouvoir être transformé en un processus stationnaire à l'ordre 2. Ce travail ouvre la voie aux techniques de tranformation non paramétrique des coordonnées de l'espace afin de se ramener à processus de points stationnaire dans le nouvel espace. Cette transformation permet alors d'utiliser tout l'arsenal des méthodes développées dans le cadre stationnaire sur l'espace transformé.

Définitions et notations

On considérera un domaine D, en général dans $\mathbf{R}^2$, bien que les différents travaux présentés ci-dessous s'étendent sans difficulté aux dimensions supérieures. Un processus de points de $\mathbf{R}^2$, noté $\mathbf{x}$ ou $\mathbf{y}$, peut être vu comme une collection de points $\{x_1, x_2, x_3, ...\}$, ou comme une mesure aléatoire $\mathbf{x} = \sum_i \delta_{x_i}$ où δ_x est la mesure de Dirac en x. Pour une fonction mesurable ϕ, on notera $\phi_{\mathbf{x}} = \int \phi(s)\mathbf{x}(ds) = \sum_i \phi(x_i)$. Lorsque $\phi = \mathbf{1}_B$ est la fonction indicatrice d'un borélien de $\mathbf{R}^2$,

on notera $N_{\mathbf{x}} = \mathbf{1}_{B,\mathbf{x}}$ la mesure de comptage associée. $N_{\mathbf{x}}(B)$ sera donc le nombre de points de $\mathbf{x}$ dans B. Souvent, on notera s un point courant de D, avec $s \notin \mathbf{x}$.

5.2 Classification non supervisée d'un mélange de processus de points

Problématique générale

De façon très générale, le problème considéré ici peut s'énoncer de la façon suivante. On observe sur une portion de plan un processus de points, par exemple la position de arbres dans une forêt, et on cherche à détecter des agrégats de points. Un agrégat sera une région du plan où la densité de points est significativement supérieure au reste du plan. On notera K le nombre d'agrégats. Il peut y avoir un seul agrégat (dans ce cas, $K = 1$), ou plusieurs ($K > 1$). On peut connaître le nombre d'agrégats à l'avance, ou au contraire K peut faire partie des grandeurs à estimer.

Approche directe

Une approche par maximum de vraisemblance non paramétrique est proposée dans [15] pour aborder ce problème. Le modèle est le suivant : on fait l'hypothèse que le processus de point $\mathbf{x}$ observé dans D est issu du mélange de deux processus de points uniformes : un processus (U_D), uniforme dans D, avec une probabilité p, et un processus (U_A) uniforme dans $A \subset D$ avec la probabilité complémentaire $1-p$. Le processus U_A est le processus d'intérêt. Dans le contexte de détection d'agrégats, il correspond aux K agrégats recherchés ; dans un contexte de segmentation en présence de bruit, il correspond au signal, qui peut être constitué de plusieurs composantes connexes. Le processus U_D est un processus de nuisance correspondant au bruit de fond. Dans ce modèle, la vraisemblance s'écrit :

$$L(A, p; \mathbf{x}) = \left(\frac{p}{|A|} + \frac{1-p}{|D|}\right)^{N_{\mathbf{x}}(A)} (1-p)^{N - N_{\mathbf{x}}(A)} \tag{24}$$

où $N = N_{\mathbf{x}}(D)$ et $|A|$ est l'aire de A. Supposons que A soit connu ; dans ce cas, l'estimateur de maximum de vraisemblance de p est

$$\hat{p} = \frac{N_{\mathbf{x}}(A)|D| - N|A|}{N(|D| - |A|)}, \tag{25}$$

et la vraisemblance partielle de A, obtenue en remplaçant p par $\hat{p}$ dans (24), est

$$L(A; \mathbf{x}) \propto \left(\frac{N_{\mathbf{x}}(A)}{|A|}\right)^{N_{\mathbf{x}}(A)} \left(\frac{N - N_{\mathbf{x}}(A)}{|D| - |A|}\right)^{N - N_{\mathbf{x}}(A)}. \tag{26}$$

L'estimation de A est par contre impossible en l'absence d'informations supplémentaires sur ce que peut constituer le domaine A. On pose que A s'écrit comme la réunion des cellules de Voronoï construites à partir de $\mathbf{x}$. La méthode consiste à rechercher les cellules de Voronoï

(dont l'union forme le domaine A) qui maximise (26). En l'absence de contrainte, on montre facilement que la solution recherchée est la réunion des M plus petites cellules, pour un nombre M à trouver. Cette solution mène en général à un ensemble A très peu régulier. Aussi est-il préférable d'imposer des contraintes de régularité sur A. On imposera que A est connexe (nombre de composantes connexes égal à 1) et que la frontière de A doit être régulière au sens où A est stable par rapport aux opérations de filtrage de la morphologie mathématique. Dans [15], plusieurs algorithmes sont proposés et testés sur des séries de simulation. Un algorithme qui donne de bonnes performances consiste à rechercher séquentiellement des solutions sous-optimales en partant des solutions optimales à m éléments, $m = 1, \ldots, N$, puis en utilisant des algorithmes de morphologie mathématique sur le graphe de Delaunay associé à la partition de Voronoï. On garde finalement la solution transformée qui maximise (26). Il faut noter toutefois que cet algorithme perd beaucoup en efficacité si on autorise plusieurs composantes connexes.

Des exercices de simulation ont montré que cette approche donne d'excellentes performances en termes d'erreur de classification ; en particulier, elle donne des erreurs de classification plus faibles que de nombreuses méthodes paramétriques telles que celles proposées dans Das Gupta & Raftery (1998) (voir également Fraley & Raftery, 2002).

Dans la présentation initiale [15], on impose que A est connexe. Dans [9], on relâche cette contrainte, et on recherche le nombre de composantes connexes de A, noté K, optimal pour un critère BIC où le nombre de paramètres est $K + 1$ (une densité par composante connexe) :

$$\mathrm{BIC}(K) = 2 \ln L(\hat{A}; \mathbf{x}) - (K + 1) \ln(N).$$

Des exercices de simulations présentés dans [9] ont montré que ce critère choisit le nombre de composantes connexes correct dans la très grande majorité des cas.

