Université de Montpellier II
Sciences et Techniques du Languedoc

THÈSE

pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER II

Discipline : Biostatistique

Ecole doctorale : Information, Structures, Systèmes

présentée et soutenue publiquement le 04 décembre 2009

par

Pierrette CHAGNEAU

Modélisation bayésienne hiérarchique pour la prédiction multivariée de processus spatiaux non gaussiens et processus ponctuels hétérogènes d'intensité liée à une variable prédite

Application à la prédiction de la régénération en forêt tropicale humide

Composition du jury

M. Jean-Noël BACRO	Université de Montpellier 2	*Directeur de thèse*
M. Alain FRANC	INRA Bordeaux	*Examinateur*
M. Carlo GAETAN	Université de Venise	*Rapporteur*
Mme Chantal GUIHENNEUC-JOUYAUX	Université Paris-Descartes	*Rapporteur*
M. Jean-Michel MARIN	Université de Montpellier 2	*Examinateur*
M. Frédéric MORTIER	CIRAD Montpellier	*Co-directeur*

Remerciements

L'heure est venue d'adresser mes remerciements à toutes les personnes qui, de près ou de loin, m'ont aidée et accompagnée pour mener à bien ce travail de thèse.

Je tiens, tout d'abord, à remercier mon trio d'encadrants : Jean-Noël Bacro, Nicolas Picard et Frédéric Mortier.

Merci à Jean-Noël d'avoir accepté de diriger cette thèse. Loin de l'image que l'on véhicule du directeur de thèse absent et arrogant, Jean-Noël s'est toujours montré disponible et accueillant. Il a su me faire progresser grâce à ses critiques constructives, sans jamais oublier de me glisser un mot d'encouragement. Je mesure la chance que j'ai eu d'effectuer ma thèse sous sa direction. Les moments partagés lors de notre petite escapade au Chili resteront sans aucun doute parmi les bons souvenirs de ces trois années.

Merci à Nicolas d'avoir toujours suivi attentivement mon travail malgré les milliers de kilomètres qui nous séparaient. Cela n'a pas toujours été facile, surtout quand, trop occupée, j'oubliais de le tenir au courant de l'avancée de mes travaux. Son éclairage sur les questions forestières a contribué à enrichir mon travail de recherche. Son efficacité redoutable en informatique, sa capacité à débusquer la moindre petite erreur ou imprécision et son pouvoir à sortir de ses tiroirs la référence bibliographique qui vous fait défaut m'ont été d'une aide précieuse. Et que dire de ses truculents jeux de mots qui nous font toujours sourire !

Un merci tout particulier à Frédéric. Assurer le suivi d'une thèse au quotidien n'est pas toujours de tout repos. Le temps que Frédéric a consacré à mon encadrement, parfois au détriment de ses soirées ou de ses congés, les nombreux conseils qu'il m'a donnés et les services qu'il m'a rendus traduisent à mes yeux sa grande générosité. Frédéric m'a donné la chance de partager son quotidien de chercheur. J'y ai vu quelqu'un de passionné et jamais à cours d'idées. Et même si parfois, il vous poursuit de ses assiduités statistiques jusqu'à la cantine, son enthousiasme pour la recherche ne serait être que communicatif. Merci Fred de t'être montré exigeant et de m'avoir poussée à donner le meilleur de moi-même pendant ces trois années.

Je tiens également à remercier les rapporteurs de cette thèse, Carlo Gaetan et Chantal Guihenneuc-Jouyaux, pour la rapidité avec laquelle ils ont lu mon manuscrit et pour l'intérêt qu'ils ont porté à mon travail. Merci également à Jean-Michel Marin d'avoir accepté de présider mon jury de soutenance et à Alain Franc pour y avoir participé en tant qu'examinateur. Leurs remarques et les discussions qu'elles ont amenées me permettront, j'en suis

sure, d'approfondir certaines pistes de travail et d'aborder des questions restées en suspens.

Deux comités de thèses sont venus ponctuer ces trois années. Nicolas Bez, Joël Chadœuf, Sophie Gerber et Vivien Rossi ont gentiment accepté d'y participer. Le regard bienveillant qu'ils ont porté sur mon travail m'a rassuré et leurs nombreuses idées sont venues nourrir ma réflexion. Je les en remercie.

Merci à Sylvie Gourlet-Fleury et à Jean-Marc Bouvet de m'avoir accueillie au CIRAD au sein leur unité respective, l'UR Dynamique des forêts naturelles et l'UR Diversité génétique et amélioration des espèces forestières, et d'avoir tout mis en œuvre pour que ma thèse se déroule dans les meilleures conditions possibles. Les moments agréables passés auprès des membres de l'UR 37 et de l'UR 39 m'ont fait trouver ces trois années beaucoup moins longues. Partager votre quotidien a été fort enrichissant. Sans mon séjour au CIRAD, je n'aurais jamais entendu parler de théorie neutre, des forêts de Boukoko et La Lolé. Je n'aurais jamais vu de pédocomparateur. Je ne saurais pas placer Pointe Noire sur une carte. PCR, QTL et SNP seraient des acronymes vides de sens, *urophylla*, *grandis* et *globulus* de vagues mots à consonances latines. Merci à chacun d'entre vous de m'avoir fait partager les sujets qui lui tenaient à cœur.

J'aimerais de nouveau remercier Sylvie d'avoir fait en sorte que les angéliques et les wacapous ne restent pas seulement des petits points verts et rouges sur mon écran d'ordinateur en m'emmenant en Guyane. Ce fut une aventure formidable. Ma thèse n'aurait peut être pas été tout à fait la même sans ce petit tour à Paracou.

Merci à Guillaume de m'avoir sauvée à plusieurs reprises des affres du C. Merci à Evelyne et à Roselyne de s'être toujours montrées disponibles quand j'avais un petit service à leur demander.

Je ne peux clore cette petite incursion ciradienne sans remercier Ciré. Tout d'abord simples collègues, les longues heures de labeur passées ensemble au bureau ont fait de nous des amis. J'aimerais lui dire au combien j'ai apprécié sa présence à mes côtés, dans les bons comme dans les mauvais moments, tout au long de ces trois années.

Je tiens à remercier l'ensemble des membres du département MMIP d'AgroParisTech et en particulier, ceux de l'UMR MIA 518. Leur accueil sur Paris m'a permis de finir de rédiger ma thèse dans de bonnes conditions. Merci pour vos encouragements et tous les bons conseils que vous m'avez donnés pour la soutenance.

Merci enfin à ma famille et à mes amis, les matheux(ses) et le clan des Montéglésiens. Les moments conviviaux passés avec vous lors de mes retours en Anjou m'ont aidée à tenir le coup. Merci à Hacène d'avoir écouté patiemment mes misères algorithmiques et mes histoires de FFT et surtout d'avoir toujours cru en moi, même de l'autre côté de l'Atlantique. Merci à Caro d'avoir toujours sa maison grande ouverte pour accueillir sa bonne vieille copine. Merci à Tonton Jean et Tante Jeannette d'avoir pris régulièrement des nouvelles de leur Grenouille.

Enfin un merci tout particulier à Blandine et à mes parents pour leur patience, leur soutien et leurs encouragements.

Table des matières

Introduction

Forêts tropicales et biodiversité

A l'heure où la pression humaine sur l'environnement est de plus en plus forte et face aux questions liées aux changements climatiques, le devenir des forêts tropicales est devenu un enjeu majeur. Les forêts tropicales s'étendent sur 1 700 millions d'hectares, soit plus de 10 % des terres émergées. Elles abritent une faune et une flore d'une grande richesse. La forêt tropicale humide, qui ne couvre que 6 % de la planète, renferme à elle seule 50 % à 80 % des espèces animales et végétales terrestres : 80 % des insectes, 84 % des reptiles, 91 % des amphibiens, 90 % des primates, 70 % des espèces végétales connues dont près de 50 000 espèces d'arbres[1]. Les forêts tropicales fournissent une grande diversité de biens (bois, gibier, fruits, médicaments, etc) et de services dont l'homme a largement su tirer parti sans toujours mesurer les risques de dégradation, voire de disparition, qu'il faisait peser sur elles. Au cours des années 1990, environ 15 millions d'hectares de forêts tropicales ont été perdus chaque année[1]. La forêt tropicale humide voit, quant à elle, sa surface diminuer d'environ 0.6 % chaque année (Achard et al., 2002). Depuis le sommet de Rio en 1992, la protection des forêts, des forêts tropicales en particulier, figure parmi les grandes priorités environnementales au niveau mondial. Des stratégies visant à conserver la biodiversité et à exploiter de manière durable les forêts tropicales sont mises en place dans le cadre de la Convention sur la Diversité Biologique[2] et du Forum des nations unies sur les forêts, ainsi qu'au sein de l'Organisation Internationale des Bois Tropicaux[3]. La diversité biologique ou biodiversité désigne, ici, le nombre, la variété des organismes vivants et les relations qu'ils entretiennent entre eux. Elle est généralement décrite suivant trois niveaux d'organisation (Burley, 2002; Levrel, 2007) ; elle comprend :

- la variabilité génétique au sein des populations et entre les populations d'une espèce donnée (étendue et mécanismes de variation des populations, variation des génotypes, des fréquences alléliques, effets et flux de gènes)
- la diversité spécifique des communautés (nombre, abondance ou rareté, endémisme des espèces)
- la diversité fonctionnelle des écosystèmes, c'est-à-dire la variation entre les écosystèmes et la façon dont les espèces intéragissent entre elles et avec leur environnement (composition, structure et fonctionnement des écosystèmes).

[1] Source : www.onf.fr

[2] Le texte complet de la Convention sur la Diversité Biologique peut être consulté à l'adresse suivante : http ://www.cbd.int/doc/legal/cbd-un-fr.pdf

[3] http ://www.itto.int

La modélisation comme outil d'aide à la décision

L'enjeu consiste aujourd'hui à conserver la biodiversité des forêts tropicales et à les gérer durablement, c'est-à-dire à exploiter leurs ressources en préservant à long terme leurs fonctions écologiques, économiques et sociales. Élaborer des règles de gestion compatibles avec un renouvellement des ressources nécessite de mieux comprendre le fonctionnement des écosystèmes forestiers afin de prévoir leur évolution (modification de structure, de composition floristique et génétique). Les tentatives effectuées en matière d'aménagement et de gestion des forêts tropicales se heurtent, en autre, à une compréhension insuffisante des phénomènes qui régissent la dynamique des populations des espèces d'arbres qui les constituent. La prise en compte de la diversité génétique dans les plans d'aménagement forestiers est elle aussi problématique. Ceci est notamment dû à une méconnaissance des interactions existant entre la dynamique démographique d'un peuplement et sa dynamique génétique (Lourmas, 2003). De plus, les effets des perturbations, notamment anthropiques, sur le peuplement sont encore insuffisamment connus. Pour mieux appréhender l'évolution des écosystèmes forestiers tropicaux, des dispositifs expérimentaux ont été mis en place dans la zone intertropicale à partir des années 70 (Gourlet-Fleury et al., 2004). Ces dispositifs de suivi permettent de mesurer l'impact de différentes interventions sylvicoles sur les peuplements (Favrichon, 1997). Leur mise en place doit notamment permettre d'estimer le taux de reconstitution du stock de bois après exploitation avec une précision donnée (Chagneau et al., 2009) pour répondre aux attentes des exploitants forestiers. Cependant, l'étude de la dynamique forestière nécessite souvent des analyses à une échelle de temps et d'espace supérieure à celle qu'il est possible d'observer sur le terrain ; c'est pourquoi il est nécessaire d'avoir recours à la modélisation (Lourmas, 2003). Les modèles forestiers sont calibrés et validés à partir des mesures effectuées sur les dispositifs expérimentaux. Ces modèles permettent non seulement d'intégrer des phénomènes couvrant une large échelle d'espace et de temps, mais aussi de prédire l'évolution des peuplements dans des cas de figure relativement complexes (comme celui des peuplements hétérogènes). Ils rendent également possible une comparaison rapide entre plusieurs scénarios sylvicoles, ce qui aurait demandé auparavant de longues années d'expérimentation (Goreaud et al., 2005). Des logiciels de simulation intégrant ces modèles ont été développés par la communauté scientifique, comme la plateforme Capsis (de Coligny et al., 2004) dédiée à la simulation d'écosystèmes forestiers. Ces outils informatiques ont permis une plus large diffusion de ces modèles. Plusieurs modèles forestiers sont aujourd'hui utilisés par les gestionnaires comme outil d'aide à la décision pour garantir une gestion durable des ressources exploitables.

Les modèles de dynamique forestière

Les modélisateurs forestiers ont développé un grand nombre de modèles pour comprendre l'évolution des arbres et des peuplements. La notion de « modèle » peut prendre différentes significations suivant les disciplines (Pavé, 1994). Le terme de « modèle » désigne ici un ensemble d'équations mathématiques permettant de décrire et de simuler des relations entre les variables d'un peuplement forestier. Les modèles de dynamique forestière peuvent être classés en trois groupes suivant le niveau de description du peuplement sur lequel ils reposent. On distingue (Vanclay, 1995; Franc et al., 2000) :

- *les modèles de population appelés aussi modèles globaux*
  Ces modèles ne mettent en jeu que des variables qui décrivent globalement la population (densité, diamètre moyen). Chaque arbre est considéré comme une réalisation de l'arbre moyen du peuplement. Ces modèles ne prennent pas en compte l'hétérogénéité entre individus au sein de la population. Ils sont particulièrement adaptés pour décrire des peuplements dits « homogènes » ou « réguliers » (c'est-à-dire monospécifiques et équiennes).

- *les modèles de distribution*
  Dans ces modèles, la population n'est plus décrite par une variable moyenne comme dans les modèles globaux, mais est résumée par une fonction de distribution sur une ou plusieurs variables (diamètre, hauteur de l'arbre, etc). La modélisation consiste à suivre l'évolution de cette fonction dans le temps. Dans cette classe de modèles, on différencie plusieurs types de modèles suivant que la fonction de distribution et le temps sont considérés comme discrets ou continus. Les modèles matriciels (Caswell, 2001) qui correspondent à des modèles à espace d'états et temps discret sont les plus utilisés en foresterie.

- *les modèles individuels* (DeAngelis et Gross, 1992)
  La description du peuplement se fait au niveau des individus. La trajectoire de la variable étudiée est suivie pour chaque arbre. Les modèles individuels sont spatialisés ou non. Ce type de modèles permet de prendre en compte l'hétérogénéité au sein du peuplement.

Les modèles individuels sont basés sur trois grandes composantes synthétisant la démographie du peuplement :

- un modèle de croissance intégrant l'effet du milieu et de la compétition (interactions entre arbres),
- un modèle de mortalité décrivant pour chaque individu la probabilité de mourir ou de survivre en fonction de différents facteurs biotiques ou abiotiques,
- un modèle de régénération ou de recrutement décrivant l'apparition de nouveaux individus dans le peuplement.

Alors que les comportements en croissance des espèces tropicales sont relativement bien connus d'un point de vue biologique, les processus de mortalité et de recrutement sont peu documentés. Les processus de mortalité et de recrutement sont plus délicats à modéliser car les pas de temps nécessaires à leur observation ne sont pas du même ordre de grandeur que les durées habituelles de suivi sur les dispositifs expérimentaux (Franc et al., 2000). Une amélioration des modèles de dynamique forestière passe donc par une meilleure compréhension de ces deux phénomènes. Dans ce travail de thèse, nous nous intéressons plus particulièrement au processus de régénération.

La régénération, un phénomène complexe

La régénération désigne l'ensemble des processus allant de la floraison d'un arbre adulte à l'apparition d'un nouvel individu dans le peuplement en passant par la production et la dissémination des graines, la germination et l'établissement d'une plantule autotrophe (Figure 1) (Franc et al., 2000). La régénération est donc un phénomène complexe résultant d'une succession de processus biologiques (Wang et Smith, 2002; Baraloto, 2003). On distingue la régénération du recrutement. Un arbre est dit recruté lorsque sa taille (diamètre ou hauteur) franchit un certain seuil, appelé seuil de recrutement ; il fait dès lors partie du peuplement. Le seuil de recrutement correspond à la taille limite minimale prise en compte dans les inventaires ; il est souvent fixé à 10 cm de diamètre à hauteur de poitrine (« dbh » en abrégé). Modéliser le recrutement consiste à comptabiliser le nombre d'arbres ayant franchi le seuil de recrutement sur une période donnée.

La succession des différents phénomènes constituant le processus de régénération conserve encore des zones d'ombre. Beaucoup d'études ont été réalisées sur la dispersion des graines (Howe et Smallwood, 1982; Clark et al., 1999; Nathan et Muller-Landau, 2000). Cependant, il reste encore difficile d'établir des liens de cause à effet entre le phénomène de dispersion des graines et la répartition spatiale des juvéniles, c'est ce que Houle (1995) nomme les « missing links ». La reconstitution de ces chaînons manquants s'avère difficile en raison de la multiplicité des facteurs biotiques ou abiotiques qui interviennent à tous les stades de la régénération et parce que des mécanismes sous-jacents très différents peuvent conduire à un même type de répartition spatiale sur le terrain. La répartition spatiale observée des juvéniles résulte d'une combinaison de leur dispersion et de leur survie. Cette dernière est difficile à décrire car le taux de survie est généralement un nombre petit résultant du rapport de deux grands nombres : beaucoup de morts parmi beaucoup de graines produites. De plus, la survie fait souvent intervenir des phénomènes rares et non stationnaires dans le temps (perturbations de grande ampleur), de sorte que les espèces se maintiennent sur le long terme bien qu'ayant une régénération erratique sur le court terme.

Étant donné les difficultés rencontrées pour décrire précisément les différents processus intervenant dans la régénération, sa modélisation peut s'avérer délicate. Elle est plus ou moins complexe suivant que le modèle est spatialisé ou non. En effet, dans les modèles spatialisés, les relations entre la répartition spatiale des juvéniles et celle des adultes reproducteurs doivent être prises en considération, ce qui constitue une réelle difficulté. Une solution simplificatrice consiste à modéliser le recrutement plutôt que la régénération (Vanclay, 1992). Ce choix est naturel pour les modèles individuels indépendants des distances. Cependant, la modélisation de la régénération offre plusieurs avantages par rapport à la simple modélisation du recrutement, en particulier, lorsque les modèles sont spatialisés. Les prévisions de l'évolution du peuplement sur le long terme sont plus fiables grâce à la prise en compte de la position des adultes reproducteurs. La modélisation de la régénération permet l'investigation des mécanismes sous-jacents. Elle offre également la possibilité d'établir des liens avec d'autres disciplines comme la zoologie (rôle des disséminateurs du pollen et des prédateurs), la pédologie (étude des effets environnementaux) et la génétique des populations (étude de la diversité génétique des juvéniles).

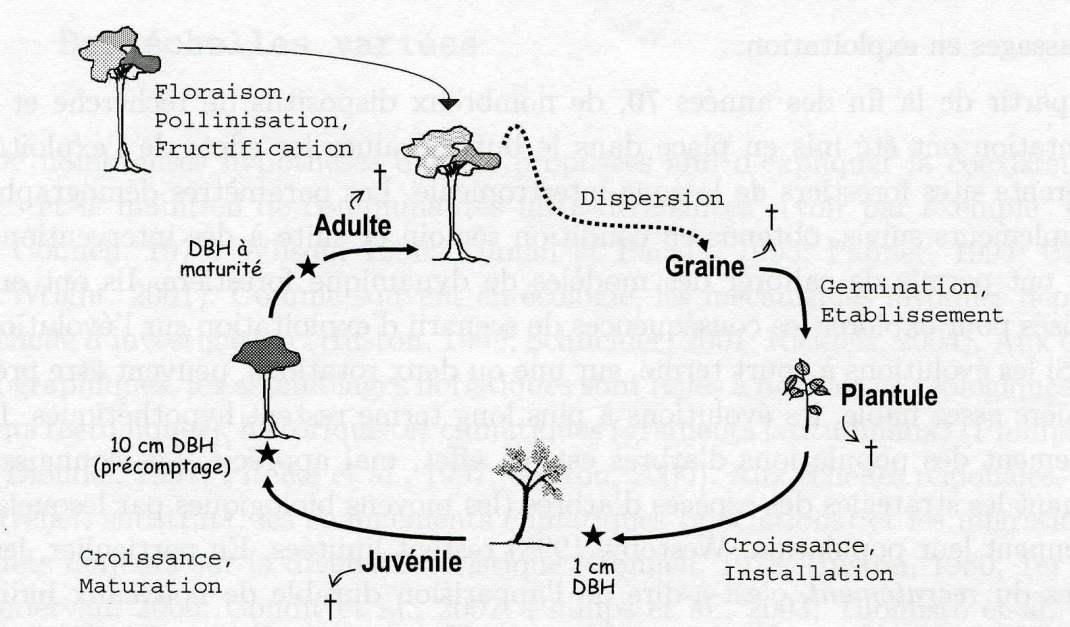


FIG. 1 – Cycle de la vie des arbres extrait de la thèse de Flores (2005). Les graines produites par les adultes sont dispersées dans le milieu. Après germination et épuisement des réserves des graines, les plantules acquièrent l'autotrophie et s'établissent dans le milieu. L'installation correspond au passage à un stade de développement ayant « évacué » les causes de mortalité biotique (prédateurs, pathogènes). Les juvéniles installés se développent ensuite jusqu'à l'acquisition de la capacité de floraison et de fructification (maturation). Le recrutement est l'entrée d'individus dans la population. Empiriquement, il est défini par le passage au dessus d'un diamètre de précomptage à partir duquel on considère les nouveaux individus.

La répartition spatiale des juvéniles et des jeunes adultes installés résultant de la régénération, joue un rôle important dans la dynamique de la population. Plusieurs facteurs sont susceptibles d'expliquer cette répartition. En premier lieu, vient la localisation des adultes reproducteurs, au moins pour les essences à dispersion de graines limitée. Connaissant la position des adultes, les mécanismes de dispersion permettent alors d'expliquer la répartition spatiale des graines dispersées. La répartition spatiale des juvéniles dépend aussi de l'environnement (Getzin et al., 2008). En fonction de leur niche écologique et de leur plasticité, les essences sont susceptibles ou non de s'installer dans différents types d'environnement. L'hétérogénéité spatiale liée aux conditions environnementales affecte la germination des graines, la croissance et la survie des plantules. Elle joue le rôle de filtre et agit différemment suivant les espèces (Beckage et Clark, 2003). Les principaux filtres environnementaux ayant

fait l'objet d'étude sont l'accès à la lumière, l'accès à l'eau (humidité du sol), la quantité de litière, la composition chimique du sol (Clark et al., 1999; Baraloto, 2001; Herrera, 2002; Messaoud et Houle, 2006; Madelaine et al., 2007). Ces facteurs déterminent le nombre de plantules qui vont s'installer et survivre ainsi que leurs localisations. D'autres phénomènes encore sont susceptibles d'agir comme des filtres sur la répartition spatiale des juvéniles : prédateurs, pathogènes, phénomènes de mortalité densité-dépendante, de compétition intra et inter-spécifique (Herrera, 2002). En particulier la combinaison de la dispersion des graines et des phénomènes de mortalité densité-dépendante peut résulter en des patterns non triviaux liant la position des juvéniles à celle de leurs parents (modèle de Janzen-Connell, (Clark et Clark, 1984) ; (HilleRisLambers et al., 2002)). La modélisation de la régénération doit donc prendre en compte l'effet de ces différents facteurs pour prédire la répartition spatiale des juvéniles.

Les études sur la régénération ne sont pas envisageables à grande échelle.

Pour améliorer la compréhension du processus de recrutement, l'idéal serait de pouvoir suivre l'ensemble du peuplement et des semis sur une zone suffisamment grande pendant plusieurs années. On désigne par semis l'ensemble des tiges de 1 à 10 cm de diamètre par opposition à l'ensemble des arbres de plus de 10 cm dbh qui constituent le peuplement. Il faudrait pouvoir répertorier les plantules et les jeunes stades d'arbres de manière exhaustive. En pratique, les données d'inventaire recueillies pour l'étude du recrutement le sont suivant deux méthodes, par recensement ou par échantillonnage. La première méthode est limitée à de petites surfaces, pour des questions de temps et d'argent. Pour donner un ordre de grandeur, en Guyane vénézuélienne, on a trouvé, en moyenne, environ 100 000 semis supérieurs à 10 cm de haut à l'hectare, toutes espèces confondues (Rollet, 1969). Il est donc hors de question d'inventorier autant de semis sur de grandes surfaces. La deuxième méthode permet d'obtenir de l'information sur de plus grandes surfaces, mais ne permet pas une étude détaillée des phénomènes sous-jacents à la régénération. Les nouvelles techniques (marquages radioactifs des graines, marqueurs moléculaires) employées pour l'étude de la dispersion des graines (Wang et Smith, 2002), ne peuvent être mises en place que sur un nombre limité d'individus, étant donné leur coût et l'effort d'échantillonnage qu'elles demandent. L'étude de l'impact des facteurs environnementaux sur la régénération nécessite elle aussi un effort d'échantillonnage important. La description de ces facteurs doit être suffisamment fine pour refléter leur structure spatiale. Pour toutes ces raisons, une étude approfondie du processus de régénération n'est pas envisageable à grande échelle. Seule la modélisation rend possible une prédiction de la régénération, à une plus grande échelle d'espace et de temps, à partir d'un échantillonnage raisonnable du peuplement, des semis et de l'environnement.

Objectifs

Bien qu'il joue un rôle prépondérant dans la dynamique des populations, le processus de régénération reste, de par sa complexité, un des points faibles des modèles de dynamique forestière. La mise en place d'études qui permettraient d'envisager des investigations plus poussées sur la régénération se heurte aux problèmes liés au temps et au coût d'échantillon-

nage. L'objectif de ce travail de thèse est de proposer un modèle mathématique permettant de prédire la régénération à grande échelle à partir d'un échantillonnage raisonnable des adultes, des juvéniles et de l'environnement. Le modèle sera appliqué à la prédiction de la régénération en forêt tropicale. Ce travail fait suite à une étude réalisée par Flores (2005) sur le déterminisme de la régénération de 15 espèces d'arbres tropicaux en forêt guyanaise, dans laquelle ce dernier a montré l'influence des conditions du milieu et des distances aux arbres adultes sur le succès d'installation des juvéniles.

Dans le travail de Flores (2005), la mise en évidence des conditions du milieu sur la répartition spatiale des juvéniles a été effectuée à partir d'une description très fine des facteurs environnementaux. Dans ce cas, l'environnement est « connu » sur toute la zone d'étude. En pratique, étant donné les coûts d'échantillonnage, certaines variables caractérisant le milieu ne sont échantillonnées qu'en un nombre limité de sites. Elles ne sont connues que partiellement, ce qui rend difficile, voire impossible, l'étude de leurs effets sur la régénération. Une solution consiste à extrapoler l'environnement, c'est-à-dire à prédire les variables dans les zones non échantillonnées, avant de modéliser la régénération.

De la prédiction de l'environnement va découler notre première question de recherche. L'environnement est caractérisé par plusieurs variables et ces variables n'ont été mesurées qu'en quelques points. Pour utiliser au mieux l'information disponible, il semble naturel de ne pas traiter chaque variable séparément, mais de se placer dans un cadre multivarié permettant de prendre en compte la structure de dépendance entre les variables. Or toutes les variables environnementales ne sont pas de même nature. Certaines sont continues comme la pente, l'altitude ou la teneur du sol en minéraux, d'autres sont discrètes comme le drainage ou la couleur du sol. Les méthodes de prédiction classiquement utilisées en géostatistique ne peuvent donc pas s'appliquer. La première partie de ce travail a pour objectif de proposer un modèle spatial multivarié permettant de prédire simultanément des variables de nature différente. D'un point de vue mathématique, ce problème soulève deux difficultés : la prédiction de variables discrètes dans le cas multivarié et la modélisation de la dépendance entre variables discrètes et continues.

La seconde partie de ce travail est consacrée à la modélisation de la régénération. L'objectif est de proposer un modèle de régénération spatialement explicite qui, en se basant sur un échantillonnage raisonnable des juvéniles, permette d'étudier l'effet des conditions du milieu sur la régénération. Alors que l'étude sur la régénération réalisée par Flores ne s'intéresse qu'à l'organisation spatiale des juvéniles et n'intègre aucune information génétique, nous souhaitons ici pouvoir prédire non seulement la répartition spatiale mais aussi le génotype des juvéniles. Intégrer de l'information génétique à la modélisation de la régénération semble pertinent à plusieurs titres. Cela va permettre de décrire la diversité génétique des juvéniles, d'améliorer la compréhension des mécanismes de dispersion et d'envisager, à terme, l'étude des interactions génotype-environnement sur la régénération. L'étape qui consiste à prédire l'environnement avant de modéliser la régénération crée une difficulté supplémentaire car elle est source de variabilité. Les variables environnementales prédites sont entachées d'erreur, alors qu'elles sont considérées comme connues dans les modèles de régénération existants. Ces erreurs vont se répercuter sur l'estimation des paramètres du modèle de régénération. L'impact des erreurs de prédiction des variables environnementales sur les éléments qui caractérisent la régénération doit donc être considéré.