Approche indirecte

Comme nous l'avons déjà vu à la section 3.3, l'inverse de l'aire d'une cellule de Voronoï est un estimateur sans biais de la densité locale de points. La distribution de la surface des cellules de Voronoï construites à partir de $\mathbf{x}$ peut être modélisée comme un mélange de deux densités gamma (Hinde & Miles, 1980), $g(\theta)$ et $g_A(\theta_A)$ dont on estime les paramètres par l'algorithme EM.

L'approche indirecte développée dans [9] est de construire de nouvelles partitions de Voronoï à partir d'un processus poissonien indépendant générant une partition en M cellules. Chacune de ces M cellules est ensuite classée comme appartenant à A ou non. La répétition sur un grand nombre, J, de réalisations du processus poissonien permet d'attribuer à chaque point s de D une probabilité d'appartenir à A. Notons $(T_1^j, \ldots, T_M^j)$ la $j$ième partition en M cellules disjointes. Notons $\lambda_m^j = N_{\mathbf{s}}(T_m^j)/|T_m^j|$ la densité locale de points de $\mathbf{x}$ dans la cellule T_m^j. Alors, pour un

point s_i appartenant à une grille de discrétisation, on calcule

$$\pi_{ij}^{A} = \frac{P(s_i \in A \mid s_i \in T_m^j)}{P(s_i \in A \mid s_i \in T_m^j) + P(s_i \in K \backslash A \mid s_i \in T_m^j)} = \frac{g_A((\lambda_m^j)^{-1}; \hat{\theta}_A)}{(g_A(\lambda_m^j)^{-1}; \hat{\theta}_A) + g((\lambda_m^j)^{-1}; \hat{\theta})}. \quad (27)$$

Finalement, on calcule $\pi_i^A = 1/J \sum_{j=1}^{J} \pi_{ij}^A$, ce qui nous permet de cartographier la probabilité d'appartenir à A. Une segmentation peut s'obtenir en seuillant au-dessus de 0.5.

Par construction, cette méthode ne nécessite pas de fixer K à l'avance ; ici K est plutôt le résultat de l'algorithme. Des simulations ont montré que cette approche tend à trouver le même nombre de composantes connexes que l'approche directe, situées globalement aux mêmes lieux, mais qu'elles sont plutôt plus petites et plus régulières. L'approche indirecte est une alternative intéressante à la méthode directe dès que K est élevé.

Dans [9], cette méthode a été appliquée à des données forestières pour lesquelles il a été possible de segmenter le domaine d'étude en zones de fortes et de faibles densités. La méthode directe ne pouvait pas être utilisée en raison du nombre trop élevé de composantes connexes. La segmentation obtenue était conforme aux attentes des gestionnaires de la forêt.

5.3 Tests locaux d'indépendance

Le problème

Bien souvent, des variables sont attachées aux points d'un processus de points. C'est le cas par exemple dans les inventaires forestiers, dans lesquels on associe à chaque arbre sa variété, son diamètre à 1 m, ou toute autre variable quantitative. Plus généralement en écologie, les jeux de données sont par essence multivariés car les questions étudiées concernent principalement les peuplements, et donc les interactions et les dynamiques entre les différentes espèces qui constituent ce peuplement.

Une modélisation multivariée à l'ordre deux d'un processus ponctuel avec p types de points nécessite l'estimation et la modélisation des $p(p-1)/2$ fonctions d'interaction. Afin de réduire cette étape aux seules interactions significatives, il est nécessaire de disposer d'une méthode rapide et efficace permettant de tester, à l'échelle du domaine d'étude, l'indépendance entre processus de points. Toutefois, même si deux processus de points peuvent être considérés comme indépendants à petite échelle, il peuvent présenter des corrélations significatives à grande échelle, par exemple s'ils sont sous la dépendance commune de covariables externes (altitude, variables de sol, ...). Dans le jeu de données traité dans [11], deux espèces végétales du Burkina-Faso sont étudiées. Elles présentent des alternances de densité pseudo-périodiques appelés « brousse tigrée ». Un test global de co-occurrence des deux espèces rejettera donc l'hypothèse d'indépendance. A l'échelle locale cependant, ces espèces semblent indépendantes. Il est donc nécessaire de proposer une démarche faite de l'agrégation, à l'échelle globale, d'une multitude de tests locaux. Ceux-ci permettent par ailleurs de cartographier le niveau d'association locale entre les processus de points étudiés. C'est

ce qui a été proposé pour les processus de points dans [11] et pour des processus d'objets (fibres ou disques dans le plan) dans [13].

Tester l'indépendance locale entre deux processus de points

Notons $\mathbf{x}$ et $\mathbf{y}$ deux processus de points observés dans un domaine D. Notons $\mathbf{x}_{s,\delta}$ et $\mathbf{y}_{s,\delta}$ leur restriction dans une boule $\mathcal{B}_{s,\delta}$ centrée en s de rayon δ et notons $D^\delta = \{s \in D : \mathcal{B}_{s,\delta} \subset D\}$. Afin de mesurer le degré d'association entre $\mathbf{x}$ et $\mathbf{y}$, on définit la notion d'indépendance locale approximative. Cette notion correspond à l'idée intuitive que deux processus de points sont indépendants dans une petite région, sachant que tout le reste est connu (par exemple, des covariables).

Définition 1 *Les processus* $\mathbf{x}$ *et* $\mathbf{y}$ *sont localement approximativement indépendants (l.a.ind.) si, pour tout* $\delta > 0$ *et tout* $s \in D^\delta$*, il existe deux collections d'événements* $U_{s,\delta}$ *et* $V_{s,\delta}$*, tels que pour tout* $A, B \subset \mathcal{B}_{s,\delta}$*,*

$$\begin{aligned} & P(N_{\mathbf{x}_{s,\delta}}(A) = n, N_{\mathbf{y}_{s,\delta}}(B) = m \mid U_{s,\delta}, V_{s,\delta}) \\ = \;& P(N_{\mathbf{x}_{s,\delta}}(A) = n \mid U_{s,\delta}) P(N_{\mathbf{y}_{s,\delta}}(B) = m \mid V_{s,\delta})(1 + O(\delta)). \end{aligned}$$

On peut de la même manière définir la notion d'isotropie locale approximative. Cette notion correspond à l'idée intuitive que dans une petite région, après rotation aléatoire, le processus de points est distribué de façon identique au processus initial. Notons $\phi_{s,\delta}(\mathbf{x})$ une rotation aléatoire de $\mathbf{x}_{s,\delta}$ autour de s, indépendante de $\mathbf{x}$ et de $\mathbf{y}$.

Définition 2 *Le processus* $\mathbf{x}$ *est localement approximativement isotropique (l.a.is) si, pour tout* $\delta > 0$ *et tout* $s \in D^\delta$*, il existe une collection d'événements* $U_{s,\delta}$*, telle que pour tout* $A \subset \mathcal{B}_{s,\delta}$*,*

$$P(N_{\phi_{s,\delta}(\mathbf{x}_{s,\delta})}(A) = n \mid U_{s,\delta}) = P(N_{\mathbf{x}_{s,\delta}}(A) = n \mid U_{s,\delta})(1 + O(\delta)).$$

On peut facilement montrer que les deux composantes d'un processus de Cox bivarié sont l.a.ind. Elles sont également l.a.is sous des hypothèses de régularité sur les intensités. On peut également montrer que les composantes d'un processus de points markoviens sont l.a.ind. ssi la fonction d'interaction de paire croisée, notée $h_{\mathbf{x},\mathbf{y}}(s, s')$, tend vers 1 lorsque $|s - s'|$ tend vers 0. Elles sont également l.a.is si toutes les fonctions d'interaction de paires $h_{\mathbf{x},\mathbf{x}}(s, s')$ et $h_{\mathbf{y},\mathbf{y}}(s, s')$ sont continues. Les preuves de ces propositions sont dans [11].

Lorsque les composantes d'un processus de points bivariable sont à la fois l.a.ind et l.a.is, la distribution du couple $(\phi_{s,\delta}(\mathbf{x}_{s,\delta}), \mathbf{y}_{s,\delta})$ est approximativement identique à celle du couple $((\mathbf{x}_{s,\delta}), \mathbf{y}_{s,\delta})$. En d'autres termes,

$$P(N_{\phi_{s,\delta}(\mathbf{x}_{s,\delta})}(A) = n, N_{\mathbf{y}_{s,\delta}}(B) = m) = P(N_{\mathbf{x}_{s,\delta}}(A) = n) P(N_{\mathbf{y}_{s,\delta}}(B) = m)(1 + O(\delta)).$$

Les tests locaux d'indépendance peuvent donc être basés sur des rotations locales d'un processus de points par rapport à l'autre.

La statistique de test utilisée est une distance symétrique entre les plus proches voisins des processus $\mathbf{x}$ et $\mathbf{y}$ dans la boules $\mathcal{B}_{s,\delta}$:

$$d_{s,\delta}(\mathbf{x},\mathbf{y}) = \frac{1}{N_{\mathbf{x}_{s,\delta}}} \sum_{\mathbf{x}_i \in \mathbf{x}_{s,\delta}} \min_{\mathbf{y}_j \in \mathbf{y}_{s,\delta}} |\mathbf{x}_i - \mathbf{y}_j| + \frac{1}{N_{\mathbf{y}_{s,\delta}}} \sum_{\mathbf{y}_j \in \mathbf{y}_{s,\delta}} \min_{\mathbf{x}_j \in \mathbf{x}_{s,\delta}} |\mathbf{x}_i - \mathbf{y}_j|. \tag{28}$$

Des valeurs faibles ou élevées de cette statistique indiquent une interaction locale, au point s, entre les processus $\mathbf{x}$ et $\mathbf{y}$ dans la boule de rayon δ centrée en s. Dorénavant, on se fixe dans une boule $\mathcal{B}_{s,\delta}$, et, pour alléger les notations, on n'indique plus la référence à s et δ.

Afin de réaliser un test local, la statistique (28) est comparée à la statistique calculée pour des rotations aléatoires d'un processus par rapport à l'autre. On peut montrer que sous les hypothèses l.a.ind et l.a.is, les statistiques $d(\mathbf{x},\mathbf{y})$ et $d(\phi(\mathbf{x}),\mathbf{y})$ sont approximativement identiquement distribuées. Alors les grandeurs

$$r = P(d(\phi(\mathbf{x}),\mathbf{y}) \geq d(\mathbf{x},\mathbf{y}) \mid \mathbf{x},\mathbf{y})$$

et

$$t = P([d(\phi(\mathbf{x}),\mathbf{y}) - \mu_d]^2 \geq [d(\mathbf{x},\mathbf{y}) - \mu_d]^2 \mid \mathbf{x},\mathbf{y})$$

sont respectivement les p-valeurs de $d(\mathbf{x},\mathbf{y})$ et $[d(\mathbf{x},\mathbf{y}) - \mu_d]^2$ sous les hypothèses l.a.ind et l.a.is.

Les tests sont construits en calculant les p-valeurs approchées $\hat{r}_s$ et $\hat{t}_s$ à partir de K rotations uniformes et indépendantes. Une petite valeur de $\hat{t}_s$ indique que la distance $d(\mathbf{x},\mathbf{y})$ s'écarte de la moyenne au point s, ce qu'on interprète comme un écart en s aux hypothèses l.a.ind et l.a.is. La statistique $\hat{r}_s$ indique comment les processus sont associés : une grande valeur de $\hat{r}_s$ indique que la distance $d(\mathbf{x},\mathbf{y})$ est plus faible que la distance randomisée. C'est donc une indication d'une association positive (attraction). A l'inverse, une petite valeur de $\hat{r}_s$ indique une association négative (répulsion).

A partir du calcul des statistiques calculées aux M points d'une grille, on peut construire un test global (agrégé) d'association entre les processus de points.

Cette approche a été validée sur un très grand nombre de simulations, pour des processus poissoniens non stationnaires (qui vérifient les hypothèses l.a.ind et l.a.is) et pour des processus markoviens non stationnaires (qui ne les vérifient pas) ainsi que des versions perturbées par un bruit indépendant permettant d'estimer la puissance. Ces simulations montrent de bons résultats en terme de puissance. En l'absence de bruit, les hypothèses l.a.ind et l.a.is sont toujours rejetées pour les processus markoviens dès que le nombre de points est suffisant pour faire apparaître l'interaction entre processus (en pratique quelques dizaines de points, mais cela dépend de la distance d'interaction).

Des données d'une végétation de type « brousse tigrée » ont été analysées selon cette approche. Les positions de deux espèces végétales ont été relevées dans un domaine d'environ 10ha :

793 points pour la première espèce et 308 pour la seconde. La répartition est fortement non stationnaire, montrant des alternances pseudo-périodiques de régions à forte densité et de régions à très faible densité. L'analyse a permis de mettre en évidence une interaction positive entre les deux espèces que des analyses utilisant des outils classiques n'avaient pas pu déceler en raison de la non-stationnarité des données.