Plan du mémoire

Ce mémoire comprend quatre chapitres et deux annexes.

Le chapitre 1 est consacré à la prédiction de l'environnement. Après avoir fait un tour d'horizon des principales méthodes de prédiction existantes, nous proposons un modèle hiérarchique spatial multivarié permettant de prédire simultanément des variables de nature différente.

Le chapitre 2 traite de l'inférence du modèle hiérarchique spatial multivarié. L'analyse bayésienne pour l'estimation des paramètres du modèle et la méthode de prédiction sont présentées avant de valider le modèle par simulation.

La modélisation de la régénération est abordée au chapitre 3. Un état des lieux des modèles de recrutement intégrant de l'information génétique est exposé. Nous étendons un de ces modèles pour prendre en compte les variables environnementales comme variables explicatives de la survie des juvéniles. Une étude par simulations est réalisée. Pour clore ce chapitre, nous discutons de la prise en compte de la variabilité liée à la prédiction de l'environnement.

Le chapitre 4 est dédié aux applications. Les données pédologiques, démographiques et génétiques dont nous disposons pour cette étude sont brièvement décrites. Après avoir prédit l'environnement sur la partie sud du dispositif, nous nous intéressons à la régénération d'une essence forestière tropicale, l'angélique (*Dicorynia guianensis* Amshoff).

En conclusion, nous récapitulerons les principaux résultats obtenus et nous dégagerons quelques perspectives auxquelles ce travail peut donner suite.

L'annexe A présente une extension possible du modèle hiérarchique spatial multivarié au cas de variables nominales. L'annexe B apporte un complément d'information sur le calcul des probabilités intervenant dans la définition de la fonction d'intensité du processus ponctuel modélisant la répartition spatiale des juvéniles.

Chapitre 1

Modèle hiérarchique spatial multivarié pour la prédiction de l'environnement

Sommaire

1.1 Problématique

Nous souhaitons reconstruire un environnement « réaliste » à grande échelle pour étudier l'effet de l'hétérogénéité spatiale sur le processus de régénération. Plusieurs variables ont été mesurées pour caractériser l'environnement : altitude, pente, drainage du sol, teneur du sol en minéraux, couleur du sol, etc. Ces variables ont été échantillonnées de manière aléatoire et en un nombre limité de sites. En chaque point d'échantillonnage, toutes les variables ont été mesurées ; les données dont nous disposons sont isotopes. Bien que toutes ces variables ne soient pas indépendantes, nous pourrions envisager de prédire chacune d'elles séparément. Mais, étant donné la quantité limitée de données disponibles, il semble préférable de se placer dans un cadre multivarié pour pouvoir prendre en compte la dépendance entre les variables, et ainsi obtenir un maximum d'information des données recueillies. L'environnement sera donc représenté par un champ spatial multivarié. Le problème de la prédiction de champs spatiaux multivariés concerne de nombreux domaines : pédologie (McBratney et al., 2000), épidémiologie (Golam Kibria et al., 2002), économie (Chica-Olmo, 2007; Gelfand et al., 2007). Ici, le champ multivarié a la particularité d'être constitué de variables de différente nature. Certaines variables comme la teneur du sol en minéraux sont continues, d'autres comme le drainage sont ordinales, d'autres encore comme la couleur du sol sont nominales. Dans certains cas, le champ spatial considéré peut également comporter des variables de comptage. Comme nous le verrons dans la partie 1.2.1, les méthodes classiques de prédiction ne s'appliquent pas dans de telles situations. Comment prédire alors un champ spatial multivarié composé de variables de nature différente ? C'est à cette question que nous tentons de répondre dans ce premier chapitre. Ce problème soulève deux difficultés majeures : la prédiction des variables discrètes dans le cas multivarié et la modélisation de la dépendance entre les variables discrètes et les variables continues.

1.2 État de l'art

1.2.1 Méthodes de prédiction

La méthode de prédiction la plus utilisée en géostatistique est le krigeage. Développée par Matheron (1963) à partir des travaux de Krige (1951), le krigeage est une méthode stochastique d'interpolation spatiale qui prédit la valeur d'une variable en des sites non échantillonnés par une combinaison linéaire sans biais et de variance minimale des observations de cette variable en des sites voisins (Baillargeon, 2005). Le krigeage dit ordinaire concerne les champs aléatoires stationnaires d'ordre deux, d'espérance finie inconnue.

Définition 1 Stationnarité d'ordre deux
Un champ aléatoire $Z(.)$ est dit ***stationnaire d'ordre deux*** *(Journel et Huijbregts, 1978), si ces deux premiers moments existent et sont invariants par translation, c'est-à-dire :*

(i) que l'espérance mathématique $\mathbb{E}[Z(\mathbf{s})]$ existe et ne dépende pas de $\mathbf{s}$; ainsi

$$\mathbb{E}[Z(\mathbf{s})] = m, \forall \mathbf{s} \in \mathbb{R}^2$$

(ii) et que, pour tout couple de variables aléatoires $(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h}))$*, la covariance existe et ne dépende que de* $\mathbf{h}$ *:*

$$\mathrm{Cov}(\mathbf{h}) = \mathbb{E}[Z(\mathbf{s})Z(\mathbf{s}+\mathbf{h})] - m^2,\ \forall \mathbf{s} \in \mathbb{R}^2, \forall \mathbf{h} \in \mathbb{R}^2.$$

La stationnarité de la covariance entraîne la stationnarité du variogramme :

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathbb{E}[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2] = \mathrm{Cov}(\mathbf{0}) - \mathrm{Cov}(\mathbf{h}),\ \forall \mathbf{s} \in \mathbb{R}^2, \forall \mathbf{h} \in \mathbb{R}^2.$$

Les hypothèses de stationnarité portent sur les moments de la distribution de la variable d'intérêt et non sur la loi de distribution elle-même. On peut donc croire que la qualité des prédictions est indépendante de la loi de distribution des données, ce qui n'est pas le cas. Rappelons que la prédiction par krigeage est linéaire, sans biais et qu'elle minimise la variance de prédiction. Une prévision non contrainte à être linéaire et minimisant l'erreur quadratique moyenne serait l'espérance conditionnelle $\mathbb{E}[Z_0|Z_1, \ldots, Z_n]$ où Z_i, $i = 0, 1, \ldots, n$ désigne la variable aléatoire Z considérée au site $\mathbf{s}_i$. Théoriquement, cette prédiction spatiale est meilleure ou équivalente à celle obtenue par krigeage. Cependant, elle dépend de la loi de la distribution de $Z(.)$. Si $Z(.)$ est un champ gaussien, l'espérance conditionnelle est une combinaison linéaire des données Z_i. L'hypothèse de linéarité est donc appropriée dans le cas gaussien (Chilès et Delfiner, 1999; Baillargeon, 2005). Pour les autres variables, le krigeage reste le meilleur estimateur linéaire, mais utiliser un estimateur linéaire ne conduit pas nécessairement à des prédictions de bonne qualité. Des méthodes plus appropriées ont donc été mises en place pour traiter des données non gaussiennes, et notamment des variables discrètes (variables ordinales, nominales, de comptage).

Quelques méthodes de prédiction univariées pour des variables non gaussiennes

Méthodes de géostatistique non-linéaire Nous cherchons, en général, à déterminer une prédiction de Z_0, Z_0^*, qui minimise l'erreur de prédiction. Si l'erreur de prédiction considérée est l'erreur quadratique moyenne, le meilleur prédicteur est alors l'espérance conditionnelle de Z_0^* connaissant les données :

$$Z_0^* = \mathbb{E}[Z_0|Z_i,\ i = 1, \ldots, n].$$

L'espérance conditionnelle ne peut pas, en général, s'exprimer comme combinaison linéaire des données, le cas gaussien mis à part, et un calcul direct de l'espérance conditionnelle est rarement envisageable car il nécessite de connaître la distribution du $n+1$ uplet $(Z_0, Z_1, \ldots, Z_n)$. Des méthodes dites de géostatistique non linéaire ont été développées, visant à approcher cette espérance conditionnelle par des fonctions non linéaires des données (Journel et Huijbregts, 1978; Cressie, 1991). C'est, par exemple, le cas du krigeage disjonctif. La mise en oeuvre de ces méthodes est, en général, plus complexe que celles relevant de la géostatistique linéaire.

• *Krigeage disjonctif*

Contrairement au krigeage où le prédicteur est construit comme combinaison linéaire des données ($Z_K^* = \sum_{i=1}^n \lambda_i Z_i$), le krigeage disjonctif (Matheron, 1973; Armstrong et

Matheron, 1986a,b; Chilès et Delfiner, 1999) cherche à déterminer le meilleur prédicteur s'écrivant comme une combinaison linéaire de fonctions univariées des données, ainsi

$$Z^*_{KD} = \sum_{i=1}^{n} f_i(Z_i),$$

où les fonctions f_i sont des fonctions mesurables de carré intégrable.

Les équations du krigeage disjonctif sont déterminées en considérant le krigeage disjonctif en terme de projection (Journel et Huijbregts, 1978). On suppose que les variables aléatoires $Z_i, i = 1, \ldots, n$ et Z_0 admettent un moment d'ordre 2, autrement dit, elles appartiennent à un espace de Hilbert muni d'un produit scalaire $< X, Y >= \mathbb{E}[XY]$. Soit H_i le sous-espace engendré par les variables de la forme $f_i(Z_i)$ ayant un moment d'ordre 2 fini. Soit H le sous-espace engendré par les fonctions aléatoires de la forme $\sum_{i=1}^{n} f_i(Z_i)$. Le krigeage disjonctif de Z_0, Z^*_{KD}, est la projection de Z_0 sur le sous-espace H. L'estimateur Z^*_{KD} vérifie les propriétés suivantes :

(i) $Z^*_{KD} \in H$ i.e. $Z^*_{KD} = \sum_{i=1}^{n} f_i(Z_i)$,

(ii) le vecteur $Z^*_{KD} - Z_0$ est orthogonal à tout vecteur Y de H :

$$< Z^*_{KD} - Z_0, Y >= 0, \forall Y \in H. \tag{1.1}$$

Cette égalité est vraie pour toute variable Y de H et donc, en particulier, pour toute variable Y de H_i. Par linéarité du produit scalaire, l'équation 1.1 s'écrit sous la forme :

$$< Z^*_{KD}, Y >=< Z_0, Y >$$

pour tout Y de la forme $f_i(Z_i)$. Cette équation est équivalente à :

$$\mathbb{E}[Z^*_{KD}|Z_j] = \mathbb{E}[Z_0|Z_j],\ j = 1, \ldots, n$$

(La preuve est donnée dans le livre de Journel et Huijbregts (1978), page 569). Par conséquent, les fonctions f_i intervenant dans la définition de Z^*_{KD} sont caractérisées par le système d'équations suivant :

$$\sum_{i=1}^{n} \mathbb{E}[f_i(Z_i)|Z_j] = \mathbb{E}[Z_0|Z_j],\ j = 1, \ldots, n. \tag{1.2}$$

Le krigeage disjonctif peut aussi être utilisé pour estimer n'importe quelle fonction mesurable $\varphi(Z_0)$ d'inconnue Z_0. Cette fonction $\varphi(Z_0)$ est projetée sur H. L'estimateur $\varphi^*_{KD} = \sum_{i=1}^{n} f_i(Z_i)$ est alors caractérisé par le système suivant :

$$\sum_{i=1}^{n} \mathbb{E}[f_i(Z_i)|Z_j] = \mathbb{E}[\varphi(Z_0)|Z_j],\ j = 1, \ldots, n.$$

L'estimateur du krigeage disjonctif est un moins bon estimateur que l'espérance conditionnelle, mais sa détermination est plus simple. En effet, la résolution du système 1.2

vérifié par les f_i nécessite seulement de connaître les distributions bivariées de tous les couples (Z_i, Z_j) et (Z_0, Z_i), alors que la détermination de l'espérance conditionnelle nécessite de connaître la distribution du $n+1$ uplet de variables $(Z_0, Z_1, \ldots, Z_n)$. La taille du système d'équations 1.2 pour déterminer Z^*_{KD} croît avec le nombre de données ; sa résolution peut donc rapidement devenir compliquée. Il existe des champs aléatoires pour lesquels la résolution du système se simplifie ; ce sont les champs dont les distributions bivariées sont isofactorielles.

Définition 2 Distributions isofactorielles (Chilès et Delfiner, 1999)
Les distributions bivariées d'un champ aléatoire $Z(.)$ de distribution marginale $G(dz)$ sont dites isofactorielles si, pour tout couple de sites $\mathbf{s}$ et $\mathbf{s}+\mathbf{h}$, la distribution bivariée du couple $(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h}))$ peut être factorisée sous la forme

$$G_{\mathbf{h}}(dz, dz') = \sum_{q=0}^{+\infty} T_q(\mathbf{h})\chi_q(z)\chi_q(z')G(dz)G(dz')$$

où $\{\chi_q, q = 0, 1, \ldots\}$ est une base hilbertienne de l'espace $L^2(G)$ et où $T_q(\mathbf{h})$ est la covariance du champ aléatoire $\chi_q(Z(.))$, c'est-à-dire que $T_q(\mathbf{h}) = \mathbb{E}[\chi_q(Z(\boldsymbol{s})\chi_q(Z(\boldsymbol{s}+\boldsymbol{h}))]$. Les facteurs χ_q sont les mêmes quel que soit $\mathbf{h}$ d'où la dénomination « isofactorielle ».

Les plus connus de ces champs sont ceux ayant des distributions bivariées gaussiennes. La résolution du système 1.2 se simplifie également pour tous les champs $Z(.)$ qui s'écrivent comme la transformation d'un champ aléatoire stationnaire $Y(.)$ de distributions bivariées isofactorielles.

Supposons que le champ aléatoire d'intérêt $Z(.)$ ait des distributions bivariées isofactorielles. Les variables Z_0 et $f_i(Z_i)$, $i = 1, \ldots, n$ se décomposent dans la base hilbertiennne correspondante $\{\chi_q, q = 0, 1, \ldots\}$ sous la forme :

$$\begin{aligned} Z_0 &= \sum_{q=0}^{+\infty} \psi_q \chi_q(Z_0), \\ f_i(Z_i) &= \sum_{q=0}^{+\infty} f_{iq}\chi_q(Z_i),\ i = 1, \ldots, n. \end{aligned}$$

Les coefficients ψ_q sont connus, alors que les coefficients f_{iq} sont à déterminer. En utilisant les propriétés de l'espérance conditionnelle et des bases hilbertiennes (Cressie, 1991; Chilès et Delfiner, 1999), on peut montrer que l'estimateur du krigeage disjonctif Z^*_{KD} est égal à :

$$Z^*_{KD} = \sum_{i=1}^{n}\sum_{q=0}^{+\infty} f_{iq}\chi_q(Z_i)$$

où les coefficients $\{f_{iq},\ i = 1, \ldots, n\}$ sont les solutions du système :

$$\sum_{j=1}^{n} f_{jq}T_q(\mathbf{s}_j - \mathbf{s}_i) = \psi_q T_q(\mathbf{s}_0 - \mathbf{s}_i),\ q = 0, 1, \ldots,\ \text{et } i = 1, \ldots, n.$$

Dans le cas d'un modèle isofactoriel, déterminer l'estimateur du krigeage disjonctif se ramène à la résolution de systèmes du même type que celui du krigeage. On pourra trouver des exemples de mise en oeuvre du krigeage disjonctif dans les articles de Webster et Oliver (1989) et de von Steiger et al. (1996).

Ainsi le krigeage disjonctif permet-il de prédire toutes les variables non gaussiennes qui peuvent s'écrire comme la transformation d'un champ dont les distributions bivariées sont isofactorielles.

• *Krigeage d'indicatrices*

Dans certains domaines, on ne s'intéresse pas vraiment à l'estimation de la valeur de la variable d'intérêt, mais plutôt à la probabilité que cette valeur soit inférieure ou supérieure à un seuil donné (Rivoirard, 1991) ; c'est par exemple le cas en science de l'environnement (dépassement d'un seuil de concentration en polluant (Lark et Ferguson, 2004)), en agriculture (excès ou déficit en élément nutritif (Atkinson et Lloyd, 2001)), etc. Le krigeage d'indicatrices a été développé pour répondre à ce type de problème. Nous souhaitons estimer $I_0(z) = \mathbb{1}_{\{Z_0<z\}}$ à partir des données $\{Z_i,\, i = 1, \ldots, n\}$. Le krigeage d'indicatrices consiste à estimer $I_0(z)$ en krigeant le champ aléatoire d'indicatrices $\mathbb{1}_{\{Z(s)<z\}}$ correspondant aux données (Journel et Huijbregts, 1978; Cressie, 1991; Chilès et Delfiner, 1999). Le processus se divise en deux étapes :

(i) le remplacement des données initiales Z_i par des données sous forme d'indicatrices $\mathbb{1}_{\{Z_i<z\}}$,

(ii) le krigeage ordinaire des données binaires obtenues.

L'estimateur s'écrit donc :

$$I_0^*(z) = \sum_{i=1}^{n} \lambda_i(z) \mathbb{1}_{\{Z_i<z\}},$$

avec la condition de non biais $\sum_{i=1}^{n} \lambda_i(z) = 1$ et les poids $\{\lambda_i(z),\, i = 1, \ldots, n\}$ qui satisfont le système

$$\sum_{i=1}^{n} \lambda_i(z)\gamma_z(\mathbf{s}_i - \mathbf{s}_j) + m(z) = \gamma_z(\mathbf{s}_0 - \mathbf{s}_j)\; j = 1, \ldots, n,\; z \in \mathbb{R}.$$

$2\gamma_z(.)$ désigne le variogramme d'indicatrices. Il est indépendant de $\mathbf{s}$ et est défini par :

$$2\gamma_z(\mathbf{h}) \equiv \mathrm{Var}(\mathbb{1}_{\{Z(\mathbf{s}+\mathbf{h})<z\}} - \mathbb{1}_{\{Z(\mathbf{s})<z\}}],\; \mathbf{h} \in \mathbb{R}^2,\; z \in \mathbb{R}.$$

Cette méthode nécessite de modéliser autant de variogrammes et de résoudre autant de systèmes d'équations que de seuils considérés. De plus, elle ne garantit pas d'obtenir une estimation comprise entre 0 et 1.

Le krigeage d'indicatrices permet, entre autres, pour les variables ordinales et nominales, de déterminer la probabilité d'observer chacune des modalités en un point s_0 non échantillonné (Miller et Franklin, 2002; Bourrennane et al., 2003; D'Or et Bogaert, 2004).

En pratique, le choix entre krigeage disjonctif et krigeage d'indicatrices est souvent dicté par des considérations pratiques plutôt que théoriques (Lark et Ferguson, 2004). En général, le krigeage disjonctif, ou toutes les autres techniques de krigeage non linéaires

qui utilisent une transformation continue des données, donnent des résultats plus précis en terme de prédiction que le krigeage d'indicatrices pour lequel le seuillage engendre une perte d'information importante par rapport aux données initiales.

Modèles linéaires généralisés spatiaux La modélisation des problèmes géostatistiques proposée par Diggle et al. (1998), appelée *model-based geostatistics*, repose sur le principe suivant : les données sont considérées comme issues d'un modèle stochastique connu explicitement. Les méthodes classiques de la statistique peuvent donc être appliquées à ce modèle. Comme nous l'avons vu précédemment, dans les modèles de géostatistique classiques, les hypothèses concernant la distribution des données ne sont, en général, pas explicites.

Pour le krigeage, par exemple, l'estimateur S_0^* de la variable Y au point $\mathbf{s}_0$ est un combinaison linéaire des données :

$$S_0^* = \sum_{i=1}^{n} \lambda_i y_i$$

où y_i désigne la valeur observée de la variable Y au point $\mathbf{s}_i$. On note respectivement Y_i et S_i les variables $Y(\mathbf{s}_i)$ et $S(\mathbf{s}_i)$. On considère que les données ont été générées par le modèle suivant :

$$Y_i = \mu + S_i + \epsilon_i,\ i = 1, \ldots, n$$

où μ est un effet moyen constant, $S(.)$ est un processus stationnaire gaussien tel que $\mathbb{E}[S(\mathbf{s})] = 0$ et $\mathrm{Cov}[S(\mathbf{s}), S(\mathbf{s}+\mathbf{h})] = \sigma^2\rho(\mathbf{h})$ et où les résidus ϵ_i sont mutuellement indépendants de loi $\mathcal{N}(0, \tau^2)$. $\mathcal{N}(m, \sigma^2)$ désigne la loi normale de moyenne m et de variance σ^2. Autrement dit, conditionnellement à $S(.)$, les Y_i sont mutuellement indépendantes, avec

$$Y_i | S_i \sim \mathcal{N}(\mu + S_i, \tau^2).$$

Dans cette approche, la méthodologie du krigeage est intégrée dans un ensemble plus large de méthodes basées sur la connaissance des distributions, de la même manière que le modèle linéaire gaussien est intégré dans la famille plus large des modèles linéaires généralisés. On suppose que le modèle dont sont issues les données vérifie les hypothèses suivantes :

1. S est un processus stationnaire gaussien avec $\mathbb{E}[S(\mathbf{s})] = 0$ et $\mathrm{Cov}[S(\mathbf{s}), S(\mathbf{s}+\mathbf{h})] = \sigma^2\rho(\mathbf{h})$;
2. conditionnellement à $S(.)$, les variables aléatoires Y_i, $i = 1, \ldots, n$ sont mutuellement indépendantes de densité $f_i(y|S_i) \equiv f(y; M_i)$ ne dépendant que des valeurs des espérances conditionnelles $M_i = \mathbb{E}(Y_i|S_i)$;
3. $l(M_i) = S_i + \mathbf{d}_i{}'\boldsymbol{\beta}$, pour une fonction de lien l, des variables explicatives $\mathbf{d}_i = \mathbf{d}(\mathbf{s}_i)$ et des paramètres $\boldsymbol{\beta}$.

Sous ces hypothèses, le prédicteur de $S(\mathbf{s})$, appelé prédicteur linéaire généralisé, est défini par $S^*(\mathbf{s}) = \mathbb{E}[S(\mathbf{s})|Y]$.

Soit $\mathbf{S} = (S(\mathbf{s}_1), \ldots, S(\mathbf{s}_n))'$ le vecteur des valeurs de la variable S aux n points échantillonnés et $\mathbf{S}^* = (S(\mathbf{s}_1^*), \ldots, S(\mathbf{s}_m^*))'$ le vecteur des valeurs de la variable S aux m points où l'on souhaite effectuer des prédictions. Dans la suite, on désigne par $\mathbf{s}_i, i = 1, \ldots, n+m$ le $i^{\text{ème}}$ élément du vecteur $(\mathbf{S}, \mathbf{S}^*)'$. Soit $g_k(\mathbf{s}_1, \ldots, \mathbf{s}_k)$ la densité des k premières composantes du vecteur $(\mathbf{S}, \mathbf{S}^*)'$. Cette densité est celle d'une loi normale multivariée.

Le prédicteur linéaire généralisé et sa variance au point $\mathbf{s}_j^*$ sont donnés par :

$$S_j^* = \frac{E_{1,j}}{f(\mathbf{y})}$$

et

$$V_j^* = \frac{E_{2,j}}{f(\mathbf{y})}$$

où, pour tout $r \in \mathbb{N}^*$ et $j = 1, \ldots, m$,

$$E_{r,j} = \int \mathbf{s}_{n+j}^r \left(\prod_{i=1}^{n} f_i(y_i|\mathbf{s}_i) \right) g_{n+m}(\mathbf{s}_1, \ldots, \mathbf{s}_{n+m}) \mathrm{d}\mathbf{s} \mathrm{d}\mathbf{s}_{n+j},$$

$$f(\mathbf{y}) = \int \left(\prod_{i=1}^{n} f_i(y_i|\mathbf{s}_i) \right) g_n(\mathbf{s}_1, \ldots, \mathbf{s}_n) \mathrm{d}\mathbf{s}$$

et $\mathrm{d}\mathbf{s} = \prod_{i=1}^{n} \mathrm{d}\mathbf{s}_i$.

Étant donné leur dimension, les intégrales intervenant dans la prédiction ne peuvent pas être calculées explicitement et doivent donc être approchées. Sous les hypothèses 1 à 3, la distribution de S et la distribution de Y sachant S sont connues. Il est donc possible d'utiliser des méthodes de Monte Carlo par chaînes de Markov (MCMC) pour simuler des réalisations de la distribution conditionnelle de S sachant Y et ainsi estimer les espérances associées à la distribution conditionnelle (paragraphe 2.4).

Les modèles linéaires généralisés spatiaux permettent donc de prédire, dans un cadre univarié, tout type de variables qui pourrait être modélisé par un modèle linéaire généralisé dans un contexte non spatial (Banerjee et al., 2004) : variable gaussienne, variable de Poisson (Christensen et Waagepetersen, 2002), variable ordinale, etc. Nous reviendrons plus en détails sur le cas des variables ordinales dans la partie 1.3.

Méthodes de prédiction multivariées

Lorsque les données à traiter sont multivariées, il est bien sûr possible de prédire chaque variable séparément. Mais il est préférable d'adopter des méthodes de prédiction spécifiques permettant de prendre en compte les relations existants entre les variables, pour améliorer la qualité des prédictions. Suivant la nature des variables, différentes méthodes peuvent être utilisées.

Le cokrigeage Le cokrigeage est une extension du krigeage au cas multivarié (Chilès et Delfiner, 1999; Wackernagel, 2003). Il permet notamment de prédire simultanément plusieurs variables gaussiennes.

Comme pour le krigeage, il existe différents types de cokrigeage. Nous nous intéressons ici au cokrigeage ordinaire, c'est-à-dire à un cokrigeage de processus stationnaires d'ordre deux, d'espérance finie inconnue.

Soient $Z_1(.), Z_2(.), \ldots, Z_K(.)$ K champs aléatoires stationnaires. Chaque variable Z_k, $k = 1, \ldots, K$ a été mesurée en n_k points. L'objectif est de prédire la variable Z_m en un point $\mathbf{s}_0$ du domaine en tenant compte des informations données par les autres champs aléatoires. L'indice m désigne une variable particulière dans l'ensemble des K variables.

L'estimateur de cokrigeage ordinaire est une combinaison linéaire des variables localisées en différents points d'échantillonnage dans le voisinage de s_0 :

$$Z_m^*(\mathbf{s}_0) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \lambda_i^k Z_k(\mathbf{s}_i).$$

La condition de non-biais nous donne :

$$\sum_{i=1}^{n_k} \lambda_i^k = \delta_{km} = \begin{cases} 1 & \text{si } k = m \\ 0 & \text{sinon} \end{cases} .$$

L'estimateur du cokrigeage est l'estimateur pour lequel la variance de l'erreur de prédiction est minimale compte tenu des deux conditions ci-dessus. Cela conduit à la résolution du système linéaire d'équations suivant, où les ζ_k sont des multiplicateurs de Lagrange :

$$\begin{cases} \forall k, k = 1, \ldots, K, \forall i, i = 1, ..., n_k \\ \sum_{p=1}^{K} \sum_{j=1}^{n_p} \lambda_j^p \operatorname{Cov}[Z_k(\mathbf{s}_i), Z_p(\mathbf{s}_j)] + \zeta_k = \operatorname{Cov}[Z_m(\mathbf{s}_0), Z_k(\mathbf{s}_i)] \\ \\ \sum_{j=1}^{n_k} \lambda_j^k = \delta_{km}, \ \forall k, k = 1, ..., K \end{cases}$$

La variance de l'erreur ou variance de cokrigeage peut s'écrire :

$$\sigma_{CKO}^2 = \operatorname{Cov}[Z_m(\mathbf{s}_0), Z_m(\mathbf{s}_0)] + \sum_{k=1}^{K} \sum_{i=1}^{n_k} \lambda_i^k \operatorname{Cov}[Z_k(\mathbf{s}_i), Z_m(\mathbf{s}_0)].$$

Le cokrigeage comme le krigeage ne manipule que des combinaisons linéaires des variables étudiées. Comme nous l'avons vu précédemment, considérer des combinaisons linéaires des données n'est pas toujours satisfaisant, en particulier lorsque les variables d'intérêt ne sont pas gaussiennes. Il est alors possible d'avoir recours à des méthodes de prédiction multivariées non linéaires comme le cokrigeage disjonctif.

Le cokrigeage disjonctif Le cokrigeage disjonctif est la généralisation du krigeage disjonctif au cas multivarié. La généralisation est simple à condition que l'on puisse déterminer un ensemble de distributions bivariées cohérent qui permette de simplifier les systèmes d'équations à résoudre (Chilès et Delfiner, 1999). La mise en œuvre du cokrigeage disjonctif reste fastidieuse, il est donc peu utilisé.