Une adaptation de cette méthode a permis d'analyser des relevés de deux espèces de mauvaises herbes dans une parcelle agricole [13]. Les mauvaises herbes étaient exhaustivement relevées dans 48 rectangles de 20 cm $\times$ 30 cm, régulièrement répartis dans la parcelle. Il a donc fallu adapter le type de déplacement à ce contexte pour tenir compte de la structure particulière de l'échantillonnage et de la direction privilégiée induite par le labour de la parcelle. Les rotations ont donc été remplacées par des translations parallèles au labour. Sur ces données, l'hypothèse d'indépendance locale n'est pas rejetée. Cela étant, l'adaptation de notre méthode à ce type particulier d'échantillonnage entraîne une perte importante de puissance, qui ne nous permet pas de conclure de façon ferme.

Tester l'indépendance locale entre structures spatiales

Il est possible d'étendre cette approche à des processus de fibres ou à des processus booléens. Dans [13], on montre l'association négative entre la présence de racines et la densité du sol, confirmant par là l'hypothèse que les réseaux racinaires se développent préférentiellement dans les zones de faible densité. Les déplacements employés dans cet exemple ne sont pas des rotations aléatoires, mais des translations aléatoires, l'idée générale étant que si la densité et le processus de racines sont indépendants, la densité moyenne du sol observée à une distance d des racines doit avoir une densité identique quel que soit d. Le test de cette hypothèse se fait en déplaçant aléatoirement un processus par rapport à l'autre. Les objets occupent une fraction, p, non négligeable du domaine d'étude (13 % dans l'exemple ci-dessus). Aussi, il existe une probabilité p qu'un point servant au calcul de la densité soit recouvert par une racine après une translation aléatoire. Si on utilise n translations, la probabilité de non-recouvrement décroît très vite, en $(1-p)^n$, ce qui affecte fortement la puissance de la méthode. Il faut donc tenir compte de ce problème de censure. Notons $Z(x)$ la variable densité, modélisée par un champ aléatoire stationnaire à l'ordre 2 et isotropique, et $B(s)$ le processus d'objet, modélisé comme un processus booléen (Lantuéjoul, 1999), observé dans un domaine circulaire D centré en 0. On considère comme statistique d'intérêt la valeur moyenne de Z calculée à une distance d de B. On note T une translation, et $T(B)$ le processus B translaté de T : $T(B) = B \oplus T$.

Afin de tenir compte de la censure générée par une translation aléatoire T, on note $Z_d(B)$ la moyenne des valeurs de $Z(s)$ lorsque s est à une distance d de B, et n'est ni dans B ni dans $T(B)$; on note $Z_d(T(B))$ la moyenne des valeurs de $Z(s)$ lorsque s est à une distance d de $T(B)$, et n'est ni dans B ni dans $T(B)$. On utilise comme statistique d'intérêt la différence $H_T(d) = Z_d(B) - Z_d(T(B))$. Sous l'hypothèse H_0 d'indépendance entre Z et B, l'espérance de $H_T(d)$ est

nulle. Si on procède à n translations aléatoires indépendantes, $T_1, \ldots, T_n$, les statistiques $H_{T_i}(d)$ sont indépendantes et sous H_0, on a $P(0 \leq \inf_i\{H_{T_i}(d)\}) = 1/(n+1)$ et $P(0 \geq \sup_i\{H_{T_i}(d)\}) = 1/(n+1)$. On obtient donc un intervalle de confiance au niveau $1/(n+1)$ en représentant les enveloppes inférieures et supérieures de $H_{T_i}(d)$ en fonction de d.

Perspectives

On pourra noter que la problématique de l'agrégation globale de tests locaux est commune avec celle rencontrée lors de la détection de Zones de Changement Abrupt présentée à la Section 3.1. Pour autant, la méthode mise en œuvre diffère sensiblement, car ici nous ne connaissons pas les propriétés mathématiques du champ des p-valeurs locales. Une piste de recherche intéressante consisterait à établir certaines propriétés de la loi du champ de p-valeurs sous H_0.

5.4 Déformation de l'espace

Dans la section 5.3, j'ai présenté comment on pouvait en quelque sorte filtrer le problème de l'éventuelle non-stationnarité des processus ponctuels dans la recherche d'un test d'association en travaillant sur les p-valeurs d'un ensemble de tests locaux. Une autre façon de procéder consiste à faire une transformation des coordonnées dans un nouvel espace dans lequel le processus de points transformé serait stationnaire. Cette approche, connue et utilisée depuis longtemps pour les processus de points indicés dans le temps, est plus difficile à mettre en œuvre dans l'espace, car les transformations y sont plus complexes. Transformer (ou déformer) l'espace afin de rendre un processus stationnaire est un thème de recherche ancien dans l'unité, mais les recherches ont plutôt concerné les champs aléatoires continus (Meiring *et al.* , 1997 ; Perrin & Senoussi, 2000) que les processus de points. Il existe pourtant une unité mathématique profonde puisqu'ils peuvent tous deux s'écrire sous la forme de mesure aléatoire. Le travail dans [12] étudie les conditions que doit vérifier une transformation Φ pour transformer un processus de points $\mathbf{x}$ en un processus $\tilde{\mathbf{x}}$ stationnaire à l'ordre 1 et/ou à l'ordre 2.

Il est nécessaire d'introduire quelques notations supplémentaires. On note $M_{\mathbf{x}}(ds) = E[N_{\mathbf{x}}(ds)]$ la mesure du premier moment de $\mathbf{x}$. Si cette mesure est absolument continue (ce que nous supposerons toujours), alors $M_{\mathbf{x}}(ds) = m_{\mathbf{x}}(s)ds$. On note $R_{\mathbf{x}}$ la mesure d'un des seconds moments de $\mathbf{x}$, par exemple la mesure covariance :

$$R_{\mathbf{x}}(ds, ds') = E[N_{\mathbf{x}}(ds)N_{\mathbf{x}}(ds')] - M_{\mathbf{x}}(ds)M_{\mathbf{x}}(ds').$$

Si la densité associée existe (ce que nous supposons également), on a : $R_{\mathbf{x}}(ds, ds') = r_{\mathbf{x}}(s, s')dsds'$. Un processus de points $\mathbf{x}$ est stationnaire à l'ordre 2 si le premier moment est invariant par translation et si le second moment ne dépend que $(s - s')$, i.e. $m_{\mathbf{x}}(s) = m$ et $r_{\mathbf{x}}(s, s') = r_{\mathbf{x}}(s - s')$ pour tout s et s' de D.