Méthode BME (Bayesian Maximum Entropy) Alors que les méthodes classiques de prédiction ne permettent de prendre en compte que les données observées, l'approche BME développée par Christakos (1990), permet d'intégrer la connaissance générale G (lois physiques, relation empirique, moments statistiques à tous les ordres, etc) que l'on a sur la variable étudiée et la connaissance spécifique aux sites S (mesures réelles, observations incertaines, information secondaire) pour prédire la densité *a posteriori* en tout point $\mathbf{s}_0$. Le principe de la méthode est expliqué dans le cadre univarié.

Soit $Y(.)$ un champ aléatoire. On note Y_i la variable aléatoire $Y(\mathbf{s}_i)$ et y_i une de ses réalisations. Soit Y_0^* l'estimateur de $Y(\mathbf{s}_0)$ en un point non échantillonné $\mathbf{s}_0$. Le vecteur des variables aléatoires $Y_{\text{map}} = (Y_1, \ldots, Y_{n+m}, Y_0)$ représente le vecteur aléatoire en tous les points $\mathbf{s}_{\text{map}} = (\mathbf{s}_{\text{données}}, \mathbf{s}_0)$, où $\mathbf{s}_{\text{données}} = (\mathbf{s}_1, \ldots, \mathbf{s}_{n+m})$ sont les points pour lesquels on dispose d'information et $\mathbf{s}_0$ un point à prédire. On note y_{map} une réalisation de Y_{map}.

Une première étape consiste à déterminer une fonction de densité f_G *a priori* qui soit la plus générale possible tout en vérifiant différentes contraintes imposées par la connaissance générale G que l'on a sur le phénomène. La connaissance générale G consiste le plus souvent à donner la moyenne et la covariance de la variable étudiée. La fonction f_G est obtenue par maximisation de l'entropie :

$$H(f_G(y_{\text{map}})) = -\int \ln[f_G(y_{\text{map}})] f_G(y_{\text{map}}) \mathrm{d}y_{\text{map}}. \tag{1.3}$$

La maximisation s'effectue sous un ensemble de $N_C + 1$ contraintes :

$$\mathbb{E}[g_\alpha] = \int g_\alpha(y_{\text{map}}) f_G(y_{\text{map}}) \mathrm{d}y_{\text{map}},\ \alpha = 0, 1, \ldots, n_C.$$

Les fonctions g_α sont choisies de telle sorte que les espérances $\mathbb{E}[g_\alpha]$ correspondent aux différents moments de la variable constituant la connaissance générale G sur le phénomène étudié. Si la connaissance générale G est constituée par les deux premiers moments, on prend :

- $g_0(y_{\text{map}}) = 1$, d'où $\mathbb{E}[g_0] = 1$ ce qui définit la constante de normalisation,
- $g_\alpha(y_i) = Y_i, i = 1, \ldots, (m+n)+1$, d'où $\mathbb{E}[g_\alpha] = \mathbb{E}[Y_i]$,
- $g_\alpha(y_i, y_j) = [y_i - \mathbb{E}(Y_i)][y_j - \mathbb{E}(Y_j)]$
  avec $\alpha = (m+n)+2, \ldots, (m+n+1)(m+n+4)/2$, d'où $\mathbb{E}[g_\alpha] = \mathbb{E}\{[Y_i - \mathbb{E}(Y_i)][Y_j - \mathbb{E}(Y_j)]\}$.

La seconde étape consiste à recueillir l'information spécifique aux sites S. On distingue deux types de données : les données considérées comme des mesures exactes appelées données « hard » et les autres appelées données « soft ». Par exemple, si l'on a effectué des mesures répétées en certains sites d'échantillonnage, les distributions $f_S(y_{\text{soft}})$ de ces mesures en chaque point peuvent constituer des données dites « soft » (Savelieva et al., 2005). Les données $y_{\text{données}}$ se décomposent en $(y_{\text{hard}}, y_{\text{soft}})$, où $y_{\text{hard}} = (y_1, \ldots, y_n)$ et $y_{\text{soft}} = (y_{n+1}, \ldots, y_{n+m})$.

La troisième étape consiste à mettre à jour la distribution *a priori* en considérant les informations spécifiques aux sites S recueillies à la seconde étape. Si les données « soft »

sont constituées par des densités de probabilité $f_S(y_{\text{soft}})$, la densité *a posteriori* au point de prédiction $\mathbf{s}_0$ s'écrit alors :

$$f_{G\cup S}(y_0|y_{\text{hard}}, f_S(y_{\text{soft}})) = \frac{\int f_G(y_{\text{map}}) f_S(y_{\text{soft}}) \mathrm{d}y_{\text{soft}}}{\int f_G(y_{\text{données}}) f_S(y_{\text{soft}}) \mathrm{d}y_{\text{soft}}}.$$

On s'intéresse à cette densité ou à la fonction de Bayes définie par :

$$B_{\mathbf{y}}(\mathbf{s}_0) = \ln\lfloor f_{G\cup S}(y_0|y_{\text{hard}}, f_S(y_{\text{soft}}))\rfloor$$

où $\lfloor x \rfloor$ désigne la partie entière de x. La densité *a posteriori* ou la fonction de Bayes doit être maximisée par rapport à y_0. La valeur de y_0 ainsi déterminée est la valeur de l'estimateur Y_0^* au point $\mathbf{s}_0$. De plus amples détails sur la procédure d'estimation sont donnés dans les articles de Christakos (1990) et de Bogaert (2002).

Développée tout d'abord pour traiter des variables continues comme ci-dessus, l'approche BME a ensuite été étendue pour traiter des variables discrètes (Bogaert, 2002; Bogaert et D'Or, 2002; D'Or et Bogaert, 2004). Plus récemment, l'approche BME a été utilisée pour la prédiction simultanée de variables catégorielles (ordinales ou nominales) et continues (Wibrin et al., 2006). L'approche BME a l'avantage de ne faire appel à aucune hypothèse sur la linéarité de l'estimateur, la normalité des variables ou l'homogénéité de la distribution spatiale (Christakos, 1998). En revanche, cette méthode est beaucoup plus gourmande en ressources informatiques que les méthodes de krigeage classiques. Cela est due en partie à la résolution de systèmes faisant intervenir des équations non linéaires et au nombre important de contraintes à vérifier. Le nombre de contraintes peut, en effet, devenir rapidement très élevé, en particulier lorsque l'on traite des variables ordinales ou nominales (Wibrin et al., 2006). L'approche BME nécessite d'avoir une connaissance générale G du phénomène étudié. Cette connaissance est constituée le plus souvent par la donnée des différents moments. Les moments théoriques de la ou des variables doivent donc être connus ou doivent pouvoir être inférés avec une précision suffisante. Notons que si la connaissance générale G se limite à la connaissance théorique des deux premiers moments et si les données observées ne sont que des données « exactes » (données « hard »), ce qui est notre cas, l'approche BME revient à réaliser un krigeage ordinaire des données. Étant donné les calculs qu'elle nécessite, l'approche BME ne s'avère donc être intéressante que si l'on dispose de données « soft » permettant d'améliorer la qualité des prédictions par un krigeage ordinaire.

Les méthodes de prédiction nécessitent d'avoir préalablement modélisé la structure de dépendance existant entre les variables, c'est ce que nous allons voir dans la partie suivante.

1.2.2 Modélisation de la structure dépendance entre les variables

Les méthodes de prédiction multivariées prennent en compte la dépendance existant entre les variables. Cette structure de dépendance peut être modélisée soit par une matrice de covariance, soit par les distributions bivariées de tous les couples de variables. Le choix

du modèle employé est dicté par la méthode de prédiction utilisée.

Soient $Z_1(.), Z_2(.), \ldots, Z_K(.)$ les K champs spatiaux étudiés et $\mathbf{s}_1, \ldots, \mathbf{s}_n$ les n points d'échantillonnage situés dans le domaine d'étude $\mathcal{D}$. On note $\mathbf{Z}(\mathbf{s}) = (Z_1(\mathbf{s}), Z_2(\mathbf{s}), \ldots, Z_K(\mathbf{s}))'$ le vecteur de toutes les variables aléatoires au point $\mathbf{s}, \mathbf{s} \in \mathcal{D}$ et $\mathbf{Z} \equiv (\mathbf{Z}(\mathbf{s}_1)', \mathbf{Z}(\mathbf{s}_2)', \ldots, \mathbf{Z}(\mathbf{s}_n)')'$ le vecteur de toutes les variables aléatoires considérées en tous les points d'échantillonnage.

La dépendance est modélisée par les lois bivariées de tous les couples $(Z_k(\mathbf{s}_i), Z_m(\mathbf{s}_j))$**.**

Dans le cas du krigeage ou du cokrigeage disjonctif, la structure de dépendance entre les variables se traduit au travers des lois bivariées des couples $(Z_k(\mathbf{s}_i), Z_m(\mathbf{s}_j))$. Dans le cas univarié, par exemple, il faut spécifier le modèle isofactoriel décrivant la distribution du couple $(Z_k(\mathbf{s}), Z_k(\mathbf{s}+\mathbf{h}))$ pour pouvoir résoudre le système de krigeage.

Il existe différents types de modèles isofactoriels. Ces modèles diffèrent par le choix de la base hilbertienne $\{\chi_q, q = 0, 1, \ldots\}$ utilisée pour la décomposition des distributions bivariées et par le choix de la forme de la covariance des facteurs $\chi_q(Z_k(\mathbf{s}))$, $T_q(\mathbf{h}) = \mathbb{E}[\chi_q(Z_k(\mathbf{s}))\chi_q(Z_k(\mathbf{s}+\mathbf{h}))]$, pour tout $q > 0$ ($T_0(\mathbf{h}) \equiv 1$). Le choix de la base hilbertienne dépend de la nature de la distribution marginale de Z_k (Chilès et Delfiner, 1999). On utilise principalement le modèle hermitien pour les distributions symétriques, le modèle laguerrien pour des distributions continues mais non symétriques et le modèle de Meixner pour des distributions de type discret. Pour la forme des covariances T_q, on se limite, en pratique, à des modèles où les covariances $T_q(\mathbf{h})$ sont fonctions de $\rho(\mathbf{h})$ et où $\rho(\mathbf{h}) \geq 0$ est le corrélogramme de $Z_k(\mathbf{s})$. Les quatre formes les plus utilisées (Chilès et Delfiner, 1999) sont :

- $T_q(\mathbf{h}) = \rho^q(\mathbf{h}), q \geq 0$, le modèle isofactoriel est de type diffusif pur ;
- $T_q(\mathbf{h}) = \rho(\mathbf{h}), q > 0$, le modèle isofactoriel est de type mosaïque ;
- $T_q(\mathbf{h}) = \beta\rho^q(\mathbf{h}) + (1-\beta)\rho(\mathbf{h}), q > 0$, où $0 < \beta < 1$, le modèle isofactoriel est de type barycentrique ;
- $T_q(\mathbf{h}) = \dfrac{\Gamma(\beta)}{\Gamma(\beta+q)} \dfrac{\Gamma(\beta\rho(\mathbf{h})+q)}{\Gamma(\beta\rho(\mathbf{h}))}, q \geq 0$ où $\beta > 0$, le modèle isofactoriel est de type beta.

La dépendance spatiale est modélisée par une matrice de covariance.

Une seconde façon de modéliser la dépendance entre les variables est de construire la matrice de covariance du vecteur $\mathbf{Z}$. C'est sous cette forme que l'on modélise la dépendance dans le cas du krigeage et du cokrigeage. Cette matrice se décompose en différents blocs de la forme $\mathrm{Cov}[\mathbf{Z}(\mathbf{s}_i), \mathbf{Z}(\mathbf{s}_j)]$. Ces blocs sont des matrices de dimension $K \times K$ qui ne sont pas forcément symétriques. En revanche, la matrice de covariance de $\mathbf{Z}$, de dimension $Kn \times Kn$, doit être définie positive quels que soient le nombre et le choix des points échantillonnés. La difficulté est de proposer des modèles qui définissent des matrices de covariance « valides ». Nous décrivons ici plusieurs modèles admissibles.

Modèle à covariance proportionnelle ou modèle de corrélation intrinsèque Le modèle de corrélation intrinsèque (Wackernagel, 2003) consiste à écrire $\mathbf{Z}(\mathbf{s})$ sous la forme $\mathbf{Z}(\mathbf{s}) = \mathbf{A}\mathbf{Y}(\mathbf{s})$ où les composantes $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}), \ldots, Y_K(\mathbf{s}))'$ sont des processus spatiaux indépendants et identiquement distribués. Si les $Y_k(\mathbf{s})$ sont de moyenne nulle, stationnaires, de variance 1 et de fonction de corrélation $\rho(\mathbf{h})$, alors

- $\mathbb{E}[\mathbf{Z}] = \mathbf{0}$,
- et la matrice de covariance croisée est définie par :

$$\mathrm{Cov}[\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s}+\mathbf{h})] \equiv \mathbf{C}(\mathbf{h}) = \rho(\mathbf{h})\mathbf{A}\mathbf{A}'.$$

Ainsi, en prenant $\mathbf{A}\mathbf{A}' = \mathbf{V}$ ($\mathbf{V}$ est la matrice de covariance de $\mathbf{Z}(\mathbf{s})$), on a : $\mathbf{C}(\mathbf{h}) = \mathbf{V}\rho(\mathbf{h})$. Les covariances simples et croisées sont toutes proportionnelles à une même fonction de corrélation de base.

Soit $\mathbf{R}$ la matrice de taille $n \times n$ définie par $[\mathbf{R}]_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j)$. La matrice de covariance de $\mathbf{Z}$ s'écrit alors sous la forme : $\mathbf{\Sigma_Z} = \mathbf{R} \otimes \mathbf{V}$ où $\otimes$ désigne le produit de Kronecker. Le modèle à covariance proportionnelle permet donc de travailler avec des matrices de taille $n \times n$ et $K \times K$ au lieu de manipuler des matrices de taille $Kn \times Kn$, c'est là l'un de ces principaux avantages. En revanche, ce modèle est assez rigide. Toutes les composantes de $\mathbf{Z}(\mathbf{s})$ ont la même portée étant donné qu'une seule fonction de corrélation a été introduite. Pour remédier à ce problème, ce modèle a été étendu en un modèle appelé modèle linéaire de corégionalisation.

Modèle linéaire de corégionalisation Le modèle linéaire de corégionalisation (Grzebyk et Wackernagel, 1994; Wackernagel, 2003) est le modèle le plus fréquemment utilisé. Il consiste à définir un processus avec un modèle de covariance emboîtée. On écrit $\mathbf{Z}(\mathbf{s})$ sous la forme :

$$\mathbf{Z}(\mathbf{s}) = \sum_{u=1}^{U} \mathbf{Z}_u(\mathbf{s}) = \sum_{u=1}^{U} \mathbf{A}_u \mathbf{Y}_u(\mathbf{s})$$

où les processus $\mathbf{Z}_u$ sont des spécifications indépendantes du modèle intrinsèque et où chaque processus $\mathbf{Y}_u$ a pour fonction de covariance ρ_u. La matrice de covariance croisée entre $\mathbf{Z}(\mathbf{s})$ et $\mathbf{Z}(\mathbf{s}+\mathbf{h})$ est de la forme : $\mathbf{C}(\mathbf{h}) = \sum_{u=1}^{U} \mathbf{B}_u \rho_u(\mathbf{h})$ avec $\mathbf{B}_u = \mathbf{A}_u {\mathbf{A}_u}'$.

Autrement dit, l'ensemble des fonctions aléatoires $\{Z_k(\mathbf{s}); k = 1, 2, ..., K\}$ peut être décomposé en K sous-ensembles $\{Z_k^u(\mathbf{s}); u = 0, ..., U\}$ de composantes spatialement non corrélées tel que :

$$Z_k(\mathbf{s}) = \sum_{u=0}^{U} Z_k^u(\mathbf{s}) + m_k, \text{ où pour toutes valeurs de } k, m, u \text{ et } v$$

$$\begin{aligned} \mathbb{E}[Z_k(\mathbf{s})] = m_i \quad &\text{et} \quad \mathbb{E}[Z_k^u(\mathbf{s})] = 0, \\ \mathrm{Cov}[Z_k^u(\mathbf{s}), Z_m^u(\mathbf{s}+\mathbf{h})] &= \mathbb{E}[Z_k^u(\mathbf{s}) Z_m^u(\mathbf{s}+\mathbf{h})] = C_{km}^u(\mathbf{h}), \\ \mathrm{Cov}[Z_k^u(\mathbf{s}), Z_m^v(\mathbf{s}+\mathbf{h})] &= 0 \text{ quand } u \neq v. \end{aligned}$$

Aux composantes spatiales sont associées des covariances croisées $C^u_{km}(\mathbf{h})$ de palier b^u_{km} et de fonction de corrélation spatiale $\rho_u(\mathbf{h})$:

$$C_{km}(\mathbf{h}) = \sum_{u=0}^{U} C^u_{km}(\mathbf{h}) = \sum_{u=0}^{U} b^u_{km}\rho_u(\mathbf{h}).$$

En regroupant les paliers b^u_{km} en $U+1$ matrices $\mathbf{B}_u$ d'ordre $K \times K$, on obtient le modèle de covariance suivant :

$$\mathbf{C}(h) = \sum_{u=0}^{U} \mathbf{B}_u\rho_u(\mathbf{h}).$$

Les matrices $\mathbf{B}_u$ sont nécessairement semi-définies positives. Si l'une d'entre elles ne l'est pas, alors le modèle n'est pas un modèle linéaire de corégionalisation et l'on ne sait pas si le modèle est admissible. La structure de dépendance peut être décrite de manière équivalente en terme de variogramme sous la forme :

$$\Gamma(\mathbf{h}) = \sum_{u=0}^{U} \mathbf{B}_u\gamma_u(\mathbf{h}),$$

où γ_u désigne un variogramme normé.

Le principal avantage du modèle linéaire de corégionalisation est qu'il permet de décrire la dépendance entre les variables à différentes échelles spatiales. En revanche, la détermination des variogrammes élémentaires γ_u ou de manière équivalente des corrélations ρ_u, reste délicate. Ces variogrammes élémentaires sont à la base du calcul de tous les variogrammes et de tous les variogrammes croisés ; leur rôle est donc important. Le choix de ces variogrammes élémentaires parmi toutes les fonctions variogrammes est libre (Goulard et Voltz, 1992). Il n'existe aucune règle pour les déterminer. Le modèle doit reproduire au mieux les éléments caractérisant le variogramme, notamment le comportement à l'origine. Si plusieurs modèles présentent un bon ajustement aux données, on choisira le plus parcimonieux.

Matrice de covariance construite à partir de la méthode moyenne mobile Le problème de la modélisation de la dépendance entre les variables est de déterminer une matrice de covariance, ou de manière équivalente un variogramme, qui soit valide. Les modèles de covariance décrits ci-dessus sont valides, mais peu flexibles car ils sont basés sur un nombre restreint de fonctions élémentaires (corrélations ou variogrammes normés). Ver Hoef et Barry définissent une nouvelle famille de variogrammes valides basés sur des fonctions de carré intégrable dites « fonctions moyennes mobiles » (Barry et Ver Hoef, 1996; Ver Hoef et Barry, 1998). Ces variogrammes, outre le fait qu'ils soient valides, présentent l'avantage d'être très flexibles. Cette méthode, permet, en effet, d'approcher d'aussi près que l'on veut n'importe quel variogramme, à condition que les fonction moyennes mobiles soient bien choisies. Le modèle de covariance basé sur la construction moyenne mobile offre la possibilité de traiter aussi bien des données isotropes qu'anisotropes (Ver Hoef

et al., 2004). Ce modèle de covariance décrit la structure de dépendance existant entre des processus aléatoires spatiaux Z_k, $k = 1, 2, \ldots, K$ construits par intégration du produit de convolution d'une fonction moyenne mobile f_k et d'un mélange de bruits blancs, d'où le nom de la méthode. Une description plus détaillée de la construction de ces processus spatiaux Z_k est donnée ci-dessous. Il est possible de déterminer la matrice de covariance de tels processus grâce aux propriétés des bruits blancs.

Définition 3 Bruit blanc gaussien (Kuo, 2001)
Un bruit blanc gaussien continu bidimensionnel *est un processus* $(W(x))_{x\in\mathbb{R}^2}$ *décrit par l'intégrale stochastique suivante :*

$$\forall f \in L^2(\mathbb{R}^2),\ \int f(x)W(x)dx =< f, W >\sim \mathcal{N}\left(0, \int f^2\right).$$

Soit $W_k(\mathbf{x})$, $k = 0, 1, \ldots, K$ un bruit blanc centré défini sur $\mathbb{R}^d$, autrement dit :
- $\mathbb{E}[W_k(\mathbf{x})] = 0$,
- $\text{Var}\left[\int_A W_k(\mathbf{x})\mathrm{d}\mathbf{x}\right] = |A|$ où $|A|$ désigne la mesure de Lebesgue de l'ensemble A,
- $\text{Cov}\left[\int_A W_k(\mathbf{x})\mathrm{d}\mathbf{x}, \int_B W_k(\mathbf{x})\mathrm{d}\mathbf{x}\right] = 0$ quand $A \cap B = \emptyset$,
- $W_k(\mathbf{x})$ est indépendant de $W_m(\mathbf{x})$ quand $k \neq m$.

On considère la combinaison linéaire de bruits blancs définie comme suit :

$$V_k(\mathbf{x}|\rho_k, \boldsymbol{\Delta}_k) = \sqrt{1-\rho_k^2}W_k(\mathbf{x}) + \rho_k W_0(\mathbf{x} - \boldsymbol{\Delta}_k) \tag{1.4}$$

où $-1 \leq \rho_k \leq 1$ pour $k = 1, 2, \ldots, K$. Le vecteur $\boldsymbol{\Delta}_k = (\boldsymbol{\Delta}_{k1}, \ldots, \boldsymbol{\Delta}_{kd})'$ permet d'introduire un « décalage » spatial dans la corrélation entre deux processus $V(.)$, c'est-à-dire que, si $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_m = \mathbf{0}$, les variables $V_k(\mathbf{x})$ et $V_m(\mathbf{t})$ sont indépendantes pour $\mathbf{x} \neq \mathbf{t}$. Les combinaisons linéaires V_k, $k = 1, 2, \ldots, K$ partagent un processus commun W_0 ; elles sont donc spatialement dépendantes. Chaque processus V_k vérifie :
- $\mathbb{E}[V_k(\mathbf{x})] = 0$,
- $\text{Var}\left[\int_A V_k(\mathbf{x})\mathrm{d}\mathbf{x}\right] = |A|$,
- $\text{Cor}\left[\int_A V_k(\mathbf{x}+a)\mathrm{d}\mathbf{x}, \int_A V_m(\mathbf{x}+b)\mathrm{d}\mathbf{x}\right] = \rho_k\rho_m \equiv \rho_{km},\ a, b \in \mathbb{R}^d$.

Soit U_k, $k = 1, 2, \ldots, K$ un bruit blanc tel que :
- $\mathbb{E}[U_k(\mathbf{x})] = 0$,
- $\text{Var}[U_k(\mathbf{x})] = 1$,
- $U_k(\mathbf{x})$ soit indépendant de $U_k(\mathbf{t})$ pour tout $\mathbf{x} \neq \mathbf{t}$,
- $U_k(\mathbf{x})$ soit indépendant de $U_m(\mathbf{t})$ pour tout $k \neq m$ et pour tous $\mathbf{x}$ et $\mathbf{t}$.

Les processus spatiaux $Z_k(.)$, $k = 1, 2, \ldots, K$ auxquels nous nous intéressons sont définis, pour tout $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d$, par :

$$Z_k(\mathbf{s}|\nu_k, \mu_k, \rho_k, \boldsymbol{\Delta}_k) \equiv \mu_k + \int_{\mathbb{R}^d} f_k(\mathbf{x} - \mathbf{s})V_k(\mathbf{x}|\rho_k, \boldsymbol{\Delta}_k)\mathrm{d}\mathbf{x} + \nu_k U_k(\mathbf{s}) \tag{1.5}$$

où f_k désigne une fonction définie sur $\mathbb{R}^d$ de carré intégrable et où μ_k et ν_k sont des réels (Ver Hoef et Barry, 1998). Les processus ainsi définis sont des processus stationnaires d'ordre 2. Le moment d'ordre un du processus $Z_k(.)$ est donné par :

$$\mathbb{E}[Z_k(\mathbf{s})] = \mu_k.$$

Lemme 1 (Yaglom, 1987)
Soit $W(.)$ un bruit blanc défini sur $\mathbb{R}^d$.
Soient f_k et f_m deux fonctions de carré intégrables définies sur $\mathbb{R}^d$.
On a la relation suivante :

$$\mathrm{Cov}\left[\int_{\mathbb{R}^d} f_k(\mathbf{x})W(\mathbf{x})\,d\mathbf{x}, \int_{\mathbb{R}^d} f_m(\mathbf{x}-\mathbf{h})W(\mathbf{x})\,d\mathbf{x}\right] = \int_{\mathbb{R}^d} f_k(\mathbf{x})f_m(\mathbf{x}-\mathbf{h})\,d\mathbf{x}.$$

En utilisant le lemme 1, il est possible de calculer la covariance et la covariance croisée des processus spatiaux $Z_k(.)$, $k = 1, 2, \ldots, K$:

$$\mathrm{Var}[Z_k(\mathbf{s})] = \int_{\mathbb{R}^d} f_k^2(\mathbf{x})\mathrm{d}\mathbf{x} + \nu_k^2 \tag{1.6}$$

$$\mathrm{Cov}[Z_k(\mathbf{s}), Z_k(\mathbf{s}+\mathbf{h})] = \int_{\mathbb{R}^d} f_k(\mathbf{x})f_k(\mathbf{x}-\mathbf{h})\mathrm{d}\mathbf{x} \text{ pour } \mathbf{h} \neq \mathbf{0} \tag{1.7}$$

$$\mathrm{Cov}[Z_k(\mathbf{s}), Z_m(\mathbf{s}+\mathbf{h})] = \rho_k\rho_m \int_{\mathbb{R}^d} f_k(\mathbf{x})f_m(\mathbf{x}-\mathbf{h}+\boldsymbol{\Delta}_m-\boldsymbol{\Delta}_k)\mathrm{d}\mathbf{x}. \tag{1.8}$$

La matrice de covariance de $\mathbf{Z}$ ainsi déterminée est définie positive. Le paramètre ν_k^2 représente l'effet de pépite associé au processus $Z_k(.)$. Les paramètres $\boldsymbol{\Delta}_m$ et $\boldsymbol{\Delta}_k$ permettent d'introduire une dissymétrie dans la modélisation de la covariance croisée.

Selon les fonctions moyennes mobiles employées, les intégrales intervenant dans les équations 1.6, 1.7 et 1.8 peuvent être plus ou moins complexes à calculer, voire ne pas avoir de solution analytique. Une solution consiste à considérer des fonctions moyennes mobiles constantes par morceaux. Les intégrales sont alors faciles à calculer mais le nombre de paramètres à estimer devient important. Une autre solution consiste à approcher les intégrales intervenant dans la matrice de covariance par des méthodes numériques, par exemple, en utilisant la transformée de Fourier rapide (Ver Hoef et al., 2004). Nous reviendrons plus en détails sur ce point au paragraphe 2.1.2.

Les processus spatiaux construits à partir de la méthode moyenne mobile sont entièrement déterminés par le choix du processus latent W et des fonctions moyennes mobiles. Ici, les processus spatiaux considérés $Z_k(.)$ sont gaussiens car ils sont construits en intégrant le produit de convolution de la fonction moyenne mobile et du mélange de bruits blancs gaussiens. L'hypothèse de normalité requise pour le cokrigeage est donc respectée. Notons cependant, que la construction basée sur les fonctions moyennes mobiles peut être étendue pour modéliser des processus spatiaux non gaussiens (Higdon, 2001). Le bruit blanc peut être remplacé par un autre processus stochastique à incréments indépendants.

Ainsi Wolpert et Ickstadt (1998) ont-ils modélisé l'intensité d'une variable de comptage en considérant l'intégration d'une fonction par rapport à un processus gamma. Tu (2006) a, quant à lui, généralisé la construction en utilisant la convolution de processus de Lévy pour modéliser des processus spatiaux temporels.