Définition 3 *Un processus de points* $\mathbf{x}$ *est dit réductible à la stationnarité d'ordre 2 s'il existe un difféomorphisme* $\Phi : D \to U$, $U \in \mathbf{R}^d$, *pour lequel* $\tilde{\mathbf{x}}$ *défini par* $\psi_{\tilde{\mathbf{x}}} = (\psi \circ \Phi)_{\mathbf{x}}$ *est stationnaire d'ordre 2 sur* U.

Il est alors aisé de montrer que l'homogénéisation à l'ordre 1 implique qu'il existe une constante μ telle que

$$m_{\mathbf{x}}(s) = \mu |J_\Phi(s)|, \tag{29}$$

où J_Φ est le Jacobien de la transformation J, et $|J_\Phi(s)|$ son déterminant calculé au point $s \in D$. L'homogénéisation à l'ordre 2 est vérifiée s'il existe une fonction $r_{\tilde{\mathbf{x}}}$ telle que

$$\frac{r_{\mathbf{x}}(s, s')}{\sqrt{r_{\mathbf{x}}(s, s) r_{\mathbf{x}}(s', s')}} = \frac{r_{\tilde{\mathbf{x}}}(\Phi(s) - \Phi(s'))}{r_{\tilde{\mathbf{x}}}(0)}. \tag{30}$$

Cette seconde équation est identique à celle obtenue dans Perrin et Senoussi (2000) pour les champs gaussiens. Il faut noter qu'un processus n'est homogénéisable à l'ordre 1 et 2 simultanément que si la condition de compatibilité $r_{\mathbf{x}}(s, s) \propto m^2_{\mathbf{x}}(s)$ est vérifiée. Pour les processus de points poissoniens inhomogènes, une homogénéisation au premier ordre est toujours suffisante, puisque la mesure de second ordre est nulle en dehors de la diagonale.

Sous des conditions de régularité assez légères, trouver les solutions de l'équation d'homogénéisation du second ordre (30) est un problème d'analyse différentielle qui possède une solution unique, voir [12] pour les détails de cette solution. A l'inverse, l'homogénéisation au premier ordre ne possède pas toujours une solution unique : il existe plusieurs façons « d'étaler » une densité afin de la rendre constante. Ceci peut par exemple se faire en « étalant » selon chaque coordonnée successivement, mais aussi en « étalant » de façon radiale à partir d'un centre, lorsque le domaine est en forme étoilée (star shaped).

Les résultats établis dans [12] ouvrent la voie à une estimation non paramétrique de la transformation des coordonnées de l'espace. Gay, Barnouin & Senoussi (2006) homogénéisent au premier ordre la densité des fermes étudiées dans le cadre d'une étude épidémiologique concernant l'infection de troupeaux bovins par la mammite. Cette homogénisation étant faite, l'analyse de l'agrégation de scores d'infection peut alors se faire dans un cadre où la densité de fermes est stationnaire.

6 Recherches en cours et futures

Recherches méthodologiques

D'un point de vue méthodologique, mes recherches se feront essentiellement dans le prolongement de mes recherches les plus récentes, avec un intérêt fort pour la modélisation spatio-temporelle des données climatiques et environnementales. Elles s'organisent essentiellement autour de deux axes :

1. La poursuite des recherches concernant la **détection de structures** pour les données spatiales : zones de changement abrupt, classification supervisée et non supervisée, détection d'agrégats. Appartiennent à ce thème la détection de ZCA pour des données de comptage sur des petits effectifs, avec des applications à la génétique des populations ; détection de ruptures, à l'intérieur de parcelles agricoles, sur des séquences temporelles d'images satellites.

   Bien que semblant réunir des thèmes dispersés, cet axe regroupe des questions statistiques communes présentes dans ce mémoire, avec entre autres :
   - la question de l'agrégation de tests locaux en un test global unique. Cette question était apparue pour la détection de ZCA et pour le test de la dépendance entre processus ponctuels. Dans le premier cas la réponse avait été apportée à partir de la géométrie d'un champ de χ^2, alors que dans le second cas une approche utilisant une approximation utilisant la notion de porté intégrale (Lantuéjoul, 2002) avait été utilisée, faute de connaître les propriétés du champ de p-valeurs considéré. Or l'approche géométrique est généralisable à un grand nombre de champs aléatoires, comme le laisse envisager une littérature récente à ce sujet (Taylor and Worsley, 2007). Il y a là une piste de recherche tout à fait intéressante.
   - la question de l'utilisation des structures détectées pour améliorer la prédiction spatiale. Cet aspect des choses, beaucoup plus appliqué, est d'un intérêt immédiat pour les praticiens.
2. Utilisation des **gaussiennes dissymétriques** pour les séries temporelles et pour les champs aléatoires (SNRF, cf. section 4 et les publications [5] et [28]). Les principales questions de recherche dans ce domaine sont : l'étude du modèle proposé dans le mémoire [M2] y compris la prédiction spatiale, l'inférence statistique dans ce cadre et l'étude du changement de support pour ces modèles en ayant en vue la question de la désagrégation des variables. Cet axe de travail est en partie lié à la thèse de Cédric Flécher [T1] en collaboration avec Philippe Naveau (LSCE, CNRS) et servira au projet CLIMATOR (voir ci-dessous).

Recherches finalisées

D'un point de vue finalisé maintenant, les recherches des trois prochaines années se feront dans le cadre d'un certain nombre de projets de recherche dont voici les principaux.

1. Le projet CLIMATOR, financé dans le cadre du prgramme ANR/vulnérabilité, explore la vulnérabilité des pratiques agricoles à l'horizon 2035-2065 dans le cadre de l'étude de l'impact du changement climatique sur le territoire de la France. Ce projet regroupe un total de 18 partenaires spécialistes de modèles agronomiques et de modèles de sol. Il s'agit, dans l'état actuel des connaissances, d'explorer les réponses d'un certain nombre de modèles de sol, de prairies et de plantes à des forçages climatiques simulant les conditions climatiques et météorologiques à l'horizon considéré.

   Cette étude nécessite la mise au point de générateurs de séries météorologiques au pas de temps horaire et à l'échelle de la parcelle pour les principales variables de forçage des modèles agronomiques : température, T, précipitation, P, rayonnement, R, humidité, H. Ces séries temporelles multivariées doivent être physiquement cohérentes ; elles doivent respecter l'auto-corrélation temporelle observée sur les séries actuelles et reproduire les comportements en tendance prédits par les modèles climatiques (GCM couplés avec ARPEGE sur la région européenne).

   Nous utiliserons pour cela une modélisation stochastique faisant appel aux vecteurs aléatoires Skew-Normal étudiés dans [5], en liaison avec la thèse en cours de Cédric Flécher [T1].
2. Le projet SADMO est un projet européen regroupant des partenaires portugais, espagnols, italiens et grecs ayant pour objet la définition et l'étude d'indicateurs de désertification sur deux zones d'études dans le Sud du Portugal et de l'Espagne. Notre rôle consiste à faire une analyse spatio-temporelle de l'évolution d'un indicateur de végétation calculé sur des images satellites à haute résolution spatiale (LANDSAT) sur ces deux zones d'étude, en investigant plus particulièrement deux points : i) tester la stationnarité temporelle de cet indicateur contre une hypothèse alternative de son aggravation ; ii) étudier le rôle de la résolution à laquelle ces données sont collectées et/ou étudiées ; en effet à quantité de végétation fixée sur un certain domaine, sa répartition spatiale plus ou moins hétérogène peut être considérée comme un indice de désertification. Nous partirons des résultats obtenus dans la thèse de S. Garrigues ([8],[7], [3], [2], [17]).
3. Un projet ECOGER dans lequel on cherche à estimer les flux de gènes entre les compartiments cultivés et sauvages des peupliers, en relation avec l'analyse génétique du parasite du peuplier. Deux zones d'études ont été définies. La première, le long de la Loire, vise à estimer la fonction de dispersion du pollen de peuplier sauvage (*P. Nigra*) ; la seconde, le long de la vallée de la Durance, vise à comprendre la propagation du parasite du peuplier, des forêts de mélèzes dont il a besoin pour se reproduire, vers les peupleraies cultivées et sauvages situées dans la vallée de la Durance.
4. Un projet d'épidémiologie végétale visant à comprendre la dynamique des populations de Carpocapses, qui est un ravageur des vergers de pommiers et de poiriers. Des larves de carpocapses ont été récoltées dans un domaine comportant une dizaine de vergers aux

pratiques culturales très diverses, durant trois saisons successives. Les données génétiques acquises sur ces larves permettent d'estimer des liens de parenté entre individus et donc les déplacements entre vergers.

Sur ces deux derniers projets, qui ne sont pas au cœur de ma thématique, j'interviens essentiellement en support méthodologique pour les équipes de biologistes concernées.

7 Annexes : étudiants encadrés, publications et autres références

Etudiants encadrés et co-encadrés

Thèses

[T1] Flecher, C. (2006 –) "Développement de méthodes statistiques pour la mise au point d'un générateur de climat adapté à l'utilisation des scénarii de changement climatique", thèse de l'Université de Montpellier II, ED SIBAGHE, (co-encadrement avec Ph. Naveau, LSCE, CNRS, et N. Brisson CSE, INRA Avignon).

[T2] Gabriel E. (2001 – 2004) "Détection de zones de changement abrupt dans des données spatiales et application à l'agriculture de précision", thèse de l'Université de Montpellier II, ED ISS, option biostatistique (co-encadrement avec M. Guérif, CSE, INRA Avignon).

[T3] Garrigues, S. (2001 – 2004) "Hétérogénéité spatiale des surfaces terrestres en télédétection ; caractérisation et influence sur l'estimation des variables biophysiques", thèse de l'ENSA-R (en co-encadrement avec F. Baret, CSE, INRA Avignon).

Etudiants encadrés en Master et DEA

[M1] Soulié, A. (2006) "Estimation et interpolation de lois dissymétriques", rapport de Master II Recherche, option biostatistique, Université de Montpellier II.

[M2] Allemand B. (2004) "Comparaison de deux modélisations probabilistes pour la classification de données géostatistiques : construction et analyse des algorithmes d'estimation", travail de fin d'étude, IUP Génie Mathématique et Informatique, Avignon (en co-encadrement avec N. Peyrard, Biométrie, INRA Avignon).

[M3] Baragnon G. (2003) "Analyse bayésienne rétrospective d'une rupture dans des séries phénologiques", rapport de DEA de Biostatistique, Université de Montpellier II.

[M4] Gabriel E. (2001) "Détection de zones de changement abrupt pour un champ gaussien", rapport de DEA de Biostatistique, Université de Montpellier II.

[M5] de Beaufort L. (2000) "Définition d'une méthode de cartographie d'indice foliaire destinée à la validation de produits de capteurs satellites", rapport de DAA *Traitement de l'Information Spatiale*, ENSA-R.

[M6] Royer C. (1999) "Krigeage avec carte thématique et classification spatiale", rapport de stage de 3ème année de l'École des Mines de Paris, Note S-376, Centre de Géostatistique, Fontainebleau.

[M7] Navarro Sanchez I. (1997), "Estimation de l'influence de l'environnement sur la température", mémoire de DEA de biostatistique, Université de Montpellier II (en co-encadrement avec P. Monestiez, Biométrie, INRA Avignon).

Références

[1] Gabriel, E., Allard, D., Mary, B. and Guérif, M., (2007) Detecting zones of abrupt change in soil data, with an application to an agricultural field, *European Journal of Soil Science*. DOI : 10.1111/j.1365-2389.2007.00920.x

[2] Garrigues, S., Allard, D., Baret, F. and Morisette, J. (2007) Multivariate Quantification of Landscape Spatial Heterogeneity using Variogram Models, *Remote Sensing of Environment*. DOI :10.1016/j.rse.2007.04.017

[3] Garrigues, S., Allard, D., Baret, F. (2007) Using first and second order variograms for characterizing landscape spatial structures from remote sensing imagery *IEEE TGRS*. **45**, 1823 - 1834.

[4] Gabriel, E. and Allard, D. (à paraître) Evaluating the Sampling Pattern When Detecting Zones of Abrupt Change, *Environmental and Ecological Statistics*.

[5] Allard D. and Naveau, P. (2007) A new spatial skew-normal random field model, *Communications in Statistics*, **36**, 1821-1834.