Le choix des fonctions moyennes mobiles est plus délicat. La méthode moyenne mobile a d'abord été développée avec des fonctions moyennes mobiles constantes par morceaux (Barry et Ver Hoef, 1996; Ver Hoef et Barry, 1998). L'avantage de ces fonctions est qu'elles permettent d'obtenir une expression explicite de la fonction de covariance et qu'elles permettent, si le découpage est suffisamment fin, d'approcher n'importe quel variogramme d'aussi près que l'on veut. En revanche, le nombre de paramètres à estimer est très important (autant de paramètres que de morceaux dans les fonctions). Il existe d'autres fonctions moyennes mobiles ayant un nombre plus restreint de paramètres. On distingue deux grandes familles de fonctions moyennes mobiles : les fonctions moyennes mobiles isotropes et les fonctions moyennes mobiles anisotropes. Parmi les fonctions moyennes mobiles isotropes, on trouve notamment (Kern, 2000; Higdon, 2001) :

- le noyau gaussien $f(\mathbf{d}) \propto e^{-\tau||\mathbf{d}||^2}$, $\tau > 0, d \in \mathbb{R}^2$ qui conduit à la covariance gaussienne,
- le noyau exponentiel $f(\mathbf{d}) \propto e^{-\tau||\mathbf{d}||}$, $\tau > 0$,
- le noyau cylindrique $f(\mathbf{d}) \propto \mathbb{1}_{\{||\mathbf{d}||<r\}}$, $r > 0$ qui conduit à la covariance cylindrique,
- le noyau sphérique $f(\mathbf{d}) \propto \left(1 - \dfrac{3||\mathbf{d}||}{2r} + \dfrac{||\mathbf{d}||^3}{2r^3}\right) \mathbb{1}_{\{||\mathbf{d}||<r\}}$, $r > 0$
- le noyau « demi-sphérique » $f(\mathbf{d}) \propto \sqrt{r^2 - ||\mathbf{d}||^2}\mathbb{1}_{\{||\mathbf{d}||<r\}}$, $r > 0$,
- le noyau d'Epanechnikov $f(\mathbf{d}) \propto \left(1 - \dfrac{||\mathbf{d}||^2}{r^2}\right) \mathbb{1}_{\{||\mathbf{d}||<r\}}$, $r > 0$,
- le noyau « tri-cubique » $f(\mathbf{d}) \propto \left(1 - \dfrac{||\mathbf{d}||^3}{r^3}\right)^3 \mathbb{1}_{\{||\mathbf{d}||<r\}}$, $r > 0$,
- le noyau conique $f(\mathbf{d}) \propto \left(1 - \dfrac{||\mathbf{d}||}{r}\right) \mathbb{1}_{\{||\mathbf{d}||<r\}}$, $r > 0$,
- le noyau proposé par Kern (2000) constitué d'un empilement de m cylindres concentriques de hauteur ϕ_i et de rayon r_i de plus en plus petit

$$f_{r,\phi}(\mathbf{d}) = \begin{cases} \sum_{j=1}^{m-i+1} \phi_j & \text{si } r_{i-1} \leq ||\mathbf{d}|| < r_i \\ 0 & \text{sinon} \end{cases}$$

  où $r_0 \equiv 0$.

Des figures illustrant les différents corrélogrammes obtenus à partir de ces fonctions moyennes mobiles sont présentées au chapitre 2 de la thèse de Kern (2000). Les variables construites à partir de ces noyaux isotropes sont elles-mêmes isotropes. Pour traiter des données anisotropiques, il faut avoir recours à des noyaux anisotropes, par exemple, Ver Hoef et al. (2004) proposent la fonction moyenne mobile suivante :

$$f(x,y) = (1 - x_t^2 + y_t^2)^{\theta_4} \mathbb{1}_{\{\sqrt{x_t^2+y_t^2}<1\}}, \ (x,y) \in \mathbb{R}^2 \tag{1.9}$$

où

$$x_t = \frac{x\cos\theta_1 - y\sin\theta_1}{\theta_2}, \ y_t = \frac{x\sin\theta_1 + y\cos\theta_1}{\theta_3},$$

$\theta_1 \in \left[-\frac{\pi}{2}; \frac{\pi}{2}\right]$ et $\theta_2, \theta_3, \theta_4 > 0$. Cette fonction a pour support une ellipse de centre $(0, 0)$ de demi-axes θ_2 et θ_3 ayant subi une rotation d'angle θ_1. Les paramètres θ_1, θ_2 et θ_3 permettent de prendre en compte l'anisotropie. Si $\theta_2 = \theta_3$, le champ modélisé est isotrope. Le paramètre θ_4 a une influence sur la portée du variogramme ; plus le paramètre θ_4 est grand, plus la portée est petite.

Quelles que soient les fonctions moyennes mobiles choisies, il faut s'assurer de l'identifiabilité des paramètres. Les paramètres de la fonction moyenne mobile 1.9 telle qu'elle est formulée ici ne sont pas identifiables ; ce point n'est pas soulevé dans l'article de Ver Hoef et al. (2004). En effet, les deux quadruplets $(\theta_1, \theta_2, \theta_3, \theta_4)$ et $(-\pi/2 + \theta_1, \theta_3, \theta_2, \theta_4)$ définissent la même ellipse et conduisent à la même matrice de covariance. Il est donc nécessaire d'ajouter une contrainte supplémentaire, $\theta_2 \leq \theta_3$, pour s'assurer de l'identifiabilité des paramètres.

Le problème du choix des fonctions moyennes mobiles est équivalent au problème du choix des corrélations dans le modèle linéaire de corégionalisation (paragraphe 1.2.2). Il n'existe pas de règle pour déterminer ces fonctions, c'est là le principal inconvénient de la méthode dite moyenne mobile. Il est parfois difficile d'établir un lien entre fonctions moyennes mobiles et corrélogramme. En revanche, le large choix de fonctions moyennes mobiles donne la possibilité de modéliser de nombreuses structures de covariance et rend la méthode très souple.

La dépendance entre les variables peut donc être modélisée par différentes méthodes. Lorsque les prédictions sont effectuées par krigeage ou cokrigeage disjonctif, la dépendance entre les variables est décrite par les distributions bivariées des couples de variables. Dans les cas plus classiques où le krigeage ou le cokrigeage ordinaire sont utilisés pour prédire les variables, la structure de dépendance entre les variables est modélisée par une matrice de covariance. Contrairement au modèle de covariance proportionnelle, le modèle linéaire de corégionalisation permet de réaliser une analyse de la structure de corrélation entre les variables qui tienne compte des différentes échelles spatiales. Le modèle de covariance basé sur la construction de processus spatiaux dépendants par la méthode moyenne mobile a l'avantage d'être très flexible. En effet, le large choix de fonctions moyennes mobiles offre la possibilité de modéliser un très grand nombre de variogrammes.

1.3 Modèle hiérarchique spatial multivarié

1.3.1 Principe du modèle

Les méthodes de prédiction permettant de traiter simultanément des variables continues et discrètes sont peu nombreuses. L'objectif est ici de proposer un modèle spatial multivarié permettant de prédire simultanément des variables de nature différente.

Le modèle que nous développons s'appuie sur les modèles linéaires généralisés spatiaux

(Diggle et al., 1998). Comme nous l'avons vu au paragraphe 1.2.1, les modèles linéaires généralisés spatiaux permettent de traiter différents types de variables, notamment des variables gaussiennes, des variables de Poisson et des variables ordinales. Ces modèles sont le plus souvent utilisés dans un cadre univarié (Christensen et Waagepetersen, 2002; De Oliveira, 2000). Nous en proposons une généralisation au cas multivarié. La généralisation est directe pour les variables gaussiennes et de Poisson, mais elle est plus délicate pour les variables ordinales. Dans un cadre non spatial, un modèle spécifique, appelé modèle probit ordinal multivarié (Définition 4), a été développé pour traiter les variables ordinales. Le principe de ce modèle consiste à tronquer des variables sous-jacentes continues pour modéliser des variables binaires ou ordinales (Ashford et Swoden, 1970; Ochi et Prentice, 1984; Albert et Chib, 1993; Chib et Greenberg, 1998).

Définition 4 Modèle probit ordinal multivarié (Chen et Shao, 1999)
Soit $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iK})'$ *un vecteur réponse de dimension* K. *On considère un échantillon* $\mathbf{Y} = (\mathbf{Y}_1{}', \mathbf{Y}_2{}', \ldots, \mathbf{Y}_n{}')'$ *constitué de* n *vecteurs indépendants. On note* $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iK})'$ *et* $\mathbf{y} = (\mathbf{y}_1{}', \mathbf{y}_2{}', \ldots, \mathbf{y}_n{}')'$ *les vecteurs observés correspondants. Pour tout* k, $1 \leq k \leq K$, *et pour tout* i, $1 \leq i \leq n$, *la variable aléatoire* Y_{ik} *prend une valeur entière comprise entre 1 et* L_k, $L_k \geq 2$. *Soient* $\mathbf{x}_{ik} = (x_{ik1}, x_{ik2}, \ldots, x_{ikj_k})'$ *et* $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kj_k})'$ *le vecteur de régression et le vecteur des coefficients de régression associés à la variable* Y_k. *On note* $\boldsymbol{\beta} = (\boldsymbol{\beta}_1{}', \boldsymbol{\beta}_2{}', \ldots, \boldsymbol{\beta}_K{}')'$. *Soit* $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)'$ *l'intercept.*

Il existe un vecteur latent de variables aléatoires de dimension K $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iK})'$ *tel que*

$$y_{ik} = l \quad si \quad \alpha_{k,l-1} \leq z_{ik} < \alpha_{k,l}$$

pour $l = 1, 2, \ldots, L_k$, *où*

$$-\infty = \alpha_{k,0} < \alpha_{k,1} \leq \ldots \leq \alpha_{k,L_k-1} < \alpha_{k,L_k} = +\infty$$

sont les seuils qui divisent la droite réelle en L_k *intervalles et où* $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})'$ *désigne une réalisation du vecteur* Z_i.

On considère que $\mathbf{Z}_i|\boldsymbol{\beta}, \mathbf{R} \sim \mathcal{N}_K(\boldsymbol{\mu} + \mathbf{X}_i\boldsymbol{\beta}, \mathbf{R})$ *où* $\mathbf{R}$ *est une matrice de corrélation de taille* $K \times K$ *et* $\mathbf{X}_i$, $1 \leq i \leq n$, *est une matrice bloc diagonale de dimension* $K \times \sum_{k=1}^{K} j_k$ *définie par* $\mathbf{X}_i = diag(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iK})$. *Alors la distribution du vecteur réponse* $\mathbf{Y}_i$ *est donnée par :*

$$\begin{aligned} \mathbb{P}(\mathbf{Y}_i = \mathbf{y}_i|\boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{R}, \boldsymbol{\alpha}) &= \mathbb{P}(Z_{i1} \in A_{i1}, Z_{i2} \in A_{i2}, \ldots, Z_{iK} \in A_{iK}|\boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{R}) \\ &= \int_{A_{i1}} \int_{A_{i2}} \ldots \int_{A_{iK}} \phi_K(\mathbf{z}_i|\boldsymbol{\mu} + \mathbf{X}_i\boldsymbol{\beta}, \mathbf{R}) d\mathbf{z}_i \end{aligned}$$

où $A_{ik} = [\alpha_{k,l-1}, \alpha_{k,l}[$ *si* $y_{ik} = l$ *(*$1 \leq l \leq L_k$*) pour* $k = 1, 2, \ldots, K$. *La notation* $\phi_K(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *désigne la fonction de densité d'une loi normale multivariée de dimension* K *de moyenne* $\boldsymbol{\mu}$ *et de matrice de covariance* $\boldsymbol{\Sigma}$.

La matrice $\mathbf{R}$ n'est pas une matrice de covariance mais une matrice de corrélation. En effet, si l'on considère la paramétrisation $(\boldsymbol{\gamma}, \boldsymbol{\Omega})$, où $\boldsymbol{\gamma}$ désigne le vecteur des coefficients

de régression et $\mathbf{\Omega}$ la matrice de covariance, on peut montrer que $\mathbb{P}(\mathbf{Y}_i = \mathbf{y}_i|\boldsymbol{\gamma}, \mathbf{\Omega}) = \mathbb{P}(\mathbf{Y}_i = \mathbf{y}_i|\boldsymbol{\beta}, \mathbf{R})$, où $\boldsymbol{\beta}_k = \omega_{kk}^{-1/2}\boldsymbol{\gamma}_k$, $\mathbf{R} = \mathbf{D}\mathbf{\Omega}\mathbf{D}'$ et $\mathbf{D} = \text{diag}(\omega_{11}^{-1/2}, \ldots, \omega_{KK}^{-1/2})$ (Chib et Greenberg, 1998; De Oliveira, 2000). Travailler avec des matrices de covariance ne permet pas d'identifier les paramètres du modèle, c'est pourquoi on impose que $\mathbf{R}$ soit une matrice de corrélation. Une autre contrainte doit être imposée pour rendre les paramètres du modèle identifiables. Soit l'intercept $\boldsymbol{\mu}$ est fixé à 0 et tous les seuils sont identifiables, soit l'intercept varie et les seuils $\alpha_{k,1}$, $k = 1, 2, \ldots, K$ sont fixés à 0 (Albert et Chib, 1993; Cowles, 1996). Dans la suite, nous nous plaçons dans cette configuration. Nous estimons l'intercept et les $L_k - 2$ seuils restants pour chaque variable ordinale.

Dans un cadre multivarié, la dépendance entre les variables doit être prise en compte pour améliorer la qualité des prédictions. Les modèles de covariance existants (paragraphe 1.2.2) s'appliquent soit à des variables dont la distribution est gaussienne, soit à des variables dont la distribution est conforme aux modèles isofactoriels. Tous les types de variables considérés ne rentrent pas dans ce cadre. Nous n'allons donc pas modéliser la structure de dépendance entre les variables elles-mêmes. Chaque variable va être modélisée suivant un modèle linéaire généralisé spatial et nous allons nous intéresser à la structure de dépendance existant entre les composantes spatiales apparaissant dans l'écriture du modèle linéaire généralisé associé à chacune d'elles. Ces composantes spatiales sont toutes gaussiennes et les modèles de covariance classiques peuvent alors être appliqués. Le modèle spatial multivarié proposé fait appel à une matrice de covariance construite à partir de la méthode moyenne mobile. L'approche moyenne mobile est privilégiée car elle s'avère être plus souple que les modèles de covariance classiques. De plus, elle conduit théoriquement à des modèles valides si les fonctions moyennes mobiles sont bien choisies.

Le modèle spatial multivarié proposé est basé sur une approche hiérarchique semblable à celle proposée par Wolpert et Ickstadt (1998). L'approche hiérarchique permet de simplifier le problème complexe consistant à traiter simultanément des variables de différente nature en le divisant en une succession de problèmes plus simples conditionnellement indépendants (Wikle, 2003; Banerjee et al., 2004). C'est sous cette forme que le modèle est présenté ci-dessous.

1.3.2 Description du modèle

Le modèle hiérarchique spatial multivarié peut être défini pour un nombre quelconque K de variables. Ici, nous nous limitons à un champ aléatoire composé de trois variables de nature différente : une variable gaussienne, une variable de Poisson et une variable ordinale. Le traitement de variables nominales peut également être envisagé mais complexifie le modèle. Par souci de clarté, le cas de la variable nominale est présenté séparément en annexe A.

Soient $\mathbf{s}_1, \ldots, \mathbf{s}_n$ les n sites échantillonnés. On désigne par $Y_1(\mathbf{s})$ la variable gaussienne au point $\mathbf{s}$, par $Y_2(\mathbf{s})$ la variable de Poisson au point $\mathbf{s}$ et par $Y_3(\mathbf{s})$ la variable ordinale à L modalités au point $\mathbf{s}$. Soit $\mathbf{Y}_k = (Y_k(\mathbf{s}_1), \ldots, Y_k(\mathbf{s}_n))'$, $k = 1, 2, 3$ le vecteur de la variable Y_k en chacun des sites échantillonnés. Soit $\mathbf{Y} = (\mathbf{Y}_1{}', \mathbf{Y}_2{}', \mathbf{Y}_3{}')'$ le vecteur de toutes les variables mesurées en tous les sites.

Le modèle est basé sur une approche hiérarchique ; tous les niveaux de la hiérarchie sont décrits successivement.

La variable gaussienne $Y_1(\mathbf{s})$ et la variable de Poisson $Y_2(\mathbf{s})$ dépendent de variables latentes $S_1(\mathbf{s})$ et $S_2(\mathbf{s})$. Les variables $S_1(\mathbf{s})$ et $S_2(\mathbf{s})$ sont les composantes spatiales intervenant dans le modèle linéaire généralisé associé à chaque variable $Y_1(\mathbf{s})$ et $Y_2(\mathbf{s})$. Conditionnellement à $S_1(\mathbf{s})$ et $S_2(\mathbf{s})$, les variables $Y_1(\mathbf{s})$ et $Y_2(\mathbf{s})$ sont indépendantes. Pour les variables gaussiennes et de Poisson, nous suivons donc le modèle linéaire généralisé proposé par Diggle et al. (1998) :

$$Y_1(\mathbf{s}_i)|\mu_1, S_1(\mathbf{s}_i), \nu_1 \sim \mathcal{N}(\mu_1 + S_1(\mathbf{s}_i), \nu_1^2), \tag{1.10}$$

$$Y_2(\mathbf{s}_i)|\mu_2, S_2(\mathbf{s}_i) \sim \mathcal{P}\{\exp(\mu_2 + S_2(\mathbf{s}_i))\} \tag{1.11}$$

où $\mathcal{P}(\lambda)$ désigne la loi de Poisson de paramètre λ. Les paramètres μ_1 et μ_2 représentent les effets moyens des variables Y_1 et Y_2. Le paramètre ν_1^2 correspond à l'effet de pépite associé à la variable gaussienne Y_1.

La modélisation de la variable ordinale se décompose en deux étapes. La première permet de définir le modèle probit ordinal multivarié (Définition 4) et la seconde de généraliser ce modèle au cas spatial en faisant apparaître la structure du modèle linéaire généralisé commune à toutes les variables :

$$\mathbb{P}(Y_3(\mathbf{s}_i) = j|\boldsymbol{\alpha}_3, S_3(\mathbf{s}_i), \mu_3) = \mathbb{P}(Z_3(\mathbf{s}_i) \in]\alpha_{3;\,l-1}, \alpha_{3;\,l}]|S_3(\mathbf{s}_i), \mu_3), \tag{1.12}$$

$$Z_3(\mathbf{s}_i)|S_3(\mathbf{s}_i), \mu_3 \sim \mathcal{N}(\mu_3 + S_3(\mathbf{s}_i), 1). \tag{1.13}$$

$\boldsymbol{\alpha}_3 = (-\infty, 0, \alpha_{3,2}, \ldots, \alpha_{3,L-1}, +\infty)$ désigne le vecteur des seuils relatifs à la variable gaussienne sous-jacente $Z_3(\mathbf{s})$. Le paramètre μ_3 est l'effet moyen associé à la variable $Z_3(\mathbf{s})$. Conditionnellement à $S_1(\mathbf{s}_i)$ (respectivement $S_2(\mathbf{s}_i)$) et $S_3(\mathbf{s}_j)$, les variables $Y_1(\mathbf{s}_i)$ (respectivement $Y_2(\mathbf{s}_i)$) et $Y_3(\mathbf{s}_j)$ sont indépendantes. Les expressions (1.10) à (1.13) constituent le premier niveau du modèle hiérarchique.

Le champ aléatoire considéré dans ce chapitre ne comporte que trois variables, toutes de nature différente, mais il se peut que le champ aléatoire étudié comprenne plusieurs variables de même nature. Nous traiterons dans le chapitre suivant (paragraphe 2.5.2) des jeux de données simulés où deux variables sont de même nature. Si le champ aléatoire comporte plusieurs variables gaussiennes, par exemple Y_k et Y_m, la distribution conditionnelle $\mathbf{Y}_k, \mathbf{Y}_m|\mathbf{S}_k, \mathbf{S}_m, \mu_k, \mu_m, \nu_k, \nu_m$ est la loi normale multivariée

$$\mathcal{N}_{2n}\left\{\begin{pmatrix} \mu_k\mathbf{1} + \mathbf{S}_k \\ \mu_m\mathbf{1} + \mathbf{S}_m \end{pmatrix}, \begin{pmatrix} \nu_k^2 & 0 \\ 0 & \nu_m^2 \end{pmatrix} \otimes \mathbf{I}_n\right\},$$

où $\mathbf{1}$ est un vecteur de longueur n dont tous les termes sont égaux à 1, $\mathbf{I}_n$ est la matrice identité d'ordre n et où $\otimes$ désigne le produit de Kronecker. Si le champ aléatoire comprend plusieurs variables de Poisson, toutes ces variables sont conditionnellement indépendantes. Pour traiter plusieurs variables ordinales, nous généralisons la définition 4. Dans le modèle

probit ordinal multivarié classique (non spatial), la dépendance entre les variables latentes gaussiennes est décrite par la matrice de corrélation $\mathbf{R}$. De la même façon, dans le cas spatial, si le modèle comprend deux variables ordinales Y_k et Y_m, la dépendance entre les deux vecteurs sous-jacents $\mathbf{Z}_k$ et $\mathbf{Z}_m$ est décrite par une matrice de corrélation de taille $2n \times 2n$. Dans la suite, nous faisons l'hypothèse simplificatrice suivante. Nous considérons que les variables latentes gaussiennes $Z_k(\mathbf{s})$ et $Z_m(\mathbf{s})$ sont indépendantes conditionnellement à $S_k(\mathbf{s})$ et $S_m(\mathbf{s})$. Cette hypothèse revient à considérer la matrice de corrélation $\mathbf{R}$ égale à l'identité d'ordre $2n$.

Le second niveau du modèle hiérarchique permet de décrire la structure de dépendance entre les variables. La dépendance spatiale entre les processus $Y_k(.)$ est portée par les variables latentes $S_k(\mathbf{s})$, $k = 1, 2, 3$. Les variables $S_k(\mathbf{s})$ sont construites suivant la méthode moyenne mobile proposée par Ver Hoef et Barry (1998) (paragraphe 1.2.2), c'est-à-dire par convolution d'une fonction dite moyenne mobile avec un mélange de bruits blancs.

Soit V_k, $k = 1, 2, 3$ une combinaison linéaire de bruits blancs

$$V_k(\mathbf{x}|\rho_k, \boldsymbol{\Delta}_k) = \sqrt{1-\rho_k^2} W_k(\mathbf{x}) + \rho_k W_0(\mathbf{x} - \boldsymbol{\Delta}_k)$$

où $W_k(.)$, $k = 0, 1, 2, 3$ est un bruit blanc, ρ_k, $k = 1, 2, 3$ appartient à l'intervalle $[-1; 1]$ et où $\boldsymbol{\Delta}_k$ appartient à $\mathbb{R}^2$. Le processus $W_0(.)$ induit une dépendance entre les processus $V_k(.)$ puisque, pour tout $k \neq m$

$$\mathrm{Cor}\left[\int_{\mathbb{R}^2} V_k(\mathbf{x} + \boldsymbol{\Delta}_k|\rho_k, \boldsymbol{\Delta}_k)\mathrm{d}\mathbf{x}, \int_{\mathbb{R}^2} V_m(\mathbf{x} + \boldsymbol{\Delta}_m|\rho_m, \boldsymbol{\Delta}_m)\mathrm{d}\mathbf{x}\right] = \rho_k\rho_m \equiv \rho_{km}.$$

Le paramètre ρ_{km} peut être considéré comme la corrélation croisée entre les mélanges de bruits blancs V_k et V_m (Ver Hoef et Barry, 1998). Soit f_k, $k = 1, 2, 3$ une fonction moyenne mobile définie sur $\mathbb{R}^2$. Soit $\boldsymbol{\theta}_k$ le vecteur des paramètres associés à f_k. La variable aléatoire $S_k(\mathbf{s}_i)$ est définie par

$$S_k(\mathbf{s}_i) = \int_{\mathbb{R}^2} f_k(\mathbf{x} - \mathbf{s}_i|\boldsymbol{\theta}_k) V_k(\mathbf{x}|\rho_k, \boldsymbol{\Delta}_k)\mathrm{d}\mathbf{x}.$$

Les variables $S_k(\mathbf{s}_i)$, $i = 1, \ldots, n$ sont dépendantes car les processus $V_k(.)$ le sont. La distribution conditionnelle $\mathbf{S} = ({\mathbf{S}_1}', {\mathbf{S}_2}', {\mathbf{S}_3}')'$ est une loi normale multivariée de moyenne nulle et de matrice de covariance $\mathbf{C}$:

$$\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}, \boldsymbol{\Delta} \sim \mathcal{N}_{3n}(\mathbf{0}, \mathbf{C})$$

où $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ et $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$. Cela constitue le second niveau de la hiérarchie. L'un des avantages de la construction moyenne mobile est que l'expression de la matrice de covariance $\mathbf{C}$ est connue :

$$C_{kk}(\mathbf{h}) = \mathrm{Cov}[S_k(\mathbf{s}), S_k(\mathbf{s} + \mathbf{h})] = \int_{\mathbb{R}^2} f_k(\mathbf{x}) f_k(\mathbf{x} - \mathbf{h})\mathrm{d}\mathbf{x}, \tag{1.14}$$

$$C_{km}(\mathbf{h}) = \mathrm{Cov}[S_k(\mathbf{s}), S_m(\mathbf{s} + \mathbf{h})] = \rho_k\rho_m \int_{\mathbb{R}^2} f_k(\mathbf{x}) f_m(\mathbf{x} - \mathbf{h} + \boldsymbol{\Delta}_m - \boldsymbol{\Delta}_k)\mathrm{d}\mathbf{x}. \tag{1.15}$$

Selon les fonctions moyennes mobiles choisies, le calcul des intégrales peut être explicite ou non. Si le calcul n'est pas explicite, chaque élément de la matrice de covariance peut être vu comme une autocorrélation en théorie du signal ou une convolution et calculé à l'aide de la transformée de Fourier rapide (Ver Hoef et al., 2004). Nous reviendrons plus en détails sur le calcul de la covariance au paragraphe 2.1.

Tous les paramètres du modèle ne sont pas identifiables, c'est le cas en particulier du vecteur $\boldsymbol{\Delta}$. En pratique, on prend $\boldsymbol{\Delta}_1 = (0, 0)$. Tous les autres décalages $\boldsymbol{\Delta}_k$, $k = 2, 3$ sont calculés relativement à $\boldsymbol{\Delta}_1$ et $\boldsymbol{\Delta}$ se réduit alors à $(\boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$. De même, le vecteur des paramètres de corrélation $\boldsymbol{\rho}$ n'est pas identifiable. Lorsque le modèle ne comprend que deux variables, seul le produit $\rho_1\rho_2$ peut être identifié. Lorsque le modèle comprend K variables ($K > 2$), les K-uplets $(\rho_1, \ldots, \rho_K)$ et $(-\rho_1, \ldots, -\rho_K)$ conduisent à la même matrice de covariance. Dans ce cas, le signe de ρ_1 doit être fixé pour assurer l'identifiabilité des paramètres de corrélation.

Le troisième niveau du modèle hiérarchique consiste à définir les distributions *a priori* sur les paramètres. Les distributions *a priori* sur les paramètres μ_1, μ_2, μ_3 sont des lois uniformes sur $[-c; c]$, où c est une constante strictement positive très grande. Pour l'effet de pépite ν_1^2 associé à la variable aléatoire gaussienne $Y_1(\mathbf{s})$, on considère une loi *a priori* inverse gamma conjuguée $\nu_1^2 \sim \mathrm{IG}(a, b)$ où a et b sont choisis suffisamment petits pour avoir une distribution *a priori* non informative. La paramétrisation utilisée pour la loi inverse gamma est la suivante

$$\nu_1^2 \sim \mathrm{IG}(a, b) \Leftrightarrow \pi(\nu_1^2) \propto (\nu_1^2)^{-(a+1)} \exp\left(-\frac{b}{\nu_1^2}\right)$$

où π désigne une fonction de densité. On attribue des lois *a priori* indépendantes aux paramètres intervenant dans la matrice $\mathbf{C}$, $\boldsymbol{\theta}_k$, $k = 1, 2, 3$, $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ et $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$:

- $\boldsymbol{\theta}_k \sim \mathcal{U}_{[-c;\,c]\times[-c;\,c]}$, $k = 1, 2, 3$,
- $\rho_1 \sim \mathcal{U}_{[0;\,1]}$,
- $\rho_k \sim \mathcal{U}_{[-1;\,1]}$, $k = 2, 3$,
- $\boldsymbol{\Delta} \sim \mathcal{U}_{[-c;\,c]\times[-c;\,c]}$.

Les $L - 2$ seuils inconnus $\alpha_{3;\,l}$, $l = 2, \ldots, L - 1$ associés à la variable ordinale sont des valeurs ordonnées. Par conséquent, la distribution *a priori* du vecteur $(\alpha_{3;2}, \ldots, \alpha_{3;L-1})$ est une distribution ordonnée de $L-2$ variables aléatoires uniformes sur $[0; u]$, où u est une valeur réelle fixée strictement positive (Dacunha-Castelle et Duflo, 1982). Cette distribution ordonnée a pour pour fonction de répartition :

$$F(\alpha_{3;2}, \ldots, \alpha_{3;L-1}) = (L-2)! \prod_{i=2}^{L-1} F_u(\alpha_{3;\,i}),$$

où F_u désigne la fonction de répartition de la loi uniforme sur $[0; u]$.

Le modèle hiérarchique spatial multivarié proposé est basé sur les modèles linéaires généralisés spatiaux. Il permet de travailler non seulement avec des variables gaussiennes et de Poisson, mais aussi avec des variables ordinales grâce à la généralisation du modèle

probit ordinal multivarié. La prise en compte de la dépendance entre les variables ne peut s'effectuer de manière directe étant donné que certaines variables sont continues et d'autres discrètes. Cette difficulté est contournée en s'intéressant à la structure de dépendance entre les composantes spatiales apparaissant dans l'expression du modèle généralisé associé à chacune des variables plutôt qu'à la structure de dépendance entre les variables elles-mêmes. Ce changement de point de vue permet de traiter de manière unifiée les différents types de variables. Le modèle spatial multivarié est relativement complexe. L'approche hiérarchique s'avère particulièrement adaptée pour traiter ce genre de modèle de grande dimension. Elle permet, en effet, de le décomposer en une série de modèles plus simples conditionnellement indépendants. La structure hiérarchique du modèle va ensuite faciliter l'estimation des paramètres, qui fait l'objet du chapitre suivant.

Chapitre 2

Inférence du modèle hiérarchique spatial multivarié

Sommaire

L'objectif de ce chapitre est de proposer une méthode d'estimation des paramètres du modèle hiérarchique spatial multivarié décrit au paragraphe 1.3.2, ainsi qu'une méthode de prédiction des variables d'intérêt. Nous considérons ici le modèle constitué d'une variable gaussienne, d'une variable de Poisson et d'une variable ordinale. Le problème consiste, tout d'abord, à estimer les paramètres μ_k, $k = 1, 2, 3$, ν_1, $\boldsymbol{\rho}$, $\boldsymbol{\Delta}$ et $\boldsymbol{\theta}_k$, $k = 1, 2, 3$ intervenant dans le modèle. Avant de s'intéresser à la procédure d'estimation à proprement parler, il est nécessaire de savoir calculer la matrice de covariance connaissant les fonctions moyennes mobiles f_k, les vecteurs de paramètres $\boldsymbol{\theta}_k$ qui leur sont associés, le vecteur des décalages $\boldsymbol{\Delta}$ et le vecteur des corrélations $\boldsymbol{\rho}$.

2.1 Calcul de la matrice de covariance

Quelle que soit la procédure d'estimation envisagée, la matrice de covariance $\mathbf{C}$ du vecteur $\mathbf{S}$ doit être déterminée. Son calcul dépend de la forme des fonctions moyennes mobiles choisies, des vecteurs de paramètres $\boldsymbol{\theta}_k$ qui leur sont associés ainsi que des vecteurs $\boldsymbol{\Delta}$ et $\boldsymbol{\rho}$. Soit $\mathbf{h} \in \mathbb{R}^2$. La matrice de covariance $\mathbf{C}$ est définie par l'équation 1.15 que nous rappelons ici :

$$C_{kj}(\mathbf{h}) = \mathrm{Cov}[S_k(\mathbf{s}), S_j(\mathbf{s}+\mathbf{h})] = \rho_k \rho_j \int_{\mathbb{R}^2} f_k(\mathbf{x}) f_j(\mathbf{x} - \mathbf{h} + \boldsymbol{\Delta}_j - \boldsymbol{\Delta}_k) \mathrm{d}\mathbf{x}. \tag{2.1}$$

avec $\rho_k \rho_m = \rho_{km}$ et $\rho_{kk} = 1$ par convention. Dans la suite, nous prenons $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_j = (0, 0)$, c'est-à-dire que la covariance croisée est considérée comme symétrique. Soit les fonctions moyennes mobiles choisies sont suffisamment simples et le calcul des intégrales intervenant dans la matrice de covariance (équation 2.1) est explicite. La matrice de covariance $\mathbf{C}$ est alors entièrement connue. C'est le cas, par exemple, lorsque les fonctions moyennes mobiles sont gaussiennes ou constantes par morceaux. Soit il n'est pas possible de calculer explicitement ces intégrales. Il faut alors les approcher à l'aide de méthodes numériques. On distingue deux types de méthodes d'approximation : les méthodes basées sur des techniques d'échantillonnage dites méthodes de Monte Carlo et les méthodes basées sur les concepts de la théorie du signal.

2.1.1 Méthodes d'approximation numériques de la matrice de covariance basées sur les méthodes de Monte Carlo

Les méthodes de Monte Carlo recouvrent une série de techniques destinées à résoudre des problèmes complexes le plus souvent déterministes, par l'introduction d'échantillonnages aléatoires. Nous avons recours à ces méthodes lorsque le problème auquel nous sommes confrontés n'admet pas de solution analytique ou lorsque la dimension du problème est trop importante pour qu'il soit résolu dans un temps acceptable. Les méthodes de Monte Carlo permettent notamment d'approcher des intégrales multiples s'écrivant sous la forme :

$$\mathcal{I} = \mathbb{E}_g[h(x)] = \int h(x) g(x) \mathrm{d}\lambda(x) \tag{2.2}$$

où g est la densité de la variable aléatoire x par rapport à la mesure λ et h est une fonction quelconque. La mesure λ est, en général, la mesure de Lebesgue. Ces méthodes consistent à réaliser des simulations de la variable aléatoire pour obtenir une approximation de l'intégrale qui converge lorsque le nombre de simulations est suffisamment grand.

Méthode d'approximation numérique de la matrice de covariance basée sur la méthode de Monte Carlo standard

Le principe de la méthode de Monte Carlo standard est le suivant.

Si $x_1, x_2, \ldots, x_N$ sont indépendants et identiquement distribués suivant la loi g, alors, d'après la loi des grands nombres, la moyenne empirique $\widehat{\mathcal{I}}_N$ des $h(x_i)$ converge presque sûrement vers $\mathcal{I}$ quand N tend vers l'infini :

$$\widehat{\mathcal{I}}_N = \frac{1}{N}\sum_{i=1}^{N} h(x_i) \xrightarrow{p.s.} \mathcal{I} = \int h(x)g(x)\mathrm{d}x.$$

De plus, si $\mathbb{E}_g[|h(x)|^2] < \infty$, le théorème central limite permet de montrer que :

$$\sqrt{N}\frac{(\widehat{\mathcal{I}}_N - \mathcal{I})}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ où } \sigma^2 = \mathrm{Var}_g[h(x)].$$

La vitesse de convergence est donc de l'ordre de $\frac{\sigma}{\sqrt{N}}$. Cette vitesse peut paraître faible en dimension 1, mais présente l'avantage d'être insensible à la dimension et de ne pas dépendre de la régularité de la fonction g, pourvu que g soit de carré intégrable (Marin et Robert, 2007).

Ici, nous souhaitons approcher chaque élément de la matrice de covariance $\mathbf{C}$ (équation 2.1) avec $\Delta_k = \Delta_j = (0,0)$. Nous utilisons la méthode de Monte Carlo standard pour calculer l'intégrale en considérant $\mathbf{x}$ comme un vecteur aléatoire.

1. Nous échantillonnons uniformément N points $\mathbf{x}_i$ dans la zone où f_k est non nulle. Soit $\mathcal{A}$ l'aire de cette zone.
2. Nous approchons le terme $C_{kj}(\mathbf{h})$ de la matrice de covariance par :

$$C_{kj}(\mathbf{h}) \approx \frac{\rho_k \rho_j \mathcal{A}}{N}\sum_{i=1}^{N} f_k(\mathbf{x}_i) f_j(\mathbf{x}_i - \mathbf{h})$$

et le terme $C_{kk}(\mathbf{h})$ par :

$$C_{kk}(\mathbf{h}) \approx \frac{\mathcal{A}}{N}\sum_{i=1}^{N} f_k(\mathbf{x}_i) f_k(\mathbf{x}_i - \mathbf{h}).$$

Méthode d'approximation numérique de la matrice de covariance basée sur l'échantillonnage préférentiel

Pour améliorer la vitesse de convergence, la méthode de Monte Carlo standard a été modifiée en remarquant que l'intégrale $\mathbb{E}_g[h(x)]$ (équation 2.2) peut aussi s'écrire sous la forme :

$$\mathbb{E}_g[h(x)] = \int \frac{h(x)g(x)}{\tilde{g}(x)}\tilde{g}(x)\mathrm{d}x, \tag{2.3}$$

où la fonction $\tilde{g}$ est une loi de densité. Autrement dit, $\mathbb{E}_g[h(x)] = \mathbb{E}_{\tilde{g}}\left[\frac{h(y)g(y)}{\tilde{g}(y)}\right]$ où la variable aléatoire y a pour densité $\tilde{g}$.

Une seconde méthode pour approcher la quantité $\mathbb{E}_g[h(x)]$ consiste à utiliser N tirages de la variable Y, $y_1, \ldots, y_N$ et à calculer l'estimateur de $\mathcal{I}$:

$$\widehat{\mathcal{I}}_N = \frac{1}{\sum_{i=1}^N w(y_i)} \sum_{i=1}^N h(y_i)w(y_i) \text{ où } w(y_i) = \frac{g(y_i)}{\tilde{g}(y_i)}.$$

Cette méthode est appelée échantillonnage préférentiel (« importance sampling » en anglais). La vitesse de convergence de cette méthode est meilleure que celle de la méthode standard si

$$\mathrm{Var}\left[\frac{h(y)g(y)}{\tilde{g}(y)}\right] < \sigma^2.$$

La seconde méthode d'approximation de la matrice de covariance $\mathbf{C}$ s'inspire de l'échantillonnage préférentiel. Elle est couplée à une procédure d'estimation basée sur un algorithme d'acceptation-rejet. A l'itération t, le vecteur de paramètres $\boldsymbol{\theta}^{(t-1)}$ est connu. Un nouveau vecteur candidat $\boldsymbol{\theta}^*$ est proposé. La matrice de covariance $\mathbf{C}$ doit être approchée connaissant $\boldsymbol{\theta}^*$, en remarquant qu'à l'itération t, chacun de ces termes peut s'écrire sous la forme :

$$\begin{aligned}
\mathcal{I}_{kj}^{(t)}(\mathbf{h}) &= \int f_k(\mathbf{x};\boldsymbol{\theta}^*) f_j(\mathbf{x}-\mathbf{h};\boldsymbol{\theta}^*)\mathrm{d}\mathbf{x} \\
&= \int \frac{f_k(\mathbf{x};\boldsymbol{\theta}^*)}{f_k(\mathbf{x};\boldsymbol{\theta}^{(t-1)})} f_j(\mathbf{x}-\mathbf{h};\boldsymbol{\theta}^*) f_k(\mathbf{x};\boldsymbol{\theta}^{(t-1)})\mathrm{d}\mathbf{x} \\
\mathcal{I}_{kj}^{(t)}(\mathbf{h}) &= \mathbb{E}_{f_k(.,\boldsymbol{\theta}^{(t-1)})}\left[\frac{f_k(\mathbf{x};\boldsymbol{\theta}^*)}{f_k(\mathbf{x};\boldsymbol{\theta}^{(t-1)})} f_j(\mathbf{x}-\mathbf{h};\boldsymbol{\theta}^*)\right].
\end{aligned}$$

On reconnaît ici une équation de la forme 2.3 où la fonction d'importance $\tilde{g}$ est la distribution $f_k(.;\boldsymbol{\theta}^{(t-1)})$. N réalisations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ du vecteur aléatoire $\mathbf{x}$ sont simulées suivant la loi $f_k(.;\boldsymbol{\theta}^{(t-1)})$. L'intégrale $\mathcal{I}_{kj}$ est approchée par :

$$\widehat{\mathcal{I}}_{kj}^{(t)}(\mathbf{h}) \approx \frac{1}{\sum_{i=1}^N w(\mathbf{x}_i)} \sum_{i=1}^N f_j(\mathbf{x}_i-\mathbf{h};\boldsymbol{\theta}^*)w(\mathbf{x}_i) \text{ où } w(\mathbf{x}_i) = \frac{f_k(\mathbf{x}_i;\boldsymbol{\theta}^*)}{f_k(\mathbf{x}_i;\boldsymbol{\theta}^{(t-1)})}$$

et

$$C_{kj}^{(t)}(\mathbf{h}) \approx \rho_k\rho_j\widehat{\mathcal{I}}_{kj}^{(t)}(\mathbf{h}).$$

On effectue un raisonnement similaire pour donner une approximation de $C_{kk}^{(t)}(\mathbf{h})$. La matrice de covariance $\mathbf{C}$ est donc déterminée grâce à l'échantillonnage préférentiel. L'algorithme d'acceptation-rejet permet alors de déterminer si le nouveau candidat $\boldsymbol{\theta}^*$ est accepté ou non. Les différentes étapes sont ensuite réitérées.

2.1.2 Méthode d'approximation numérique de la matrice de covariance basée sur la transformée de Fourier rapide

Chaque terme de la matrice de covariance (équation 2.1) est défini à partir d'une intégrale. Ces intégrales sont des convolutions ; c'est pourquoi Ver Hoef et al. (2004) préconisent d'utiliser la transformée de Fourier rapide (FFT) pour les approcher. En effet, la transformée de Fourier transforme la convolution en un produit. A partir du moment où on dispose d'un algorithme efficace pour calculer la transformée de Fourier, ce qui est la cas avec la FFT, il est alors numériquement plus rapide de calculer une transformée de Fourier, c'est-à-dire un produit, et une transformée de Fourier inverse que de calculer directement la convolution.

La transformée de Fourier rapide est un algorithme de calcul de la transformée de Fourier discrète (Brigham, 1974; Press et al., 1992). Cet algorithme est récursif : il subdivise une transformée de Fourier discrète de taille composite $N = N_1 N_2$ en plusieurs transformées de Fourier discrètes de tailles inférieures N_1 et N_2. La transformée de Fourier rapide permet de ramener le calcul de la transformée de Fourier discrète de N^2 à $N \log N$ opérations ; cette réduction de complexité facilite la résolution de nombreux problèmes.

On désigne par TFf la transformée de Fourier d'une fonction f et par $TF^{-1}f$ sa transformée de Fourier inverse. $\bar{f}$ désigne le complexe conjugué de f. Nous souhaitons approcher l'intégrale multiple figurant dans l'équation 2.1. Pour plus de clarté, nous décrivons le principe de l'approximation grâce à la transformée de Fourier rapide pour une intégrale en dimension 1 avant de le généraliser en dimension 2.

Considérons l'intégrale simple suivante :

$$\mathcal{I}_{kj}(h) = \int_{\mathbb{R}} f_k(u|\boldsymbol{\theta}_k) f_j(u-h|\boldsymbol{\theta}_j) \mathrm{d}u$$

où f_j et f_k sont des fonction moyennes mobiles définies sur $\mathbb{R}$ de paramètres respectifs $\boldsymbol{\theta}_j$ et $\boldsymbol{\theta}_k$. Il est possible de redéfinir cette intégrale en terme de produit de convolution :

$$\begin{aligned}
\mathcal{I}_{kj}(h) &= \int_{\mathbb{R}} f_k(u|\boldsymbol{\theta}_k) f_j(u-h|\boldsymbol{\theta}_j) \mathrm{d}u \\
&= \int_{\mathbb{R}} f_k(y+h|\boldsymbol{\theta}_k) f_j(y|\boldsymbol{\theta}_j) \mathrm{d}y \\
&= \int_{\mathbb{R}} f_k(y+h|\boldsymbol{\theta}_k) \overline{f_j}(y|\boldsymbol{\theta}_j) \mathrm{d}y \\
\mathcal{I}_{kj}(h) &= (f_j \star f_k)(h) \qquad (2.4)
\end{aligned}$$

où $\star$ désigne le produit de convolution. On applique la transformée de Fourier aux deux membres de l'équation 2.4 :

$$TF[\mathcal{I}_{kj}](v) = TF[f_j \star f_k](v).$$

En appliquant le théorème de corrélation croisée (Bracewell, 1965; Press et al., 1992), on obtient :

$$TF[\mathcal{I}_{kj}](v) = TFf_k(v)\overline{TFf_j}(v).$$

On applique alors la transformée de Fourier inverse à chacun des membres, d'où :

$$\mathcal{I}_{kj}(h) = TF^{-1}[TFf_k\overline{TFf_j}](h).$$

Soient $c \in]0;+\infty[$, $M \in \mathbb{N}^*$ et $p \in \mathbb{Z}$. Soit $h_p^* = \dfrac{2pc}{M}$. En développant l'expression de la transformée de Fourier inverse et en considérant la version discrète de la transformée de Fourier pour TFf_k et $\overline{TFf_j}$, on obtient finalement une approximation de l'intégrale :

$$\mathcal{I}_{kj}(h_p^*) \approx \frac{1}{M}\sum_{m=0}^{M-1} TFf_k[m]\overline{TFf_j}[m]e^{\frac{i2\pi mp}{M}}.$$

Considérons maintenant l'intégrale multiple suivante :

$$\mathcal{I}_{kj}(h) = \int_{\mathbb{R}^2} f_k(\mathbf{u}|\boldsymbol{\theta}_k)f_j(\mathbf{u}-\mathbf{h}|\boldsymbol{\theta}_j)\mathrm{d}\mathbf{u}$$

où f_j et f_k sont des fonctions moyennes mobiles définies sur $\mathbb{R}^2$ de paramètres respectifs $\boldsymbol{\theta}_j$ et $\boldsymbol{\theta}_k$. Soient $d \in]0;+\infty[$, $N \in \mathbb{N}^*$ et $q \in \mathbb{Z}$. Soit $h_q^* = \dfrac{2qd}{N}$. Par un raisonnement similaire à celui effectué en dimension 1, Ver Hoef et al. (2004) montrent que l'intégrale peut être approchée par :

$$\mathcal{I}_{kj}(h_p^*, h_q^*) \approx \frac{1}{MN}\sum_{n=0}^{N-1}\left[\sum_{m=0}^{M-1} TFf_k\left[\frac{m}{c},\frac{n}{d}\right]\overline{TFf_j}\left[\frac{m}{c},\frac{n}{d}\right]e^{i2\pi mp/M}\right]e^{i2\pi nq/N}.$$

Dans le cas bidimensionnel, la transformée de Fourier discrète de f_k, TFf_k, est obtenue de la façon suivante :

$$TFf_k\left[\frac{m}{n},\frac{n}{d}\right] = \sum_{q=0}^{N-1}\left[\sum_{p=0}^{M-1} f_k[x_p^+, y_p^+]e^{-i2\pi mp/M}\right]e^{-i2\pi nq/N}$$

pour $m = 0,1,...,M-1$ et $n = 0,1,...,N-1$, où $x_p^+ = \dfrac{2pc - c(M-1)}{M}$, pour $p = 0,1,...,M-1$ et $y_q^+ = \dfrac{2qd - d(N-1)}{N}$, pour $q = 0,1,...,N-1$. En pratique, les transformées de Fourier discrètes de f_j et f_k sont calculées à l'aide de la transformée de Fourier rapide.

Il est donc possible de déterminer la matrice de covariance en tous les points de la grille de coordonnées (h_p^*, h_q^*). L'ensemble de ces points constitue une grille de référence. La covariance et la covariance croisée définies par l'équation 2.1 sont approchées respectivement par les quantités suivantes :

$$C_{kk}(h_p^*, h_q^*) \approx \frac{1}{MN} \sum_{n=0}^{N-1} \left[\sum_{m=0}^{M-1} TFf_k \left[\frac{m}{c}, \frac{n}{d} \right] \overline{TFf_k} \left[\frac{m}{c}, \frac{n}{d} \right] e^{i2\pi mp/M} \right] e^{i2\pi nq/N},$$

$$C_{kj}(h_p^*, h_q^*) \approx \frac{\rho_k \rho_j}{MN} \sum_{n=0}^{N-1} \left[\sum_{m=0}^{M-1} TFf_k \left[\frac{m}{c}, \frac{n}{d} \right] \overline{TFf_j} \left[\frac{m}{c}, \frac{n}{d} \right] e^{i2\pi mp/M} \right] e^{i2\pi nq/N}.$$

Il faut maintenant déterminer la covariance en tout point $\mathbf{h} = (h_1, h_2)$ de $\mathbb{R}^2$. Soient $h_{1L} = \left\lfloor \frac{Mh_1}{2c} \right\rfloor \frac{2c}{M}$, $h_{1U} = \left\lceil \frac{Mh_1}{2c} \right\rceil \frac{2c}{M}$, $h_{2L} = \left\lfloor \frac{Nh_2}{2d} \right\rfloor \frac{2d}{N}$, $h_{2U} = \left\lceil \frac{Nh_2}{2d} \right\rceil \frac{2d}{N}$, où $\lfloor x \rfloor$ désigne la partie entière du réel x et $\lceil x \rceil = \lfloor x \rfloor + 1$. Soient $\alpha_1 = \frac{(h_1 - h_{1L})M}{2c}$ et $\alpha_2 = \frac{(h_2 - h_{2L})N}{2d}$. Les points (h_{1L}, h_{2L}), (h_{1L}, h_{2U}), (h_{1U}, h_{2L}) et (h_{1U}, h_{2U}) appartiennent à la grille de référence. On définit C_{kj} en tout point $\mathbf{h}$ comme interpolation linéaire des quatre valeurs de covariance $C_{kj}((h_{1L}, h_{2L})')$, $C_{kj}((h_{1L}, h_{2U})')$, $C_{kj}((h_{1U}, h_{2L})')$, $C_{kj}((h_{1U}, h_{2U})')$ (Schéma 2.1). Finalement, chaque terme de la matrice de covariance est approché par :

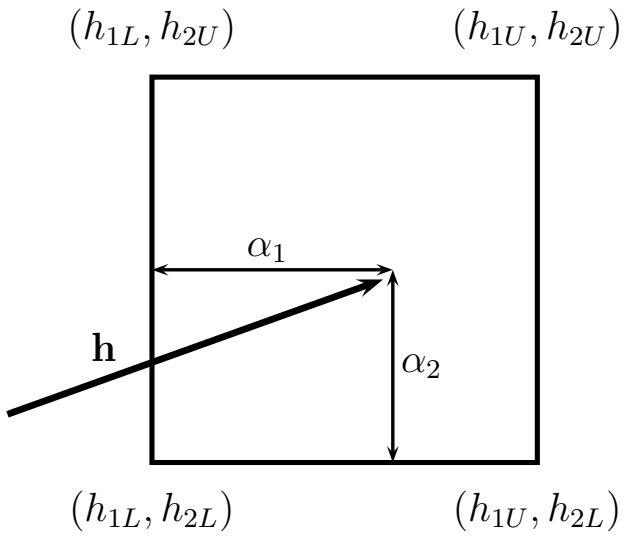


FIG. 2.1 – Schéma expliquant l'approximation du terme $C_{km}(\mathbf{h})$

$$\begin{aligned} C_{kj}(\mathbf{h}) \approx\ & (1-\alpha_1)(1-\alpha_2)C_{kj}((h_{1L},h_{2L})') + (1-\alpha_1)\alpha_2 C_{kj}((h_{1L},h_{2U})') \\ & + \alpha_1(1-\alpha_2)C_{kj}((h_{1U},h_{2L})') + \alpha_1\alpha_2 C_{kj}((h_{1U},h_{2U})'). \end{aligned}$$

La taille de la grille de référence ($M \times N$) choisie pour le calcul des $C_{kj}(h_p^*, h_q^*)$ est importante. Plus elle comporte de points, plus l'approximation de $C_{kj}(\mathbf{h})$ par interpolation est précise, mais plus le temps de calcul est long.

Les trois méthodes d'approximation décrites ci-dessus ont été implémentées. En pratique, seule la méthode d'approximation de la matrice de covariance basée sur la transformée de Fourier rapide est opérationnelle. Le calcul de la matrice de covariance par la méthode de Monte Carlo standard ou par échantillonnage préférentiel s'avère moins efficace. En effet, ces deux méthodes conduisent fréquemment à déterminer une approximation de la matrice de covariance qui n'est pas définie positive, lorsque l'échantillon utilisé n'est pas de taille suffisante. Obtenir une bonne approximation de la matrice de covariance nécessite de manipuler des échantillons de grande taille, c'est-à-dire, en général, de taille supérieure à 50 000. Le temps de calcul est alors beaucoup plus long avec ces méthodes (une à plusieurs minutes suivant la taille de l'échantillon pour la méthode de Monte Carlo standard implémentée en C) qu'avec celle basée sur la transformée de Fourier rapide (quelques centièmes de secondes avec une implémentation en R). Il n'est donc pas possible d'utiliser les méthodes d'approximation basées sur l'intégration de Monte Carlo lorsque le calcul de la covariance doit être effectué un grand nombre de fois. Ainsi, si les fonctions moyennes mobiles choisies ne permettent pas de déterminer explicitement la matrice de covariance, nous utiliserons la méthode d'approximation basée sur la transformée de Fourier rapide pour la calculer.

2.2 Les méthodes de Monte Carlo par chaînes de Markov

Le modèle hiérarchique spatial multivarié décrit au paragraphe 1.3.2 est relativement complexe. Il n'est pas possible d'écrire la vraisemblance d'un tel modèle, excepté si toutes les variables étudiées sont gaussiennes. Dans ce cas, les paramètres du modèle peuvent être estimés par la méthode du maximum de vraisemblance restreint (Ver Hoef et al., 2004). Ici, le modèle traitant de variables de nature différente, les méthodes du maximum de vraisemblance ne s'appliquent pas. Les paramètres sont estimés grâce à une approche bayésienne.

Considérons un modèle paramétrique. On désigne par $\mathbf{x}$ le vecteur des paramètres de ce modèle et par $\mathbf{y}$ le vecteur des observations. Alors que les paramètres sont considérés comme fixes dans les méthodes statistiques fréquentistes, ces derniers sont traités comme des quantités aléatoires dans les approches bayésiennes (Parent et Bernier, 2007). Nous nous intéressons donc à la distribution de ces paramètres. L'estimation bayésienne consiste, tout d'abord, à traduire la connaissance que l'on a du phénomène étudié, et donc des paramètres $\mathbf{x}$ qui lui sont associés, par une loi de distribution sur les paramètres appelée loi *a priori* et notée $\pi(\mathbf{x})$. On se donne ensuite un modèle statistique paramétré représentant le mécanisme aléatoire de génération des données $\mathbf{y}$ connaissant les paramètres $\mathbf{x}$ et la

vraisemblance qui lui est associée $f(\mathbf{y}|\mathbf{x})$. L'information *a priori* $\pi(\mathbf{x})$ est alors actualisée au vu de l'information contenue dans les observations, c'est-à-dire que l'on détermine la distribution des paramètres $\mathbf{x}$ connaissant les données $\mathbf{y}$. Cette loi est appelée loi *a posteriori* sur les paramètres. Elle est obtenue en utilisant la version continue du théorème de Bayes :

$$\pi(\mathbf{x}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})}{\int_{\boldsymbol{\chi}} f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x}}$$

où χ désigne l'ensemble de définition des paramètres.

L'inférence statistique bayésienne consiste à déterminer les caractéristiques (moyenne, médiane, mode, etc) de la distribution *a posteriori* des paramètres $\mathbf{x}$. La moyenne *a posteriori* est choisie ici comme estimateur des paramètres $\mathbf{x}$. Notons que la médiane ou le mode *a posteriori* auraient également pu être choisis comme estimateur. La moyenne *a posteriori* est l'espérance de $\mathbf{x}$ sous la loi *a posteriori* $\pi(\mathbf{x}|\mathbf{y})$:

$$\widehat{\mathbf{x}} = \mathbb{E}[\mathbf{x}|\mathbf{y}] = \int_{\boldsymbol{\chi}} \mathbf{x}\pi(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}. \tag{2.5}$$

Pour les modèles les plus simples, la loi *a priori* $\pi(\mathbf{x})$ est conjuguée (Robert, 1992), c'est-à-dire que la loi *a priori* et la loi *a posteriori* appartiennent à la même famille paramétrique de lois. La moyenne *a posteriori* peut alors être calculée analytiquement. Pour les modèles plus complexes, notamment pour les modèles en grande dimension, la loi *a priori* a généralement une structure quelconque et les propriétés induites par les lois conjuguées ne peuvent pas être exploitées (Parent et Bernier, 2007). Des méthodes numériques comme les méthodes de Monte Carlo et les méthodes de Monte Carlo par chaînes de Markov (MCMC) sont alors utilisées pour approcher la moyenne *a posteriori*.

Comme nous l'avons vu au paragraphe 2.1.1, approcher l'intégrale figurant dans l'équation 2.5 par la méthode de Monte Carlo standard nécessite de savoir générer des réalisations de la loi *a posteriori* $\pi(\mathbf{x}|\mathbf{y})$. En général, cette loi est trop complexe pour pouvoir être simulée de manière directe. L'échantillonnage préférentiel, quant à lui, ne nécessite pas de savoir simuler suivant la loi *a posteriori*. En revanche, il requiert de connaître cette loi de manière exacte. Or la loi *a posteriori* est le plus souvent connue à une constante de proportionnalité près. L'échantillonnage préférentiel ne s'applique donc que dans un nombre restreint de cas. Les méthodes d'approximation numériques les plus utilisées en inférence bayésienne sont donc les méthodes de Monte Carlo par chaînes de Markov (Robert, 1996).