[6] Allard D., Froidevaux R. and Biver, P. (2006) Conditional Simulation of Multi-Type Non Stationary Markov Object Models Respecting Specified Proportions, *Mathematical Geology*, **38**, 959-986.

[7] Garrigues, S., Allard, D., Baret, F. and Weiss, M. (2006) Influence of the spatial heterogeneity on the non linear Estimation of Leaf Area Index from moderate resolution remote sensing data, *Remote Sensing of Environment*, **105**, 286-298.

[8] Garrigues, S., Allard, D., Baret, F. and Weiss M. (2006) Quantifying spatial heterogeneity at the landscape scale using variogram models, *Remote Sensing of Environment*, **103**, 81-96.

[9] Magnussen S., Allard D., and Wulder M. (2006) Poisson Voronoï tiling for finding clusters in spatial point patterns, *Scan. J. For. Res.*, **21**, 239-248.

[10] Monestiez P., Courault D., Allard D. and Ruget F. (2001) Spatial interpolation of air temperature using environmental context : application to a crop model, *Environmental and Ecological Statistics*, **8**, 297-309.

[11] Allard D., Brix A., Chadœuf J. and Couteron P. (2001) Testing Local Independence Between two Point Processes, *Biometrics*, **57**, 508-517.

[12] Senoussi R., Chadœuf J. and Allard D. (2000) Weak Homogenization of Point Processes by Space Deformations, *Advances in Applied Probability*, **32**, 948-959.

[13] Chadœuf J., Brix A., Pierret A. and Allard D. (2000) Local tests in agricultural research, *Journal of Microscopy*, **200**, 32-41.

[14] Allard D. (1998) Geostatistical Classification and Class Kriging, *Journal of Geographical Information and Decision Analysis*, **2**, 87-101.

[15] Allard D. and Fraley Ch. (1997) Non Parametric Maximum Likelihood Estimation of Features in Spatial Point Processes Using Voronoï Tesselation, *Journal of the American Statistical Association*, **92**, 1485-1493.

[16] Allard D. (1993) Some Connectivity Characteristics of a Boolean Model, *Acta Stereologica*, **12**, 191-196.

Articles soumis ou en révision

[17] Garrigues, S., Allard, D., Baret, F. (à paraître) Modeling Temporal Changes in Surface Spatial Heterogeneity over an Agricultural site *Remote Sensing of Environment.*

[18] Bel, L., Allard, D., Laurent, J.M., Cheddadi, R. and Bar-Hen, A., Adapting the CART algorithm to spatial data : application to ecological data

[19] Gabriel, E. Allard, D., and Bacro, J.-N., On non-stationary Gaussian random fields, their derivatives and a related χ^2 field.

Edition d'ouvrage

[20] Monestiez P., Allard D. and Froidevaux R. (Eds) (1999) *geoENV III : Geostatistics for environmental applications*, Kluwer Academic Publishers, Dordrecht.

Chapitres d'ouvrages et proceedings avec comité de lecture et sélection

[21] Allard, D., Gabriel, E. (2007) Détection de zones de changement abrupt pour des variables non permanentes du sol : vers la définition de zones homogènes ?, in Guérif, M. and King, D., Coords., Editions Quae, Paris, pp. 165–76.

[22] Allard, D. (2006) Validation d'un modèle géostatistique pour l'interpolation : application à un événement pluvieux, in *Statistiques Spatiales*, Eds. Droesbeke, J.-J. et Lejeune, M., Technip, Paris, pp. 403–414.

[23] Chilès, J.-P. and Allard, D. (2005) Stochastic Simulation of Soil Variation, in *Geographic Information Technologies for Environmental Soil-Landscape Modelling*, Ed. Grunwald, S., CRC Press, Boca Raton, pp. 289-321.

[24] Bel, L., Laurent, J.M., Bar-Hen, A., Allard, D. and Cheddadi, R. (2005) A spatial extension of CART : Application to classification of ecological data, in *Geostatistics for Environmental Applications*, Eds. P. Renard H. Demougeot-Renard and R. Froidevaux, Springer, pp. 99-109.

[25] Allard D., Froidevaux R. and Biver P. (2005) Accounting for non-stationarity and interactions in object simulation for reservoir heterogeneity characterization, in *Geostatistics Banff 2004*, Eds. Leuangthong, O., Deutsch, C. V., Springer, pp. 155-164.

[26] Gabriel E. and Allard D. (2005) Assessing the power of zones of abrupt change detection test, in *Geostatistics Banff 2004*, Eds. Leuangthong, O., Deutsch, C. V., Springer, pp. 1103-1008.

[27] Garcia M., Allard D., Foulon D. and Delisle S. O. (2005) Fine scale rock properties : Towards the spatial modeling of regionalized probability distribution functions, in *Geostatistics Banff 2004*, Eds. Leuangthong, O., Deutsch, C. V., Springer, pp. 579-590.

[28] Naveau P. and Allard D. (2005) Modeling skewness in spatial data analysis without data transformation, in *Geostatistics Banff 2004*, Eds. Leuangthong, O., Deutsch, C. V. , Springer, pp. 929-931.

[29] Gabriel, E., Allard, D. and Bacro, J.N. (2004) Detecting Zones of Abrupt Change : application to soil data, in *geoENV IV : Geostatistics for Environmental Applications* Eds. Sanchez Vila, X. Carrera J. and Gomez Hernandez, J.J., Kluwer Academic Publisher, pp. 437-448.

[30] Courault, D., Olioso, A., Lagouarde J.-P., Monestiez, P. and Allard, D. (2004) Influence des cultures sur les variables climatiques in *Organisation spatiale des activités agricoles et processus environnementaux*, Eds. Monestiez, P., Lardon S. et Séguin, B, INRA Editions, pp. 303-320.

[31] Gleyze J.-F., Bacro, J.-N. and Allard D. (2001) Detecting regions of abrupt change : Wombling procedure and statistical significance, in *geoENV III : Geostatistics for environmental applications*, Eds. Monestiez P., Allard D. and Froidevaux R., Kluwer Academic Publishers, Dordrecht, pp. 311-322.

[32] Allard D. and Monestiez P. (1999) Geostatistical Segmentation of Rainfall Data, in *geoENV II : Geostatistics for Environmental Applications*, Eds. Gomez-Hernandez J., Soares A. and Froidevaux R., Kluwer Academic Publishers, Dordrecht, pp. 139-150.

[33] Monestiez P., Allard D., Navarro Sanchez I. and Courault D. (1999) Kriging with Categorical External Drift : Use of Thematic Maps in Spatial Prediction and Application to Local Climate Interpolation for Agriculture, in *geoENV II : Geostatistics for environmental applications*, Eds. Gomez-Hernandez J., Soares A. and Froidevaux R., Kluwer Academic Publishers, Dordrecht, pp. 163-174.

[34] Allard D., Armstrong M., and Kleingeld W. J. (1994) The Need for a Connectivity Index in Mining Geostatistics in *Geostatistics for the Next Century*, Ed. R. Dimitrakopoulos, Kluwer Academic Publishers, Dordrecht, pp. 293-302.

[35] Allard D. (1993) Simulating a Geological Lithofacies with Respect of Connectivity Information Using the Truncated Gaussian Model, in *Geostatistical Simulations*, Eds. M. Armstrong and P. Dowd, Kluwer Academic Publishers, Dordrecht, pp. 197-211.

[36] Allard, D. and Heresim Group (1992) On the connectivity of two random set models : the truncated gaussian and the Boolean, in *Geostatistics Troia '92*, Ed. Soares A., Kluwer Academic Publishers, Dordrecht, pp. 467-478.

Autres références

Adler, R. (2000) On excursion sets, tube formulas and maxima of random fields. *The Annals of Applied Probability*, **10**, 1-74.

Ambroise, C., and Govaert, G. (1998) Convergence Proof of an EM-type Algorithm for Spatial Clustering. *Pattern Recognition Letters*, **19**, 919-927.

Atteia, O., Dubois, J.-P., and Webster, R. (1994) Geostatistical analysis of soil contamination in Swiss Jura. *Environmental pollution*, **86**, 513–327.

Azzalini, A. (2005) The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159–188.

Banerjee, S., Gelfand, A., and Sirmans, C. (2003) Directional rates of change under spatial process models. *Journal of the American Statistician Association*, **98**, 946–954.

Banerjee, and Gelfand, A., (2006) Curvilinear Boundary Assessment under Spatial Processes Models. *Journal of the American Statistical Association*, **101**, 1487-1501 .

Barbujani, G., Oden, N., and Sokal, R. (1989) Detecting areas of abrupt change in maps of biological variables. *Systematic Zoology*, **38**, 376–389.

Besag, J. (1986) On the Statistical Analysis of Dirty Pictures *Journal of the Royal Statistical Society. Series B*, **48**, 259–302

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics*, **43**, 1572-9052.

Bocquet-Appel, J-P., and Bacro J-N. (1994) Generalized Wombling. *Systematic Biology*, **43** 442–448.

Breiman, L., Friedman, J.H. ,Olshen, R.A. and Stone, C.J. (1984) *Classification and regression trees*. Wadsworth International Group, Belmont.

Cao, J. (1999) The size of the connected components of excursion sets of χ^2, t and F fields. *Advances in Applied Probability (SGSA)*, **31**, 579–595.

Celeux, G., Forbes, F., and Peyrard, N. (2003) EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, **36**, 131–144.

Chilès, J.P., and Delfiner, P. (1999) *Geostatistics : Modeling Spatial Uncertainty*. John Wiley & Sons, New York.

Cressie, N. (1993) *Statistics for spatial data, Revised Edition*. John Wiley & Sons, New York.

Diggle, P.J. (1983) Statistical Analysis of Spatial Point Patterns. London : Academic press.

Das Gupta, A., and Raftery, A.E. (1998) Detecting Features in Spatial Point Processes with Clutter Via Model-Based Clustering. *Journal of the American Statistical Association* **93**, 294–302.

Fortin, M-J. (1994) Edge detection algorithms for two-dimensional ecological data. *Ecology*, **75**, 956–965.

Fraley, C., and Raftery, A. E. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* **97**, 611–631.

Gay E., Barnouin J., and Senoussi R. (2006) Spatial and temporal patterns of herd somatic cell score in France. *Journal of Dairy Science* **89**, 2487-s-2498.

Genz, A. (1992) Numerical computation of multivariate probabilities. *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Genton, M.G. (Ed.) (2004) *Skew-Elliptical Distributions and Their Applications : A Journey Beyond Normality*. Chapman & Hall/CRC, Boca Raton.

Hinde, A.L. and Miles, R.E. (1980) Monte-Carlo estimates of the distribution of the random polygons of the Voronoï tesselation with respect to a Poisson process. *Journal of Statistical Computation and Simulation*, **10**, 205-223.

IFEN (1996) *CORINE Landcover France : un nouvel outil au service de l'environnement et de la gestion de l'espace*. Paris.

Lantuéjoul, C. (2002) *Geostatistical simulations : models and algorithms*. Springer Verlag, Berlin.

Laurent, J.M., Bar-Hen, A., François, L., Ghislain, M., and Cheddadi, R. (2004) Refining vegetation simulation models : From Plant Functional Types to Bioclimatic Affinity Groups of plants. *Journal of Vegetation Science*, **15**, 739–746.

Meiring, W., Monestiez, P., Sampson, P.D., and Guttorp, P. (1997) Developments in the modelling of nonstationary spatial covariance structure from space-time monitoring data. In *Geostatistics Wollongong '96*, Vol. 1, eds E.Y. Baafi and N. Schofield. Kluwer Academic Publishers, Dordrecht, pp. 162-173.

Myneni, R. B., F. G. Hall, P.J. Sellers, and A.L. Marshak (1995) The interpretation of spectral vegetation indexes, *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 481-486.

Perrin, O., and Senoussi, R. (2000) Reducing non-stationary random fields to stationarity using space deformation. *Statist. Prob. Lett.*, **48**, 23-32.

Proust, M. (1927) *Le temps retrouvé*, Gallimard, Paris.

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Taylor, J.E. and Worsley, K.J. (2007) Random fields of multivariate test statistics, with applications to shape analysis and fMRI. *Annals of Statistics*, accepted.

Wackernagel, H. (2003) *Multivariate Geostatistics. An Introduction with Applications*, 3rd edition. Springer-Verlag, Heidelberg.

Womble, W. (1951) Differential systematics. *Science*, **114**, 315–322.

Worsley, K. (2001) Testing for signals with unknown location and scale in a χ^2 random field, with application to fMRI. *Advances in Applied Probability (SGSA)*, **33**, 773–793.