L'intégrale définissant la moyenne *a posteriori* (équation 2.5) et que nous souhaitons approchée est une intégrale du type $\mathcal{I} = \int h(\mathbf{x})g(\mathbf{x})\mathrm{d}\mathbf{x}$ où la fonction h est ici la fonction identité et où la densité g est la loi *a posteriori* $\pi(\mathbf{x}|\mathbf{y})$. Les méthodes MCMC permettent d'approcher ce type d'intégrales sans qu'il soit nécessaire de simuler suivant la densité g. Le principe des méthodes MCMC repose sur la construction d'une chaîne de Markov ergodique dont la loi stationnaire est g. Soit $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$ les états pris successivement par

la chaîne de Markov $(\mathbf{x}^{(t)})_{t\in\mathbb{N}}$ de loi stationnaire g. L'intégrale $\mathcal{I}$ est alors estimée par :

$$\widehat{\mathcal{I}}_N = \frac{1}{N}\sum_{i=T-N+1}^{T} h(\mathbf{x}^{(i)}).$$

Le théorème ergodique garantit la convergence de la quantité $\widehat{\mathcal{I}}_N$ vers $\mathbb{E}_g[h(\mathbf{x})]$ quand N tend vers l'infini, quelle que soit la valeur initiale $\mathbf{x}^{(0)}$ de la chaîne (Marin et Robert, 2007).

Il existe différentes méthodes MCMC. Trois d'entre elles sont présentées ci-dessous : l'algorithme de Metropolis-Hastings, l'échantillonneur de Gibbs et une version adaptative l'algorithme de Langevin-Hastings tronqué.

2.2.1 L'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings est un algorithme MCMC. Cet algorithme développé par Hastings (1970) est une généralisation de l'algorithme de Metropolis (Metropolis et al., 1953). L'algorithme de Metropolis-Hastings peut être considéré comme une extension des algorithmes de simulation standard comme les méthodes d'acceptation-rejet qui sont toutes basées sur l'utilisation d'une loi de proposition. La loi de proposition de l'algorithme de Metropolis-Hastings a la particularité d'être markovienne (Marin et Robert, 2007). On la note $q(x, y)$. L'algorithme 1 décrit l'algorithme de Metropolis-Hastings permettant de simuler des réalisations de la loi cible π.

Pseudo-code 1 Algorithme de Metropolis-Hastings

Entrées : $\mathbf{x}^{(0)}$ (valeur initiale de la chaîne de Markov), T (nombre d'itérations)
Sorties : $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$
  pour t allant de 1 à T **faire**
    Connaissant $\mathbf{x}^{(t-1)}$, générer un candidat $\tilde{\mathbf{x}} \sim q(\mathbf{x}^{(t-1)}, \mathbf{x})$.
    $r(\mathbf{x}^{(t-1)}, \tilde{\mathbf{x}}) \Leftarrow \min\left\{\dfrac{\pi(\tilde{\mathbf{x}})/q(\mathbf{x}^{(t-1)}, \tilde{\mathbf{x}})}{\pi(\mathbf{x}^{(t-1)})/q(\tilde{\mathbf{x}}, \mathbf{x}^{(t-1)})}, 1\right\}$
    Générer $U \sim \mathcal{U}_{[0;1]}$
    si $U \leq r(\mathbf{x}^{(t-1)}, \tilde{\mathbf{x}})$ **alors**
      $\mathbf{x}^{(t)} \Leftarrow \tilde{\mathbf{x}}$
    sinon
      $\mathbf{x}^{(t)} \Leftarrow \mathbf{x}^{(t-1)}$
    finsi
  fin pour
  retourner $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$

Pour mettre en œuvre l'algorithme de Metropolis-Hastings, nous n'avons besoin de connaître la loi cible π et la loi de proposition q qu'à une constante de proportionnalité près, puisque les deux constantes se compensent dans le calcul du ratio de Metropolis-Hastings r (défini dans le pseudo-code 1). La convergence de l'algorithme de Metropolis-Hastings est théoriquement garantie pour un large choix de lois de proposition q, à condition cependant

que le support de la fonction q contienne le support de la loi cible π. En pratique, le choix de la loi de proposition q influence fortement la vitesse de convergence. En effet, si la loi de proposition q ne permet que de petits déplacements dans l'espace des paramètres $\boldsymbol{\chi}$, le taux d'acceptation est élevé et la chaîne reste dans le voisinage de la valeur initiale. Au contraire, si la loi de proposition q autorise de grands sauts dans l'espace des paramètres, le taux d'acceptation est faible et la chaîne bouge difficilement. La loi de proposition q choisie doit permettre de bien explorer le support de la loi cible π. Gelman et al. (1996) ont montré que la vitesse de convergence de l'algorithme de Metropolis-Hastings est optimale si la loi de proposition q choisie conduit à un taux d'acceptation compris entre 25 % et 40 %.

2.2.2 L'échantillonneur de Gibbs

L'algorithme de Gibbs est un cas particulier de l'algorithme de Metropolis-Hastings (Gelman et al., 2004). L'algorithme de Gibbs est plus simple à mettre en œuvre et offre, en général, une meilleure vitesse de convergence que l'algorithme de Metropolis-Hastings, mais il ne s'applique que sous certaines conditions. Si la loi cible est la distribution $\pi(x_1, \ldots, x_p)$, l'algorithme de Gibbs nécessite de connaître toutes les lois conditionnelles complètes $\pi_1, \ldots, \pi_p$ où π_j désigne la distribution de x_j conditionnellement à $\mathbf{x}_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$. L'échantillonneur de Gibbs modifie tour à tour les composantes du vecteur de paramètres $\mathbf{x}$ en simulant de nouvelles valeurs suivant chacune des lois conditionnelles complètes (Marin et Robert, 2007). La description de l'algorithme de Gibbs est donnée par l'algorithme 2.

Pseudo-code 2 Algorithme de Gibbs

Entrées : $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_p^{(0)})$ (valeur initiale de la chaîne de Markov), T (nombre d'itérations)
Sorties : $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$

pour t allant de 1 à T **faire**
  Générer $x_1^{(t)} \sim \pi_1(x_1 | x_2^{(t-1)}, \ldots, x_p^{(t-1)})$.
  Générer $x_2^{(t)} \sim \pi_2(x_2 | x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)})$.
  $\vdots$
  Générer $x_p^{(t)} \sim \pi_p(x_p | x_1^{(t)}, \ldots, x_{p-1}^{(t)})$.
fin pour
retourner $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$

Lorsque le modèle est complexe, il arrive fréquemment que toutes les lois conditionnelles complètes π_j, $j = 1, \ldots, p$ ne puissent pas être simulées de manière directe. Dans ce cas, nous faisons appel à un algorithme MCMC hybride appelé algorithme de « Metropolis-Hastings within Gibbs ». La structure de l'algorithme de Gibbs tel qu'il est présenté dans le pseudo-code 2 est conservée, mais les lois conditionnelles complètes ne relevant pas des propriétés de conjugaison sont simulées grâce à une étape de l'algorithme de Metropolis-Hastings (Banerjee et al., 2004). Le remplacement d'une simulation de π_j par une étape de Metropolis-Hastings ne modifie pas la loi stationnaire de la chaîne, et est donc valable d'un point de vue théorique (Robert, 2006). La convergence de ce type d'algorithme peut

être lente si les paramètres sont fortement corrélés (Gelman et al., 2004).

2.2.3 Version adaptative de l'algorithme de Langevin-Hastings tronqué

L'algorithme de Langevin-Hastings, appelé aussi en anglais « Metropolis-adjusted Langevin algorithm » (MALA), a été introduit par Besag (1994), puis étudié plus en détails par Roberts et Tweedie (1996). Cet algorithme est un algorithme de Metropolis-Hastings pour lequel la loi de proposition est donnée par :

$$\mathcal{N}_d\left(\mathbf{x} + \frac{\sigma^2}{2}\nabla \ln \pi(\mathbf{x}), \sigma^2 \mathrm{I}_d\right)$$

où d est la dimension de l'espace des paramètres $\boldsymbol{\chi}$ et où le terme $\nabla \ln \pi(\mathbf{x})$, appelé dérive, désigne le gradient de $\ln \pi(\mathbf{x})$. L'utilisation du gradient dans la loi de proposition permet d'obtenir de meilleures propriétés de convergence qu'avec un algorithme de Metropolis-Hastings où la loi de distribution serait $\mathcal{N}_d(\mathbf{x}, \sigma^2 \mathrm{I}_d)$ (Christensen et al., 2001; Christensen et Waagepetersen, 2002). La variance de proposition σ^2 ($\sigma > 0$) est spécifiée par l'utilisateur (Møller et Waagepetersen, 2004). Des résultats théoriques obtenus par Roberts et Rosenthal (1998) et Breyer et Roberts (2000) suggèrent de choisir σ de façon à obtenir un taux d'acceptation 0,574. Pour éviter des problèmes de dégénérescence dans le taux de convergence de l'algorithme, on se limite, en général, à l'utilisation de dérives qui soient des fonctions bornées. En pratique, la façon la plus simple d'obtenir une fonction de dérive bornée est de tronquer la fonction prise pour dérive. Nous obtenons alors une version tronquée de l'algorithme MALA en remplaçant la quantité $\nabla \ln \pi(\mathbf{x})$ dans la loi de proposition par :

$$D_{MALA}(\mathbf{x}) = \frac{\delta}{\max(\delta, |\nabla \ln \pi(\mathbf{x})|)} \nabla \ln \pi(\mathbf{x})$$

où $\delta > 0$ est une constante fixée.

Quel que soit l'algorithme employé, algorithme de Metropolis-Hastings ou algorithme MALA, l'utilisateur doit choisir la loi de proposition et, en particulier, déterminer les paramètres qui lui sont associés. Ici, l'utilisateur doit fixer la variance de proposition σ^2. Déterminer la valeur optimale de ces paramètres n'est pas un problème facile. Les algorithmes MCMC adaptatifs (Gilks et al., 1998; Haario et al., 2001; Atchade et Rosenthal, 2005; Andrieu et Moulines, 2006) offrent une solution élégante à ce problème de calibration des paramètres. Ces algorithmes permettent, en effet, d'ajuster automatiquement les paramètres au cours du déroulement de l'algorithme. Nous nous intéressons plus particulièrement à la version adaptative de l'algorithme MALA tronqué (Atchade, 2006). Le paramètre d'échelle σ est ajusté de manière à atteindre le taux d'acceptation τ recommandé. Si, à l'itération t, le ratio de Metropolis-Hastings est plus grand que τ, la variance de proposition σ^2 est augmentée à l'itération suivante, dans le cas contraire, elle est diminuée. Nous présentons dans l'algorithme 3 une version simplifiée de l'algorithme MALA tronqué de Atchade (2006). Dans cet algorithme, la variance de proposition σ^2 appartient à l'intervalle $[\epsilon, A]$, avec $0 < \epsilon < A < \infty$, et (γ_t) désigne une suite de nombres positifs telle que $\sum \gamma_t = \infty$ et $\gamma_t = O(t^{-\lambda})$, où $\frac{1}{2} < \lambda \leq 1$. Rappelons que $\phi_d(x|\boldsymbol{\mu}, \Sigma)$ désigne la densité de la loi normale multivariée de dimension d, de moyenne $\boldsymbol{\mu}$ et de matrice de covariance Σ.

Pseudo-code 3 Algorithme MALA tronqué

Entrées : ϵ, A, τ, $\mathbf{x}^{(0)}$ (valeur initiale des paramètres), σ_0^2 (valeur initiale de la variance de proposition), T (nombre d'itérations)
Sorties : $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$
pour t allant de 1 à T **faire**
Générer un candidat $\tilde{x} \sim \mathcal{N}_d\left(x^{(t-1)} + \frac{\sigma_{t-1}^2}{2} D_{MALA}(x^{(t-1)}), \sigma_{t-1}^2 \mathrm{I}_d\right)$
$r(x^{(t-1)}, \tilde{x}) \Leftarrow \min\left\{ \frac{\pi(\tilde{x})/\phi_d(\tilde{x}|x^{(t-1)} + \frac{\sigma_{t-1}^2}{2} D_{MALA}(x^{(t-1)}), \sigma_{t-1}^2 \mathrm{I}_d)}{\pi(x^{(t-1)})/\phi_d(x^{(t-1)}|\tilde{x} + \frac{\sigma_{t-1}^2}{2} D_{MALA}(\tilde{x}), \sigma_{t-1}^2 \mathrm{I}_d)}, 1 \right\}$
Générer $U \sim \mathcal{U}_{[0;1]}$
si $U \leq r(x^{(t-1)}, \tilde{x})$ **alors**
$x^{(t)} \Leftarrow \tilde{x}$
sinon
$x^{(t)} \Leftarrow x^{(t-1)}$
finsi
$\sigma_t = \sigma_{t-1} + \gamma_t(r(x^{(t-1)}, \tilde{x}) - \tau)$
si $\sigma_t < \epsilon$ **alors**
$\sigma_t = \epsilon$
finsi
si $\sigma_t > A$ **alors**
$\sigma_t = A$
finsi
fin pour
retourner $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$

2.3 Analyse *a posteriori*

Le modèle hiérarchique spatial multivarié peut être résumé par la distribution *a posteriori* des paramètres :

$$\begin{aligned}
&\pi(\mu_1, \mu_2, \mu_3, \mathbf{S}, \mathbf{Z}_3, \nu_1, \boldsymbol{\alpha}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho} | \mathbf{Y}) \\
&\quad \propto \pi(\mathbf{Y}_1|\mu_1, \mathbf{S}_1, \nu_1)\pi(\mathbf{Y}_2|\mu_2, \mathbf{S}_2)\mathbb{P}[\mathbf{Y}_3|\mathbf{Z}_3, \boldsymbol{\alpha}_3]\pi(\mathbf{Z}_3|\mu_3, \mathbf{S}_3)\pi(\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}) \\
&\quad \pi(\mu_1)\pi(\mu_2)\pi(\mu_3)\pi(\nu_1^2)\pi(\boldsymbol{\alpha}_3)\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_3)\pi(\boldsymbol{\rho})
\end{aligned}$$

$$
\begin{aligned}
&\pi(\mu_1, \mu_2, \mu_3, \mathbf{S}, \mathbf{Z}_3, \nu_1, \boldsymbol{\alpha}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho} | \mathbf{Y}) \\
&\quad \propto \exp\left\{ -\frac{1}{2\nu_1^2} (\mathbf{Y}_1 - \mu_1 \mathbf{1} - \mathbf{S}_1)' (\mathbf{Y}_1 - \mu_1 \mathbf{1} - \mathbf{S}_1) \right\} \\
&\quad \times \prod_{i=1}^{n} \left[\frac{\{\exp(\mu_2 + S_2(\mathbf{s}_i))\}^{Y_2(\mathbf{s}_i)} \exp\{-\exp(\mu_2 + S_2(\mathbf{s}_i))\}}{Y_2(\mathbf{s}_i)!} \right] \\
&\quad \times \prod_{i=1}^{n} \left[\exp\left\{ -\frac{1}{2} (Z_3(\mathbf{s}_i) - \mu_3 - S_3(\mathbf{s}_i))^2 \right\} \mathbb{1}(Z_3(\mathbf{s}_i) \in]\alpha_{3;Y_3(\mathbf{s}_i)-1}; \alpha_{3;Y_3(\mathbf{s}_i)}]) \right] \\
&\quad \times \exp\left\{ -\frac{1}{2} \mathbf{S}' \mathbf{C}^{-1} \mathbf{S} \right\} \pi(\mu_1) \pi(\mu_2) \pi(\mu_3) \pi(\nu_1^2) \pi(\boldsymbol{\alpha}_3) \pi(\boldsymbol{\theta}_1) \pi(\boldsymbol{\theta}_2) \pi(\boldsymbol{\theta}_3) \pi(\boldsymbol{\rho})
\end{aligned}
$$

où $\mathbb{1}$ désigne la fonction indicatrice et $\mathbf{1}$ un vecteur de longueur n où tous les termes sont égaux à 1. Cette loi est trop complexe pour pouvoir être simulée de manière directe. Nous utilisons donc les méthodes MCMC décrites au paragraphe 2.2 pour obtenir des réalisations de cette loi. Ces réalisations vont nous permettre d'estimer les paramètres en approchant la moyenne *a posteriori* (équation 2.5). Notons que la structure hiérarchique du modèle est particulièrement adaptée à la mise en œuvre de telles méthodes d'estimation.

L'algorithme MCMC utilisé pour l'estimation des paramètres du modèle est l'algorithme de « Metropolis-Hastings within Gibbs » (paragraphe 2.2.2). Les paramètres sont mis à jour successivement par tirage suivant leurs lois conditionnelles complètes. Lorsque l'une des lois conditionnelles complètes est trop complexe pour être simulée de manière directe, le paramètre correspondant est mis à jour par une étape de l'algorithme de Metropolis-Hastings (pseudo-code 1) ou de l'algorithme MALA tronqué (pseudo-code 3). Nous détaillons ci-dessous la mise à jour de chacun des paramètres.

• **Mise à jour de la moyenne μ_1 associée à la variable gaussienne Y_1** La loi *a priori* du paramètre μ_1 est une loi uniforme $\mathcal{U}_{[-c;\,c]}$. La loi de $\mathbf{Y_1}|\mathbf{S}_1, \mu_1, \nu_1$ est une loi normale multivariée de moyenne $\mu_1 \mathbf{1} + \mathbf{S}_1$ de variance $\nu_1^2 \mathbf{I}_n$. La loi conditionnelle complète de μ_1 est donc donnée par :

$$
\mu_1 | \ldots \sim \mathcal{N}\left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_1(\mathbf{s}_i) - S_1(\mathbf{s}_i)), \frac{\nu_1^2}{n} \right\} \text{ tronquée sur } [-c;\, c].
$$

• **Mise à jour de l'effet de pépite ν_1^2 associée à la variable gaussienne Y_1** La paramètre ν_1^2 a pour loi *a priori* une loi inverse gamma $IG(a, b)$. La loi *a posteriori* de ν_1^2 est une loi inverse gamma :

$$
\nu_1^2 | \ldots \sim IG\left\{ a + \frac{n}{2}, b + \frac{\sum_{i=1}^{n} (Y_1(\mathbf{s}_i) - \mu_1 - S_1(\mathbf{s}_i))^2}{n} \right\}
$$

• **Mise à jour de la moyenne μ_2 associée à la variable de Poisson** La loi *a posteriori* de μ_2 se définit à partir de la relation de Bayes suivante :

$$
\mu_2 | \ldots \propto \pi(\mathbf{Y}_2 | \mu_2, \mathbf{S}_2) \pi(\mu_2)
$$

où la loi *a priori* de $\pi(\mu_2)$ est une loi uniforme $\mathcal{U}_{[-c;c]}$. Cette loi *a posteriori* n'est pas explicite. L'algorithme de Gibbs n'est donc pas adapté pour simuler la distribution de μ_2. Le paramètre μ_2 est mis à jour grâce à une étape de l'algorithme de Metropolis-Hastings. La loi de proposition du nouveau candidat $\tilde{\mu}_2$ est une loi normale de moyenne μ_2 et variance $\sigma^2_{\mu_2}$. Le ratio de Metropolis-Hastings est donné par :

$$r(\mu_2, \tilde{\mu}_2) = \min\left\{1, \frac{\pi(\mathbf{Y}_2|\mathbf{S}_2, \tilde{\mu}_2)\pi(\tilde{\mu}_2)}{\pi(\mathbf{Y}_2|\mathbf{S}_2, \mu_2)\pi(\mu_2)} \times \frac{q(\tilde{\mu}_2, \mu_2)}{q(\mu_2, \tilde{\mu}_2)}\right\}.$$

• **Mise à jour du vecteur de la moyenne μ_3 associée à la variable ordinale Y_3** On suppose que la loi *a priori* sur la moyenne μ_3 est une loi uniforme $\mathcal{U}_{[-c;\,c]}$. La loi *a posteriori* du paramètre μ_3 est donc une loi gaussienne tronquée :

$$\mu_3|\ldots \sim \mathcal{N}\left\{\frac{1}{n}\sum_{i=1}^{n}(Z_3(\mathbf{s}_i) - S_3(\mathbf{s}_i)), \frac{1}{n}\right\} \text{ tronqueé sur}[-c;\, c].$$

• **Mise à jour de la variable latente gaussienne Z_3 associée à la variable ordinale Y_3** Nous nous intéressons à la loi des variables latentes $Z_3(\mathbf{s}_i)$. Par définition, $Z_3(\mathbf{s}_i)|S_3(\mathbf{s}_i), \mu_3 \sim \mathcal{N}(\mu_3 + S_3(\mathbf{s}_i), 1)$ et $\mathbb{P}(Y_3(\mathbf{s}_i) = l)$ si et seulement si $Z_3(\mathbf{s}_i) \in]\alpha_{3,\,l-1}, \alpha_{3,\,l}]$. On en déduit que :

$$Z_3(\mathbf{s}_i)|Y_3(\mathbf{s}_i), S_3(\mathbf{s}_i), \mu_3 \sim \mathcal{N}(\mu_3 + S_3(\mathbf{s}_i)) \text{ tronquée sur }]\alpha_{3,\,Y_3(\mathbf{s}_i)-1};\, \alpha_{3,\,Y_3(\mathbf{s}_i)}],$$

c'est-à-dire que :

$$f(Z_3(\mathbf{s}_i)|Y_3(\mathbf{s}_i), S_3(\mathbf{s}_i), \mu_3) = \\ \phi(Z_3(\mathbf{s}_i)|\mu_3 + S_3(\mathbf{s}_i), 1)\sum_{l=1}^{L}\mathbb{1}_{\{Z_3(\mathbf{s}_i)\in]\alpha_{3,\,l-1};\,\alpha_{3,\,l}]\}}\mathbb{1}_{\{Y_3(\mathbf{s}_i)=l\}}.$$

Chaque variable latente $Z_3(\mathbf{s}_i)$ est simulée de manière univariée.

• **Mise à jour du vecteur de seuils $\boldsymbol{\alpha}_3$** Nous devons déterminer la loi *a posteriori* des seuils $\boldsymbol{\alpha}_3$ afin de les mettre à jour. Rappelons que les seuils sont classés par ordre croissant. Albert et Chib (1993) proposent de simuler les seuils par échantillonnage de Gibbs suivant la distribution :

$$\alpha_{3;\,l}|\ldots \sim \mathcal{U}[\max\{\max\{Z_3(\mathbf{s}_i) : Y_3(\mathbf{s}_i) = l\}, \alpha_{3;\,l-1}\}; \\ \min\{\min\{Z_3(\mathbf{s}_i) : Y_3(\mathbf{s}_i) = l+1\}, \alpha_{3;\,l+1}\}]$$

pour $l = 2, \ldots, L-1$. Nous utilisons cette méthode pour mettre à jour les seuils. Une autre approche visant à améliorer la vitesse de convergence de l'algorithme a été proposée par Cowles (1996). Cette méthode utilise un algorithme de Metropolis-Hastings pour générer les seuils selon leur distribution marginale *a posteriori* en utilisant une loi de proposition qui soit une loi gaussienne tronquée.

Mise à jour du vecteur latent $\mathbf{S} = (\mathbf{S}_1{}', \mathbf{S}_2{}', \mathbf{S}_3{}')'$ Le vecteur latent $\mathbf{S}$ étant de grande dimension, il est mis à jour par bloc. La loi conditionnelle complète de $\mathbf{S}_1$ se décompose sous la forme :

$$\mathbf{S}_1|\ldots \propto \pi(\mathbf{Y}_1|\mathbf{S}_1, \mu_1, \nu_1^2)\pi(\mathbf{S}_1|\mathbf{S}_2, \mathbf{S}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}).$$

La loi de $\mathbf{Y}_1|\mathbf{S}_1, \mu_1, \nu_1^2$ est une loi normale multivariée $\mathcal{N}_n(\mu_1\mathbf{1}, \nu_1^2\mathbf{I}_n)$. Reste à déterminer la loi de $\mathbf{S}_1|\mathbf{S}_2, \mathbf{S}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}$. Par définition, $\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho} \sim \mathcal{N}_{3n}(\mathbf{0}, \mathbf{C})$. La distribution de $\mathbf{S}_1|\mathbf{S}_2, \mathbf{S}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}$ est une loi normale multivariée

- de moyenne $\mathbf{m}_1 = \begin{pmatrix} \mathbf{C}_{12} & \mathbf{C}_{13} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{22} & \mathbf{C}_{23} \\ \mathbf{C}_{32} & \mathbf{C}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_2 \\ \mathbf{S}_3 \end{pmatrix}$
- et de variance $\mathbf{V}_1 = \mathbf{C}_{11} - \begin{pmatrix} \mathbf{C}_{12} & \mathbf{C}_{13} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{22} & \mathbf{C}_{23} \\ \mathbf{C}_{32} & \mathbf{C}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{C}_{21} \\ \mathbf{C}_{31} \end{pmatrix}$

où $\mathbf{C}_{km}$ désigne la matrice de covariance entre le vecteur $\mathbf{S}_k$ et $\mathbf{S}_m$. Finalement, la loi conditionnelle complète du vecteur $\mathbf{S}_1$ est donnée par :

$$\mathbf{S}_1|\ldots \sim \mathcal{N}_n(\mathbf{m}_1^*, \mathbf{V}_1^*) \text{ avec } \begin{cases} \mathbf{V}_1^* = \left(\mathbf{V}_1^{-1} + \frac{1}{\nu_1^2}\mathbf{I}\right)^{-1} \\ \mathbf{m}_1^* = \mathbf{V}_1^* \left\{\mathbf{V}_1^{-1}\mathbf{m}_1 + \frac{1}{\nu_1^2}(\mathbf{Y}_1 - \mu_1\mathbf{1})\right\} \end{cases}.$$

Nous effectuons un raisonnement similaire pour déterminer la loi conditionnelle complète du vecteur $\mathbf{S}_3$. Cette loi s'écrit :

$$\mathbf{S}_3|\ldots \propto \pi(\mathbf{Z}_3|\mathbf{S}_3, \mu_3)\pi(\mathbf{S}_3|\mathbf{S}_1, \mathbf{S}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho})$$

où $\mathbf{Z}_3|\mathbf{S}_3, \mu_3 \sim \mathcal{N}_n(\mu_3\mathbf{1}+\mathbf{S}_3, \mathbf{I}_n)$. On montre que $\mathbf{S}_3|\mathbf{S}_1, \mathbf{S}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}$ suit une loi normale multivariée de

- de moyenne $\mathbf{m}_3 = \begin{pmatrix} \mathbf{C}_{31} & \mathbf{C}_{32} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{pmatrix}$
- et de variance $\mathbf{V}_3 = \mathbf{C}_{33} - \begin{pmatrix} \mathbf{C}_{31} & \mathbf{C}_{32} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{C}_{13} \\ \mathbf{C}_{23} \end{pmatrix}$.

La loi conditionnelle complète de $\mathbf{S}_3$ est la suivante :

$$\mathbf{S}_3|\ldots \sim \mathcal{N}_n(\mathbf{m}_3^*, \mathbf{V}_3^*) \text{ avec } \begin{cases} \mathbf{V}_3^* = (\mathbf{V}_3^{-1} + \mathbf{I})^{-1} \\ \mathbf{m}_3^* = \mathbf{V}_3^*\{\mathbf{V}_3^{-1}\mathbf{m}_3 + (\mathbf{Z}_3 - \mu_3\mathbf{1})\} \end{cases}.$$

Les vecteurs $\mathbf{S}_1$ et $\mathbf{S}_3$ peuvent donc être mis à jour par l'algorithme de Gibbs, ce qui n'est pas le cas du vecteur $\mathbf{S}_2$. La formule de Bayes nous donne la loi conditionnelle complète du vecteur $\mathbf{S}_2$:

$$\mathbf{S}_2|\ldots \propto \pi(\mathbf{Y}_2|\mathbf{S}_2, \mu_2)\pi(\mathbf{S}_2|\mathbf{S}_1, \mathbf{S}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}).$$

La loi de $\mathbf{S}_2|\mathbf{S}_1, \mathbf{S}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}$ est une loi normale multivariée :

$$\mathbf{S}_2|\mathbf{S}_1, \mathbf{S}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho} \sim \mathcal{N}_n(\mathbf{m}_2, \mathbf{V}_2),$$

où

- $\mathbf{m}_2 = \begin{pmatrix} \mathbf{C}_{21} & \mathbf{C}_{23} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{13} \\ \mathbf{C}_{31} & \mathbf{C}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_3 \end{pmatrix}$,

- $\mathbf{V}_2 = \mathbf{C}_{22} - \begin{pmatrix} \mathbf{C}_{21} & \mathbf{C}_{23} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{13} \\ \mathbf{C}_{31} & \mathbf{C}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{C}_{12} \\ \mathbf{C}_{32} \end{pmatrix}$.

La loi de $\mathbf{Y}_2|\mathbf{S}_2, \mu_2$ étant une loi de Poisson, la propriété de conjugaison ne s'applique pas. La loi conditionnelle complète de $\mathbf{S}_2$ ne peut donc pas être simulée de manière directe. Le vecteur $\mathbf{S}_2$ est mis à jour grâce à une étape de l'algorithme de Langevin-Hastings tronqué. Cet algorithme est préféré à l'algorithme de Metropolis-Hastings car il permet d'accélérer la vitesse de convergence. Nous utiliserons ici sa version adaptative (pseudo-code 3) qui permet de calibrer automatiquement la variance de la loi de proposition afin d'avoir une vitesse de convergence optimale. L'algorithme 3 est mis en œuvre en prenant $\tau = 0,574$, $\gamma_t = 1/t$, $\epsilon = 10^{-6}$, $A = 10^7$ et $\delta = 1000$.

• **Mise à jour du vecteur $\boldsymbol{\theta}_k$, $k = 1, 2, 3$** La loi *a posteriori* du vecteur θ_k s'obtient en appliquant la formule de Bayes :

$$\boldsymbol{\theta}_k \propto \pi(\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho})\pi(\boldsymbol{\theta}_k).$$

La loi *a priori* sur le vecteur $\boldsymbol{\theta}_k$ est une loi uniforme $\mathcal{U}_{[-c;\,c]\times[-c;\,c]}$. Nous utilisons une étape de l'algorithme de Metropolis-Hastings pour simuler le vecteur $\boldsymbol{\theta}_k$. La loi de proposition utilisée est la suivante :

$$\tilde{\boldsymbol{\theta}}_k|\boldsymbol{\theta}_k \sim \mathcal{N}_2(\boldsymbol{\theta}_k, \mathbf{U}_k) \text{ avec } \mathbf{U}_k = \begin{pmatrix} \sigma^2_{\boldsymbol{\theta}_{k_1}} & 0 \\ 0 & \sigma^2_{\boldsymbol{\theta}_{k_2}} \end{pmatrix}.$$

Autrement dit, les éléments du vecteur candidat $\tilde{\boldsymbol{\theta}}_k$ sont proposés indépendamment les uns des autres. Le ratio de Metropolis-Hastings est défini par :

$$\begin{aligned} r(\boldsymbol{\theta}_k, \tilde{\boldsymbol{\theta}}_k) &= \min\left\{ \frac{\pi(\mathbf{S}|\tilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_j, j = 1, 2, 3, j \neq k, \boldsymbol{\rho})\pi(\tilde{\boldsymbol{\theta}}_k)}{\pi(\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho})\pi(\boldsymbol{\theta}_k)} \times \frac{q(\tilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k)}{q(\boldsymbol{\theta}_k, \tilde{\boldsymbol{\theta}}_k)}, 1 \right\} \\ &= \min\left\{ \frac{\phi_{3n}(\mathbf{S}|\mathbf{0}, \tilde{\mathbf{C}})\mathbb{1}_{\{[-c;\,c]\times[-c;\,c]\}}(\tilde{\boldsymbol{\theta}}_k)}{\phi_{3n}(\mathbf{S}|\mathbf{0}, \mathbf{C})\mathbb{1}_{\{[-c;\,c]\times[-c;\,c]\}}(\boldsymbol{\theta}_k)} \times \frac{\phi_2(\boldsymbol{\theta}_k|\tilde{\boldsymbol{\theta}}_k, U_k)}{\phi_2(\tilde{\boldsymbol{\theta}}_k|\boldsymbol{\theta}_k, U_k)}, 1 \right\} \end{aligned}$$

• **Mise à jour du vecteur de corrélation $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$** Les éléments ρ_k du vecteur de corrélation $\boldsymbol{\rho}$ sont mis à jour successivement. La loi *a priori* de ρ_k est une loi uniforme sur $[-1;\,1]$ si $k = 2, 3$ et une loi uniforme sur $[0, 1]$ si $k = 1$. La loi *a posteriori* s'écrit sous la forme :

$$\rho_k| \ldots \propto \pi(\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho})\pi(\boldsymbol{\rho}).$$

L'algorithme de Metropolis-Hastings est utilisé pour simuler une nouvelle valeur de ρ_k. La loi de proposition choisie est une loi normale de moyenne ρ_k et de variance $\sigma^2_{\rho_k}$ tronquée sur $[-1;\,1]$ si $k = 2, 3$ ou sur $[0;\,1]$ si $k = 1$. Le ratio de Metropolis-Hastings s'écrit alors :

$$r(\rho_k, \tilde{\rho}_k) = \min\left\{ \frac{\pi(\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \tilde{\rho}_k, \rho_j, j = 1, 2, 3,\, j \neq k)\pi(\tilde{\rho}_k)}{\pi(\mathbf{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho})\pi(\rho_k)} \times \frac{q(\tilde{\rho}_k, \rho_k)}{q(\rho_k, \tilde{\rho}_k)}, 1 \right\}.$$

Si le modèle ne comprend que deux variables, le vecteur $\boldsymbol{\rho}$ est restreint au produit $\rho_1\rho_2$. La loi *a priori* de $\boldsymbol{\rho}$ est une loi uniforme sur $[-1; 1]$. La mise à jour de $\boldsymbol{\rho}$ s'effectue à l'aide de l'algorithme de Metropolis-Hastings en prenant pour loi de proposition une loi normale tronquée sur $[-1; 1]$.

Cet algorithme a été implémenté en C interfacé avec R.

2.4 Prédictions

2.4.1 Comment les prédictions sont-elles obtenues ?

Le but de notre étude est de prédire le champ aléatoire $Y(.)$ sur tout le domaine d'étude $\mathcal{D}$, c'est-à-dire en un nombre n_0 de sites non échantillonnés $\mathbf{u}_1, \ldots, \mathbf{u}_{n_0}$ couvrant cette zone. Les méthodes de krigeage (paragraphe 1.2.1) habituellement utilisées dans le cadre de la prédiction ne s'appliquent pas ici, étant donné que toutes les variables ne sont pas gaussiennes. L'implémentation bayésienne du modèle va permettre d'effectuer des prédictions aux points non échantillonnés. On note $\tilde{\mathbf{S}}_k = (S_k(\mathbf{u}_1), \ldots, S_k(\mathbf{u}_{n_0})'$ le vecteur de la variable latente S_k aux points non échantillonnés et $\tilde{\mathbf{S}}$ le vecteur $(\tilde{\mathbf{S}}_1', \tilde{\mathbf{S}}_3', \tilde{\mathbf{S}}_4')'$. L'algorithme de « Metropolis-Hastings within Gibbs » est lancé pour estimer les paramètres du modèle. Lorsque la chaîne a atteint sa distribution stationnaire, une étape supplémentaire est ajoutée à l'algorithme MCMC pour générer $\tilde{\mathbf{S}}$. A l'itération t, un vecteur $\tilde{\mathbf{S}}^t$ de dimension $3n_0$ est simulé suivant la distribution conditionnelle gaussienne $\tilde{\mathbf{S}}^{(t)}|\mathbf{Y}, \mu_1^{(t)}, \mu_2^{(t)}, \mu_3^{(t)}, \mathbf{S}^{(t)}, \mathbf{Z}_3^{(t)}, \nu_1^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \boldsymbol{\rho}^{(t)}$ où $x^{(t)}$ désigne la valeur courante du paramètre x à l'itération t. Connaissant $\mathbf{S}^{(t)}$, $\boldsymbol{\theta}_1^{(t)}$, $\boldsymbol{\theta}_2^{(t)}$, $\boldsymbol{\theta}_3^{(t)}$ et $\boldsymbol{\rho}^{(t)}$, le vecteur $\tilde{\mathbf{S}}$ est indépendant de $\mathbf{Y}$ et du reste des paramètres. Le vecteur $\tilde{\mathbf{S}}^{(t)}$ est donc obtenu en effectuant un tirage suivant la distribution conditionnelle suivante :

$$\tilde{\mathbf{S}}^{(t)}|\mathbf{S}^{(t)}, \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \boldsymbol{\rho}^{(t)} \sim \mathcal{N}_{3n_0}(\mathbf{C}_{12}{}'\mathbf{C}_{11}{}^{-1}\mathbf{S}^{(t)}, \mathbf{C}_{22} - \mathbf{C}_{12}{}'\mathbf{C}_{11}{}^{-1}\mathbf{C}_{12}),$$

où $\mathbf{C}_{11} = \mathrm{Var}[\mathbf{S}^{(t)}]$, $\mathbf{C}_{12} = \mathrm{Cov}[\mathbf{S}^{(t)}, \tilde{\mathbf{S}}^{(t)}]$ et $\mathbf{C}_{22} = \mathrm{Var}[\tilde{\mathbf{S}}^{(t)}]$. Les matrices $\mathbf{C}_{11}$, $\mathbf{C}_{12}$ et $\mathbf{C}_{22}$ sont calculées en utilisant les valeurs courantes des paramètres $\boldsymbol{\theta}_k^{(t)}, k = 1, 2, 3$ et $\boldsymbol{\rho}^{(t)}$ (Diggle et al., 1998; Kern, 2000; Christensen et Waagepetersen, 2002). En utilisant la valeur courante de $\mu_k^{(t)}, k = 1, 2, 3$, nous obtenons :

- une réalisation $\tilde{y}_1^{(t)}(\mathbf{u}_i)$ de la moyenne de $Y_1(\mathbf{u}_i)$: $\tilde{Y}_1^{(t)}(\mathbf{u}_i) = \mu_1^{(t)} + S_1^{(t)}(\mathbf{s}_0)$,
- une réalisation $\tilde{y}_2^{(t)}(\mathbf{u}_i)$ de la moyenne de $Y_2(\mathbf{u}_i)$: $\tilde{Y}_2^{(t)}(\mathbf{u}_i) = \exp(\mu_2^{(t)} + S_2^{(t)}(\mathbf{u}_i))$,
- une réalisation $\tilde{z}_3^{(t)}(\mathbf{u}_i)$ de la moyenne de $Z_3(\mathbf{u}_i)$: $\tilde{Z}_3^{(t)}(\mathbf{u}_i) = \mu_3^{(t)} + S_3^{(t)}(\mathbf{u}_i)$. Nous déterminons $\tilde{Y}_3^{(t)}(\mathbf{u}_i)$ en tronquant $\tilde{Z}_3^{(t)}(\mathbf{u}_i)$ suivant le vecteur de seuils courant $\boldsymbol{\alpha}_3^{(t)}$.

Lorsque la convergence de la chaîne est atteinte, cette étape supplémentaire n'est réalisée que toutes les j itérations, afin de réduire la corrélation entre les différentes réalisations. En laissant l'algorithme MCMC tourner suffisamment longtemps, nous obtenons autant de réalisations de la moyenne de $\tilde{Y}(\mathbf{u}_i)$ que nous désirons. La prédiction de la variable gaussienne au point $\mathbf{u}_i$, $\widehat{\tilde{Y}_1}(\mathbf{u}_i)$, est la moyenne des réalisations $\tilde{Y}_1^{(t)}(\mathbf{u}_i)$. La variable de Poisson Y_2 est prédite au point $\mathbf{u}_i$ en prenant la médiane des réalisations $\tilde{Y}_2^{(t)}(\mathbf{u}_i)$ (Christensen et

Waagepetersen, 2002). De même, la prédiction de la variable ordinale au point $\mathbf{u}_i$ est la médiane des réalisations $\tilde{Y}_3^{(t)}(\mathbf{u}_i)$.

2.4.2 Comment mesurer la qualité des prédictions ?

La qualité des prédictions est mesurée grâce à des prédictions réalisées sur un jeu de données de validation. Les critères de validation employés sont différents suivant la nature de la variable prédite. Soit n_V le nombre de données contenues dans le jeu de données de validation. Soit $\widehat{Y_1}(\mathbf{s}_i)$, $1 \leq i \leq n_V$, la valeur prédite pour la variable gaussienne au $i^{\text{ème}}$ site du jeu de données de validation. Soit $\widehat{\text{var}}(\widehat{Y_1}(\mathbf{s}_i))$ la variance de prédiction estimée au point $\mathbf{s}_i$. La variance $\widehat{\text{var}}(\widehat{Y_1}(\mathbf{s}_i))$ est la somme de la variance de toutes les réalisations $\tilde{y}_1(\mathbf{s}_i)$ et de l'erreur d'échantillonnage approchée par ${\widehat{\nu_1}}^2$. Pour la variable gaussienne, nous considérons les critères de validation suivants (Ver Hoef et al., 2004) :

- $\text{biais} = \dfrac{1}{n_V}\displaystyle\sum_{i=1}^{n_V}(\widehat{Y_1}(\mathbf{s}_i) - Y_1(\mathbf{s}_i))$,
- $\text{RMSPE} = \sqrt{\dfrac{\sum_{i=1}^{n_V}(\widehat{Y_1}(\mathbf{s}_i) - Y_1(\mathbf{s}_i))^2}{n_V}}$,
- $\text{RMEV} = \sqrt{\dfrac{\sum_{i=1}^{n_V}\widehat{\text{var}}(\widehat{Y_1}(\mathbf{s}_i))}{n_V}}$,
- $80\%\text{PI} = \dfrac{1}{n_V}\displaystyle\sum_{i=1}^{n_V}\mathbb{1}\{|\widehat{Y_1}(\mathbf{s}_i) - Y_1(\mathbf{s}_i)| < 1.28\sqrt{\widehat{\text{var}}(\widehat{Y_1}(\mathbf{s}_i))}\}$.

Si les variances de prédiction sont correctement estimées, alors la racine carrée de la moyenne des variances estimées (RMEV, root-mean estimated variance) doit être proche de la racine carrée de la moyenne de l'erreur quadratique de prédiction (RMSPE, root-mean-squared-prediction error). La couverture de l'intervalle de prédiction, noté 80%PI, doit être proche de 80 %. Pour la variable de Poisson, l'intervalle de prédiction $[\zeta_{2.5\%}; \zeta_{97.5\%}]$ où ζ_q est le q-quantile de la série composée de toutes les réalisations $\tilde{y_2}(\mathbf{s}_i)$ est déterminé pour chaque site $\mathbf{s}_i$, $1 \leq i \leq n_V$ (Christensen et Waagepetersen, 2002). Nous nous intéressons au biais des prédictions et à la distribution de la largeur de l'intervalle de prédiction. La largeur de l'intervalle de prédiction nous renseigne sur l'aplatissement de la distribution prédictive des variables et donc sur la variance de prédiction. Pour la variable ordinale, nous donnons le pourcentage de valeurs correctement prédites (%CP) dans le jeu de données de validation.

2.5 Simulations

La procédure d'estimation des paramètres est validée à partir de simulations. Nous travaillons d'abord avec des jeux de données simulés bivariés. Nous appliquons ensuite la méthode d'estimation à un jeu de données trivarié pour comparer la qualité des prédictions ainsi obtenues avec celle des prédictions obtenues à partir d'une procédure d'estimation univariée ou bivariée.

Le temps de calcul nécessaire à l'estimation des paramètres par l'algorithme MCMC peut être conséquent si le nombre de variables composant le champ aléatoire multivarié est important. Pour tester la validité de la procédure d'estimation, nous nous limitons donc à des jeux de données simulés comportant au maximum trois variables.

Le temps de chauffe de l'algorithme MCMC, et par conséquent le temps de calcul, dépendent des valeurs initialement choisies pour les paramètres. Nous avons d'abord testé la méthode d'estimation en choisissant aléatoirement les valeurs initiales des paramètres. Par la suite, pour réduire le temps de chauffe de la procédure d'estimation multivariée, l'initialisation des paramètres a été effectuée comme suit. Pour chaque variable étudiée, nous avons effectué une procédure d'estimation univariée en partant de valeurs initiales de paramètres choisies aléatoirement. Les estimations des paramètres ainsi obtenues ont été prises comme valeurs initiales pour les procédures d'estimation multivariées.

2.5.1 Simulation d'un jeu de données

Plusieurs jeux de données composés de différents types de variables sont simulés en utilisant la construction moyenne mobile. Nous décrivons ici la procédure pour simuler un jeu de données trivarié composé d'une variable gaussienne Y_1, d'une variable de Poisson Y_2 et d'une variable ordinale Y_3 à L modalités. Une procédure similaire peut être utilisée pour simuler des jeux de données bivariés.

350 points sont tirés uniformément dans un carré de dimension $[-10;10] \times [-10;10]$. Nous déterminons ensuite la matrice de covariance $\mathbf{C}$ du vecteur $\mathbf{S}$. Ici, les fonctions moyennes mobiles sont choisies proportionnelles au noyau gaussien :

$$f_k(\mathbf{x}|\boldsymbol{\theta}_k) = \sigma_k \exp(-\|\mathbf{x}\|^2/\phi_k) \text{ avec } \boldsymbol{\theta}_k = (\sigma_k, \phi_k).$$

Elles ne dépendent que d'un nombre limité de paramètres et conduisent à une forme explicite de la covariance. La matrice de covariance correspondante a la forme analytique suivante :

$$\begin{aligned} \mathrm{Cov}[S_k(\mathbf{s}_i), S_k(\mathbf{s}_j)] &= \frac{\sigma_k^2 \phi_k \pi}{2} \exp\left(-\frac{||\mathbf{s}_j - \mathbf{s}_i||^2}{2\phi_k}\right), \\ \mathrm{Cov}[S_k(\mathbf{s}_i), S_m(\mathbf{s}_j)] &= \frac{\rho_k \rho_m \sigma_k \sigma_m \phi_k \phi_m \pi}{\phi_k + \phi_m} \exp\left(-\frac{||\mathbf{s}_j - \mathbf{s}_i||^2}{\phi_k + \phi_m}\right). \end{aligned}$$

Remarquons que la seconde équation redonne la première lorsque $k = m$, à condition de poser $\rho_k \rho_m = \rho_{km}$ et $\rho_{kk} = 1$. Nous simulons le vecteur $\mathbf{S}$, de longueur $3n$, selon la distribution normale multivariée de moyenne nulle et de matrice de covariance $\mathbf{C}$. Le vecteur de variables aléatoires gaussiennes $\mathbf{Y}_1$ est obtenu en effectuant un tirage suivant la loi normale multivariée $\mathcal{N}_n(\mu_1 \mathbf{1} + \mathbf{S}_1, \nu_1^2 \mathbf{I}_n)$. $\mathbf{I}_n$ désigne la matrice identité d'ordre n. Pour tout i, $i = 1, \ldots n$, la variable $Y_2(\mathbf{s}_i)$ est simulée suivant la distribution $\mathcal{P}(\exp(\mu_2 + S_2(\mathbf{s}_i)))$. En ce qui concerne la variable ordinale, nous simulons d'abord le vecteur de variables latentes $\mathbf{Z}_3$ de loi $\mathcal{N}_n(\mu_3 \mathbf{1} + \mathbf{S}_3, \mathbf{I}_n)$. Pour $l = 1, \ldots, L$, la variable $Y_3(\mathbf{s}_i)$ prend la valeur l si $Z_3(\mathbf{s}_i)$ est compris entre le $(l-1)^{\text{ème}}$ et le $l^{\text{ème}}$ L-quantile de $\mathbf{Z}_3$. Pour les différentes

simulations réalisées, le nombre de modalités L est pris égal à 3. Le jeu de données est constitué de 250 données choisies aléatoirement parmi les 350 initialement simulées. Les 100 restantes sont utilisées comme jeu de données de validation.

2.5.2 Résultats pour des jeux de données bivariés

La procédure d'estimation a été appliquée sur plusieurs jeux de données simulés bivariés (jeux de données composés de deux variables gaussiennes, d'une variable gaussienne et d'une variable de Poisson, d'une variable gaussienne et d'une variable ordinale, de deux variables de Poisson, d'une variable de Poisson et d'une variable ordinale, de deux variables ordinales). Tous les jeux de données ont été obtenus par une procédure similaire à celle décrite ci-dessus. Dans la suite, Y_k désigne la $k^{\text{ème}}$ variable du jeu de données quelle que soit sa nature.

Plusieurs chaînes dont les valeurs initiales ont été choisies aléatoirement ont été lancées sur chaque jeu de données bivarié pour tester la convergence de l'algorithme MCMC. La convergence a été atteinte pour tous les jeux de données. Le temps de chauffe varie suivant la nature des variables considérées. La convergence de l'algorithme a été contrôlée graphiquement et à l'aide du diagnostic dit de Gelman et Rubin (1992). La figure 2.2 illustre la convergence des chaînes pour le jeu de données composé de deux variables ordinales. Parmi tous les jeux de données bivariés simulés, ce dernier est le plus difficile à traiter car, à chaque variable ordinale, correspondent deux niveaux de variables latentes Z et S.

La tableau 2.1 présente exclusivement les estimations des cinq paramètres communs à tous les jeux de données (σ_1, ϕ_1, σ_2, ϕ_2, ρ_{12}). Les résultats d'estimation pour l'ensemble des paramètres sont discutés ci-dessous.

Les estimations des paramètres sont cohérentes avec les valeurs utilisées pour les simulations. Les paramètres relatifs aux variables gaussiennes et de Poisson sont estimés précisément ; cela reste vrai même si le nombre d'observations est plus faible. La moyenne et les seuils associés aux variables ordinales sont difficiles à estimer et leurs variances sont élevées. Pour le jeu de données gaussien-ordinal, l'estimation du seuil $\alpha_{2,2}$ est de 31,47 avec un écart-type de 4,6 alors que la vraie valeur est de 35,95. Les seuils relatifs aux jeux de données Poisson-ordinal et ordinal-ordinal sont estimés plus précisément. Les écarts absolus entre les estimations et les vraies valeurs sont compris entre 0,11 et 0,91 et les écarts-types des estimations varient entre 0,186 et 0,711. La corrélation ρ_{12} entre les variables est correctement estimée pour tous les couples de variables excepté pour le jeu de données gaussien-Poisson où elle est surestimée. La vitesse de convergence est élevée pour les variables gaussiennes et de Poisson. L'estimation des paramètres pour les variables ordinales requiert plus d'itérations pour atteindre la distribution stationnaire. Le temps de chauffe est plus long à cause de la présence des deux vecteurs latents $\mathbf{Z}$ et $\mathbf{S}$. Les critères de validation permettant de mesurer la qualité des prédictions pour chaque jeu de données sont présentés dans le tableau 2.2.

Le biais est faible ($< 0,43$) pour les prédictions des variables gaussiennes. La RMSPE et la RMEV sont proches (différence maximale $\approx 0{,}4$) ; la variance de prédiction est donc estimée précisément. Les valeurs de la couverture de l'intervalle de prédiction comprises

TAB. 2.1 – Estimation des paramètres σ_1, ϕ_1, σ_2, ϕ_2, ρ_{12} associés à la structure de dépendance à partir des jeux de données bivariés simulés. Pour chaque jeu de données, la première ligne comporte les vraies valeurs des paramètres et la seconde les valeurs estimées. Les estimations sont les moyennes *a posteriori* des paramètres. Les écarts-types sont donnés entre parenthèses. La nature des variables composant le jeu de données est indiquée sur la gauche.

Jeu de données	σ_1	ϕ_1	σ_2	ϕ_2	ρ_{12}
Y_1 gaussien	10	2	20	4	0,5
Y_2 gaussien	10,63 (0,61)	2,02 (0,08)	19,43 (1,30)	3,73 (0,18)	0,53 (0,06)
Y_1 gaussien	10	2	1	0,2	0,5
Y_2 Poisson	9,47 (0,54)	1,93 (0,08)	1,36 (0,43)	0,11 (0,05)	0,72 (0,16)
Y_1 gaussien	10	2	20	4	0,5
Y_2 ordinal	9,06 (0,48)	1,83 (0,08)	19,72 (2,93)	2,99 (0,42)	0,42 (0,08)
Y_1 Poisson	1	0,2	0,5	2	0,3
Y_2 Poisson	0,999 (0,082)	0,27 (0,05)	0,45 (0,04)	1,51 (0,18)	0,12 (0,14)
Y_1 Poisson	0,5	2	2	4	0,5
Y_2 ordinal	0,54 (0,03)	2 (0,17)	2,22 (0,56)	3,13 (0,62)	0,48 (0,10)
Y_1 ordinal	1	3	2	3	0,5
Y_2 ordinal	0,85 (0,18)	2,85 (0,93)	1,31 (0,24)	3,25 (0,69)	0,44 (0,12)

TAB. 2.2 – Critères de validation permettant de mesurer la qualité des prédictions pour chacun des jeux de données simulés. Le biais, la RMSPE, la RMEV et l'intervalle de couverture 80%PI sont donnés pour les variables gaussiennes. Pour les variables de Poisson, nous donnons le biais et des statistiques (minimum, $q_{0,25}$, médiane, moyenne, $q_{0,75}$, maximum) résumant la distribution de la largeur de l'intervalle de prédiction. Le pourcentage de valeurs correctement prédites, noté %CP, est indiqué pour chaque variable ordinale.

Jeu de données	**Variable 1**	**Variable 2**
gaussien-gaussien	biais 0,15 RMSPE 3,10 RMEV 3,10 80%PI 0,78	biais 0,15 RMSPE 3,71 RMEV 4,13 80%PI 0,88
gaussien-Poisson	biais 0,43 RMSPE 2,18 RMEV 2,59 80%PI 0,82	biais $-0,58$ (5,53 ; 7,53 ; 8 ; 8,44 ; 9 ; 13,52)
gaussien-ordinal	biais 0,28 RMSPE 2,51 RMEV 2,24 80%PI 0,86	%CP 91
Poisson-Poisson	biais $-0,52$ (4 ; 8 ; 9 ; 9,11 ; 10 ; 15)	biais $-0,49$ (5 ; 9 ; 13 ; 14,19 ; 17 ; 33,05)
Poisson-ordinal	biais 0,04 (3 ; 9 ; 12,5 ; 16,57 ; 21,64 ; 52)	%CP 77
ordinal-ordinal	%CP 54	%CP 74

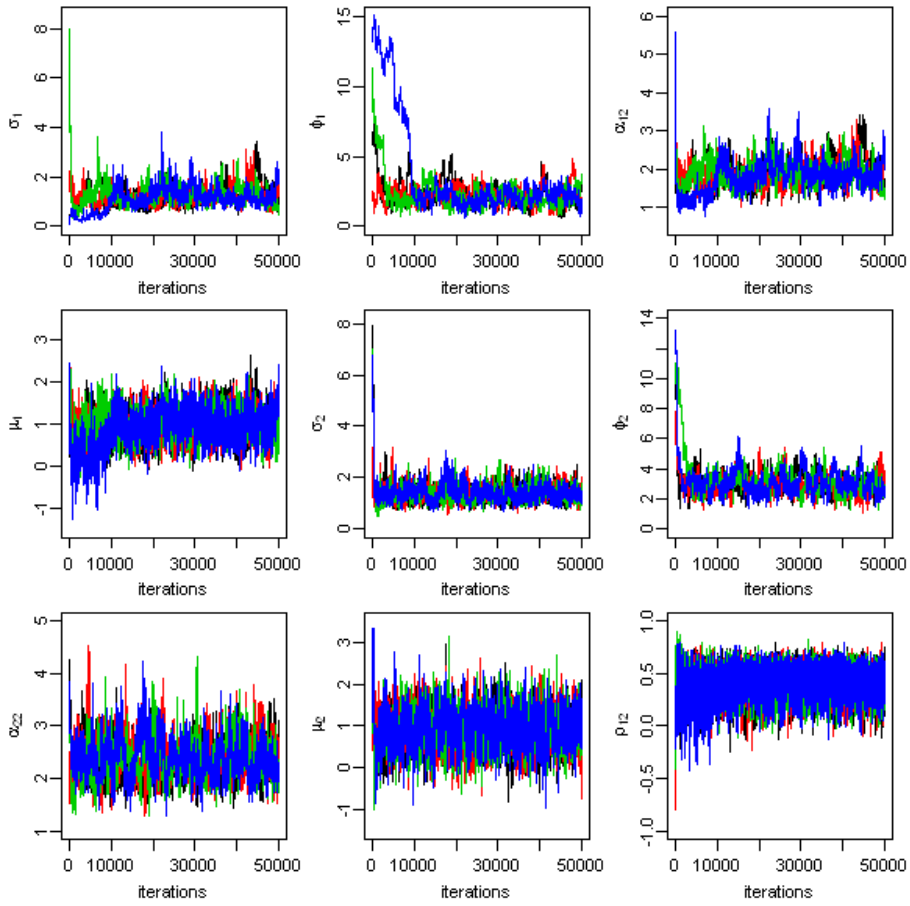


FIG. 2.2 – Échantillonnage suivant les lois conditionnelles complètes des paramètres obtenu à partir du jeu de données simulé composé de deux variables ordinales. La procédure d'estimation a été lancée quatre fois ; les chaînes correspondantes sont tracées en différentes couleurs.

entre 77 % et 85 % confirment ce résultat. La plupart des prédictions relatives aux variables de Poisson présentent un biais de l'ordre de $-0,5$, excepté pour le jeu de données Poisson-ordinal où le biais est plus faible (0,04). La largeur moyenne des intervalles de prédiction varie d'un jeu de données à l'autre. Les intervalles de prédiction les plus petits sont obtenus pour le jeu de données gaussien-Poisson (75 % des intervalles ont une largeur inférieure à 9) ; la variable gaussienne fournit suffisamment d'information pour améliorer la variance des prédictions de la variable de Poisson. Le pourcentage des valeurs correctement prédites pour les variables ordinales varie entre 54 % et 91 %. Dans le jeu de données gaussien-ordinal, l'information apportée par la variable gaussienne peut expliquer le pourcentage élevé de valeurs correctement prédites. Inversement, dans le jeu de données ordinal-ordinal, nous supposons que la seconde variable ordinale n'apporte pas suffisamment d'information pour

améliorer la qualité des prédictions de la première.

Les estimations et les critères de validation relatifs aux variables gaussiennes obtenus par une procédure d'estimation univariée et ceux obtenus par une procédure d'estimation bivariée sont quasiment similaires. En revanche, la procédure d'estimation bivariée améliore les estimations relatives aux variables ordinales, et par conséquent la qualité des prédictions.

2.5.3 Résultats pour un jeu de données trivarié

Le but de notre étude est de prédire le plus précisément possible les variables étudiées dans des sites non échantillonnés. Nous faisons l'hypothèse qu'inclure un grand nombre de variables à l'étude améliore les prédictions grâce à la prise en compte des corrélations existant entre les variables. En contrepartie, plus le modèle comprend de variables, plus le temps de calcul nécessaire à l'estimation des paramètres est conséquent. Nous souhaitons mesurer l'effet d'une procédure d'estimation multivariée sur la qualité des prédictions. La procédure d'estimation est appliquée à un jeu de données simulé composé d'une variable gaussienne, d'une variable de Poisson et d'une variable ordinale. Les prédictions obtenues avec la procédure d'estimation trivariée sont comparées à celles obtenues avec des procédures univariées ou bivariées. Les critères de validation sont données dans le tableau 2.3.

Le biais des prédictions associées à la variable gaussienne est plus faible lorsque la procédure d'estimation est multivariée. La RMSPE et la RMEV sont du même ordre de grandeur que l'on travaille en univarié ou en multivarié. La probabilité de couverture est d'environ 80 %. En ce qui concerne la variable de Poisson, la largeur moyenne de l'intervalle de prédiction diminue légèrement lorsque le nombre de variables augmente, alors que la valeur absolue du biais a, quant à elle, tendance à augmenter. Considérer un jeu de données multivarié conduit à un pourcentage de valeurs correctement prédites, pour la variable ordinale, plus élevé que lorsque cette dernière est traitée en univarié. Ainsi la prise en compte de la dépendance existant entre la variable ordinale et les autres variables permet-elle d'améliorer la qualité des prédictions relatives à la variable ordinale.

La prise en compte de la corrélation existant entre les variables grâce à une procédure d'estimation multivariée permet d'améliorer la qualité des prédictions, notamment celles de la variable ordinale. La comparaison des prédictions obtenues avec la procédure d'estimation bivariée et la procédure d'estimation trivariée montre qu'inclure un nombre plus important de variables dans le modèle ne garantit en rien d'augmenter la qualité des prédictions et augmente les temps de calcul.

2.6 Discussion

L'objectif de notre étude était de proposer un modèle spatial multivarié permettant de prédire des variables de différente nature. Le modèle spatial multivarié hiérarchique décrit au paragraphe 1.3 répond en partie à ce problème. Il permet, en effet, de traiter simultanément des variables gaussiennes, des variables de Poisson et des variables ordinales, grâce à une approche basée sur les modèles linéaires généralisés spatiaux. Les simulations réalisées ont confirmé la capacité du modèle à prédire différents types de variables et ont

TAB. 2.3 – Critères de validation permettant de mesurer la qualité des prédictions à partir d'un jeu de données simulé en utilisant successivement des procédures d'estimation univariées, bivariées et trivariées. Le biais, la RMSPE, la RMEV et l'intervalle de couverture 80%PI sont donnés pour les variables gaussiennes. Pour les variables de Poisson, nous donnons le biais et des statistiques (minimum, $q_{0,25}$, médiane, moyenne, $q_{0,75}$, maximum) résumant la distribution de la largeur de l'intervalle de prédiction. Le pourcentage de valeurs correctement prédites, noté %CP, est indiqué pour chaque variable ordinale.

Jeu de données	**Variable gaussienne**		**Variable de Poisson**	**Variable ordinale**	
Univarié	biais RMSPE RMEV 80%PI	0,47 4,43 4,95 0,81	biais $-0,34$ (5,03 ; 8 ; 9 ; 8,95 ; 10 ; 13)	%CP	76,0
Bivarié gaussien-Poisson	biais RMSPE RMEV 80%PI	0,27 4,59 4,96 0,80	biais $-0,37$ (5 ; 8 ; 9 ; 8,83 ; 10 ; 14)		
Bivarié gaussien-ordinal	biais RMSPE RMEV 80%PI	0,38 4,44 4,95 0,83		%CP	79,5
Bivarié Poisson-ordinal			biais $-0,49$ (5 ; 7,03 ; 9 ; 8,76 ; 10 ; 14)	%CP	80,5
Trivarié	biais RMSPE RMEV 80%PI	0,35 4,57 5,11 0,80	biais $-0,40$ (5 ; 7,76 ; 9 ; 8,66 ; 10 ; 13,02)	%CP	80,0

montré qu'une procédure d'estimation multivariée conduit à des prédictions de meilleure qualité qu'une procédure d'estimation univariée.

Dans le modèle, la dépendance entre les variables se traduit au travers de la dépendance de leurs composantes spatiales S_k. La structure de dépendance est modélisée par une matrice obtenue grâce à la construction moyenne mobile. Remarquons qu'il est possible d'utiliser un modèle de covariance classique, mais, en l'absence d'information sur les variables latentes S_k, l'approche moyenne mobile offre l'avantage d'être plus flexible. Comme nous l'avons vu au paragraphe 1.2.2, le choix des fonctions moyennes mobiles est délicat et reste un problème ouvert. Il pourrait être intéressant de tester la robustesse des prédictions au choix de la forme des fonctions moyennes mobiles grâce à des simulations. Une extension de la construction moyenne mobile pourrait être envisagée afin de prendre en compte la dépendance spatiale à différentes échelles comme dans le modèle linéaire de corégionalisation.

L'hypothèse simplificatrice qui consiste à prendre la matrice $\mathbf{R}$ égale à l'identité lorsque le modèle comprend plusieurs variables ordinales peut être discutée. Cette hypothèse est assez restrictive, mais permet de simplifier la procédure d'estimation : chaque variable latente $Z(\mathbf{s})$ peut être mise à jour séparément. Cette hypothèse peut être levée, c'est-à-dire que les variables latentes $Z_k(\mathbf{s})$ associées aux variables ordinales ne sont pas conditionnellement indépendantes connaissant les variables $S_k(\mathbf{s})$, comme dans la définition 4. Le modèle doit alors être complété en donnant une loi *a priori* sur les paramètres intervenant dans la définition de la matrice $\mathbf{R}$. Dans ce cas, la procédure d'inférence est plus complexe et le temps de calcul plus important.

La procédure d'estimation des paramètres est gourmande en ressources informatiques et le temps de calcul nécessaire à l'estimation peut s'avérer long. Ce dernier augmente avec le nombre de variables étudiées et le nombre de sites échantillonnés, car les matrices manipulées sont alors de grande dimension. Plusieurs alternatives peuvent être envisagées pour remédier à ce problème. Il pourrait être fait appel à des méthodes basées sur la vraisemblance composite (Varin, 2008). Une autre alternative consisterait à simplifier la procédure d'estimation en suivant l'approche proposée par Joe (1997). Cette approche consiste, dans un premier temps, à effectuer autant de procédures d'estimation univariées qu'il y a de variables dans le modèle, afin d'estimer les paramètres relatifs à chacune des variables. Dans un second temps, une procédure d'estimation multivariée est lancée. Les paramètres associés à chacune des variables sont considérés comme connus, seul le vecteur de corrélations $\boldsymbol{\rho}$ est estimé, ce qui permet de réduire la durée des calculs. Enfin, une autre solution consisterait à utiliser une inférence bayésienne approchée (Eidsvik et al., 2009; Rue et al., 2009) au lieu des simulations MCMC. L'amélioration des performances de la procédure d'estimation pourrait éventuellement permettre d'inclure des variables nominales dans le modèle.

La procédure d'estimation bayésienne des paramètres du modèle sera appliquée à un jeu de données réelles au chapitre 4.

Chapitre 3

Prédiction de la répartition spatiale et du génotype des juvéniles

Sommaire

3.1 Problématique

Comme nous l'avons vu dans l'introduction, la régénération est un phénomène complexe. Lors des différentes étapes de la régénération, de la floraison d'un arbre adulte à l'installation d'un nouvel individu dans le peuplement, l'environnement intervient à plusieurs reprises. Il agit notamment comme un filtre sur la survie des juvéniles[1]. L'idéal pour mieux comprendre le phénomène de régénération et l'effet des conditions environnementales sur l'établissement des juvéniles serait de pouvoir suivre l'évolution du peuplement à grande échelle. L'effort d'échantillonnage nécessaire est beaucoup trop important et un tel suivi n'est pas envisageable. En général, les données dont nous disposons sont plutôt du type suivant. Des relevés ont été effectués pour caractériser l'environnement sur une partie ou sur l'ensemble de la zone d'étude (Figure 3.1). Les arbres adultes peuvent être échantillonnés de manière exhaustive ou non. Les juvéniles ne sont, en général, échantillonnés que partiellement ou le sont de manière exhaustive sur de petites zones.

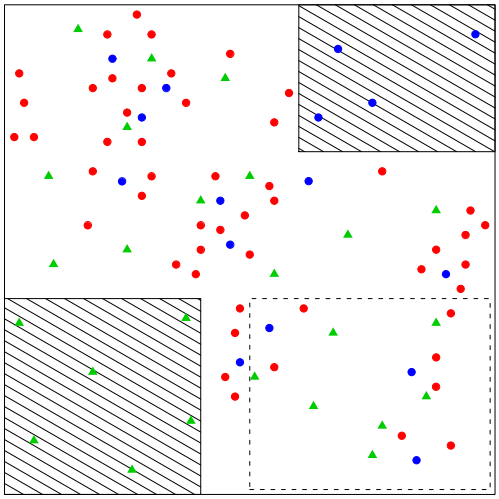


FIG. 3.1 – Schéma représentant les données disponibles avant la prédiction de l'environnement. Les triangles verts représentent les sites où l'environnement a été échantillonné. Les localisations des adultes sont représentées par des points bleus et celles des juvéniles par des points rouges. Les zones hachurées sont des zones où les juvéniles n'ont pas été échantillonnés. Les zones délimitées par des pointillés sont des zones où les juvéniles ont été partiellement échantillonnés.

Le modèle hiérarchique spatial multivarié proposé au chapitre 1 permet de prédire

[1]On considère comme juvénile tout arbre ayant un diamètre inférieur à un diamètre seuil fixé. Ce diamètre seuil dépend de l'espèce étudiée. Par exemple, pour *Dicorynia guianensis*, nous considérerons comme juvéniles tous les arbres ayant un diamètre inférieur à 5 cm.

l'environnement en n'importe quel point du domaine d'étude à partir des observations réalisées. Connaissant l'environnement sur toute la zone d'étude, le problème est alors de « prédire la régénération » sur les zones non échantillonnées (zones hachurées) ou sur les zones partiellement échantillonnées (zones délimitées par des pointillés). « Prédire la régénération » ne signifie pas déterminer la position de chaque juvénile avec précision. Il s'agit ici de caractériser la répartition spatiale et la diversité génétique des juvéniles. Nous aimerions, par exemple, déterminer la probabilité qu'il y ait des juvéniles dans les zones non échantillonnées. Si la probabilité d'observer des juvéniles est non nulle, nous aimerions pouvoir déterminer en quelle quantité ils sont présents et comment ils sont répartis. Nous souhaiterions également connaître leurs génotypes. Les réponses à ces questions passent par l'amélioration de la compréhension des phénomènes de dispersion (dissémination du pollen, dispersion des graines) (Gerber et al., 2004) à l'origine de l'organisation spatiale et de la diversité génétique des juvéniles, et par l'étude des effets de l'environnement sur la survie des juvéniles au cours du processus de régénération.

L'objectif principal, d'un point de vue mathématique est de proposer un modèle de régénération spatialement explicite qui tienne compte des variables environnementales et qui permette d'inclure l'information génétique recueillie sur les adultes et les juvéniles dans la description des phénomènes de dispersion. Ce modèle doit permettre de « prédire la régénération » à partir d'un échantillonnage raisonnable des adultes reproducteurs et des juvéniles. L'étape qui consiste à extrapoler l'environnement sur toute la zone d'étude avant de modéliser la régénération crée une difficulté supplémentaire d'un point de vue statistique. Les variables environnementales prédites sont entachées d'erreur, alors qu'elles sont considérées comme connues dans les modèles de régénération existant. La variabilité introduite dans le modèle de régénération par la prédiction de l'environnement doit donc être prise en compte et son effet sur les éléments caractérisant la régénération doit être quantifié.

3.2 Modèles de régénération incluant de l'information génétique

Plusieurs modèles mathématiques ont été développés pour prédire la régénération. Certains modèles sont, à proprement parler, des modèles de recrutement. Ils ne permettent de déterminer que le nombre d'arbres qui seront recrutés sur une zone donnée pendant un temps donné (Vanclay, 1992; Ribbens et al., 1994; Lexerød, 2005). D'autres modèles se rapprochent plus d'un modèle de régénération dans la mesure où ils font apparaître différentes étapes du processus de régénération et permettent, non seulement d'obtenir de l'information sur le nombre d'arbres recrutés, mais aussi sur les mécanismes de dispersion ou la survie des juvéniles (Sagnard et al., 2007; Eerikäinen et al., 2007). D'autres encore s'intéressent plus spécialement à la distribution spatiale des juvéniles en prenant en compte la sur-représentation de zéros dans les données (Rathbun et Fei, 2006; Rathbun et Black, 2006; Flores et al., 2009). Ces modèles offrent tous la possibilité de prendre en compte des variables environnementales. Plus récemment, le développement rapide des marqueurs mo-

léculaires a permis le recueil d'une quantité plus importante d'information génétique. Cette information génétique a été mise à profit pour améliorer la compréhension des mécanismes de dispersion (Austerlitz et al., 2004; Robledo-Arnuncio et Garcia, 2007; Jones et Muller-Landau, 2008). De nouveaux modèles de régénération incluant de l'information génétique ont vu le jour. Ils permettent d'établir un lien entre l'organisation spatiale des individus et la structuration génétique de la population. Ces modèles s'appuient essentiellement sur la description des phénomènes de dispersion. Parmi ces modèles, on trouve notamment les modèles basés sur la notion de voisinage (Adams et Birkes, 1989, 1991; Burczyk et al., 2002, 2006) et ceux utilisant les processus ponctuels (Shimatani, 2004; Shimatani et al., 2006). Ces deux types de modèles permettent d'estimer le succès reproducteur des arbres matures, le flux de graines venu de l'extérieur de la zone d'étude, mais seul le modèle proposé par Shimatani (2004) permet de déterminer la densité locale des juvéniles et d'envisager une prédiction de la régénération. C'est donc à ce type de modèles, appelé en anglais « spatial molecular ecological model », que nous nous intéressons ici.

Nous rappelons au préalable quelques notions sur les processus ponctuels nécessaires à la bonne compréhension de ces modèles.

3.2.1 Quelques notions sur les processus ponctuels

Les définitions concernant les processus ponctuels données ci-dessous sont issues de l'ouvrage de Møller et Waagepetersen (2004) sauf mention du contraire.

Définition d'un processus ponctuel

Un processus ponctuel est un processus aléatoire dont les réalisations sont des semis de points. Avant de donner une définition plus précise, nous définissons ce qu'est une configuration de points localement finie.

Définition 5 Configuration de points localement finie
Soit (E, d) un espace métrique complet et séparable muni de la distance d.
Soit $\mathbf{x} \subseteq E$ une configuration de points de E et $n(\mathbf{x})$ son cardinal.
On note $\mathbf{x}_B = \mathbf{x} \cap B$, $B \subseteq E$, la restriction de la configuration $\mathbf{x}$ à l'ensemble B.
*Une configuration $\mathbf{x} \subseteq E$ est dite **localement finie**, si $n(\mathbf{x}_B) < \infty$ pour tout ensemble borné $B \subseteq E$.*
L'ensemble de toutes les configurations localement finies de E est noté N_{lf} :

$$N_{lf} = \{\mathbf{x} \subseteq E : n(\mathbf{x}_B) < \infty \text{ pour tout ensemble borné } B \subseteq E\}.$$

Nous pouvons maintenant définir un processus ponctuel.

Définition 6 Processus ponctuel
Soit (E, d) un espace métrique complet et séparable muni de la distance d et $\mathfrak{B}$ la tribu borélienne associée. On note $\mathfrak{B}_0$ l'ensemble des boréliens bornés de E.

Soit $\mathcal{N}_{lf}$ la σ-algèbre associée à l'espace des configurations localement finies de E :

$$\mathcal{N}_{lf} = \mathcal{T}(\{\mathbf{x} \in N_{lf} : n(\mathbf{x}_{\mathcal{B}}) = m\} : \mathcal{B} \in \mathfrak{B}_0, m \in \mathbb{N}).$$

La notation $\mathcal{T}(A)$ désigne la tribu engendrée par l'ensemble A. Un ***processus ponctuel*** *X défini sur E est une application mesurable d'un espace probabilisé $(\Omega, \mathcal{F}, P)$ à valeurs dans $(N_{lf}, \mathcal{N}_{lf})$.*
La distribution P_X de X est définie par :

$$P_X(U) = P(\{\omega \in \Omega : X(\omega) \in U\}), \forall U \in \mathcal{N}_{lf}.$$

Soit $X_{\mathcal{B}}$ la restriction du processus ponctuel X à l'ensemble $\mathcal{B}$. La mesurabilité de X est équivalente au fait que la fonction de comptage $N(\mathcal{B}) = n(X_{\mathcal{B}})$ soit une variable aléatoire pour tout $\mathcal{B} \in \mathfrak{B}_0$. On retrouve alors la définition du processus ponctuel donnée par Van Lieshout (2000). Un processus ponctuel peut également être défini en terme de mesure aléatoire de comptage (Daley et Vere-Jones, 1988).

Définition 7 Processus ponctuel marqué
Soit Y un processus ponctuel sur $E \subseteq \mathbb{R}^d$. Soit M un espace. Si une « marque » aléatoire $m_\xi \in M$ est associée à chaque point ξ de Y, alors

$$X = \{(\xi, m_\xi) : \xi \in Y\}$$

est un ***processus ponctuel marqué*** *défini sur E et dont l'espace des marques est M.*

Un exemple simple de processus ponctuel marqué est le processus ponctuel multitype où $M = \{1, \ldots, k\}$ et où les marques correspondent aux k différents types de points (par exemple, différents types de cellules, différentes espèces de plantes, etc). Cela est équivalent à considérer un processus ponctuel multivarié $(X_1, \ldots, X_k)$ où chaque processus ponctuel X_k correspond à un type de points particulier.

Mesures des moments d'un processus ponctuel

De la même manière que l'on définit les moments d'une variable aléatoire, on définit les mesures des moments d'un processus ponctuel X défini sur $E = \mathbb{R}^d$. Cela revient à définir les propriétés du premier et du second ordre des variables aléatoires de comptage $N(\mathcal{B})$ pour tout borélien $\mathcal{B} \subseteq E$.

Définition 8 Mesure d'intensité
La ***mesure d'intensité*** *Λ sur $\mathbb{R}^d$ est donnée par*

$$\Lambda(\mathcal{B}) = \mathbb{E}[N(\mathcal{B})], \mathcal{B} \subseteq \mathbb{R}^d.$$

Définition 9 Fonction d'intensité
Si la mesure d'intensité Λ peut s'écrire

$$\Lambda(\mathcal{B}) = \int_{\mathcal{B}} \lambda(\xi) d\xi, \mathcal{B} \subseteq \mathbb{R}^d$$

où λ est une fonction positive, alors λ est appelée ***fonction d'intensité****. Si la fonction λ est constante, le processus ponctuel X est dit* ***homogène*** *ou* ***stationnaire du premier ordre*** *d'intensité λ, sinon le processus ponctuel X est dit* ***hétérogène****.*

La quantité $\lambda(\xi)\mathrm{d}\xi$ peut être interprétée comme la probabilité d'avoir un point dans une boule infinitésimale de centre ξ et de volume $\mathrm{d}\xi$.

Définition 10 Mesure du moment factoriel du second ordre
La ***mesure du moment factoriel du second ordre*** $\alpha^{(2)}$ *sur $\mathbb{R}^d \times \mathbb{R}^d$ est définie par :*

$$\alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = \mathbb{E}\left[\sum_{\xi,\eta \in X, \xi \neq \eta} \mathbb{1}[(\xi,\eta) \in \mathcal{B}_1 \times \mathcal{B}_2]\right], \mathcal{B}_1 \times \mathcal{B}_2 \subseteq \mathbb{R}^d \times \mathbb{R}^d.$$

Les moments d'ordre 2 des variables aléatoires $N(\mathcal{B}), \mathcal{B} \subseteq \mathbb{R}^d$ peuvent être exprimés en fonction de Λ et de $\alpha^{(2)}$:

$$\mathbb{E}[N(\mathcal{B}_1)N(\mathcal{B}_2)] = \alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) + \Lambda(\mathcal{B}_1 \cap \mathcal{B}_2),\ \mathcal{B}_1, \mathcal{B}_2 \subseteq \mathbb{R}^d.$$

Définition 11 Densité produit du second ordre
Si la mesure du moment factoriel du second ordre peut s'écrire sous la forme

$$\alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = \int\int \mathbb{1}[(\xi,\eta) \in \mathcal{B}_1 \times \mathcal{B}_2]\lambda^{(2)}(\xi,\eta)\,d\xi\,d\eta,\ \mathcal{B}_1 \times \mathcal{B}_2 \subseteq \mathbb{R}^d \times \mathbb{R}^d$$

où $\lambda^{(2)}$ est une fonction positive, alors $\lambda^{(2)}$ est appelée ***densité produit du second ordre****.*

La quantité $\lambda^{(2)}(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta$ peut être interprétée comme la probabilité d'avoir simultanément un point de X dans la boule infinitésimale de centre ξ et de volume $\mathrm{d}\xi$ et un autre dans la boule infinitésimale de centre η et de volume $\mathrm{d}\eta$.

Quelques processus ponctuels particuliers

• **Processus de Poisson** Un processus de Poisson est un processus ponctuel pour lequel la disposition des points est complètement aléatoire à chaque réalisation. Son rôle est essentiel puisqu'il sert de référence (hypothèse nulle) pour tester la structure spatiale des autres semis de points.

Définition 12 Processus de Poisson (Gaetan et Guyon, 2008)
Soit Λ une mesure positive sur $(E, \mathcal{B})$ de densité λ, Λ étant finie sur les boréliens bornés. Un ***processus ponctuel de Poisson*** *de mesure d'intensité $\Lambda(.) > 0$ et d'intensité $\lambda(.)$ est caractérisé par :*

1. *Pour tout $B \in \mathcal{B}_0$ de mesure $0 < \Lambda(B) < \infty$, $N(B)$ suit un loi de Poisson de paramètre $\Lambda(B)$.*
2. *Conditionnellement à $N(E)$, les points de $\mathbf{x}_B$ sont indépendamment et identiquement distribués et leur densité est proportionnelle à $\lambda(\xi)$, $\xi \in B$:*

$$\mathbb{P}(N(B) = n) = e^{-\Lambda(B)}\frac{(\Lambda(B))^n}{n!}$$

et

$$g_n(\{x_1, \ldots, x_n\}) \propto \prod_{i=1}^{n} \lambda(x_i).$$

Si E est borné, la densité du processus de Poisson X de fonction d'intensité λ par rapport au processus de Poisson standard[2] est donnée par :

$$\exp\left(|E| - \int_E \lambda(u)\mathrm{d}u\right) \prod_{u \in X} \lambda(u).$$

Soit X un processus de Poisson hétérogène de fonction d'intensité $\lambda(.;\boldsymbol{\eta})$ dépendant d'un vecteur de paramètre $\boldsymbol{\eta}$. La log-vraisemblance de $\boldsymbol{\eta}$ est définie par :

$$\ell(\boldsymbol{\eta}; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln \lambda(x_i; \boldsymbol{\eta}) - \int_W \lambda(u; \boldsymbol{\eta})\mathrm{d}_u. \tag{3.1}$$

• **Processus de Cox** Les processus de Cox (ou « doubly stochastic Poisson processes ») sont utilisés pour modéliser des structures de points agrégées (Diggle, 1983). Un processus de Cox peut être considéré comme une extension d'un processus de Poisson, obtenue en considérant la fonction intensité comme la réalisation d'un champ aléatoire.

Définition 13 Processus de Cox
Soit $Z = \{Z(\xi) : \xi \in E\}$ un champ aléatoire positif tel que l'application $\xi \mapsto Z(\xi)$ soit localement intégrable presque sûrement.
Si la distribution conditionnelle de X sachant Z est un processus de Poisson sur E de fonction d'intensité Z, alors X est appelé ***processus de Cox dirigé par*** *Z.*

La mesure d'intensité du processus de Poisson $X|Z$ est définie par

$$\Lambda(\mathcal{B}) = \int_{\mathcal{B}} Z(\xi)d\xi,\ \mathcal{B} \subseteq E.$$

Si Z est déterministe, le processus X est simplement le processus de Poisson d'intensité $\lambda = Z$.

3.2.2 Modèles de régénération basés sur les processus ponctuels

Dans le modèle de Shimatani (2004), la répartition spatiale des juvéniles est considérée comme la réalisation d'un processus ponctuel marqué (définition 7), les marques étant constituées par les génotypes. Suivant les éléments que l'on souhaite introduire dans le modèle (apport extérieur de graines, survie), la fonction d'intensité λ (définition 9) du processus ponctuel va prendre différentes formes.

Soit $B \subseteq \mathbb{R}^2$ la zone sur laquelle les arbres adultes vivants ont été géoréférencés et génotypés. Soit $A \subset B$ la zone sur laquelle les juvéniles ont été géoréférencés et génotypés.

[2]Processus de Poisson pour lequel la fonction d'intensité est constante égale à 1.

On note x_h la position de l'adulte h. Soit x un point de A. Soit G le génotype d'un juvénile potentiel situé au point x. La fonction d'intensité du processus ponctuel considéré est notée $\lambda(x, G)$.

Modèle sans apport extérieur de graines

Shimatani (2004) propose de modéliser la répartition spatiale des juvéniles par un processus de Poisson hétérogène de fonction d'intensité :

$$\lambda(x) = \sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h}) \tag{3.2}$$

où la somme porte sur l'ensemble des adultes situés dans B. $d_{x;\, y}$ désigne la distance entre les points x et y, g le noyau de dispersion des graines et U_h le succès reproducteur de l'arbre h. Chaque adulte reproducteur h, situé dans la zone B, participe à la reproduction de manière indépendante et produit des graines qui sont dispersées autour de lui. Ces graines peuvent donner naissance ou non à une plantule. La quantité $\lambda(x)\Delta x$ traduit la probabilité qu'un juvénile soit présent dans l'aire infinitésimale Δx centrée sur le point x. Cette probabilité dépend du succès reproducteur U_h de chaque adulte situé dans la zone B, de la localisation des adultes et du noyau de dispersion des graines g. Le succès reproducteur U_h décrit le nombre de juvéniles installés ayant pour mère h. Le noyau g est une fonction positive telle que $\int_0^{+\infty} g(r) 2\pi r \mathrm{d}r = 1$. Ce noyau est, en général une fonction qui décroît exponentiellement avec la distance à l'adulte h (Tufto et al., 1997). Des fonctions non décroissantes telle que la distribution log-normale peuvent être utilisées pour traduire un déficit de graines à proximité des adultes (Stoyan et Wagner, 2001). Dans ce modèle, tous les juvéniles sont issus d'une graine produite par un arbre adulte situé dans la zone B ; il n'y a pas d'apport extérieur de graines.

En classant les juvéniles suivant leur génotype G, le processus de Poisson hétérogène (équation 3.2) peut être étendu à un processus de Poisson multivarié composé d'autant de processus de Poisson indépendants qu'il y a de génotypes observés chez les juvéniles échantillonnés sur la zone A. Cela est équivalent à définir un processus ponctuel marqué (Van Lieshout, 2000). Chaque processus de Poisson hétérogène indépendant a pour fonction d'intensité :

$$\lambda(x, G) = \sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h}) \mathbb{P}(G|h)$$

où $\mathbb{P}(G|h)$ est la probabilité pour qu'un juvénile ayant pour mère h présente le génotype G. Autrement dit, chaque adulte h donne un nombre de juvéniles de génotype G proportionnel à $U_h \mathbb{P}(G|h)$. La probabilité $\mathbb{P}(G|h)$ est donnée par :

$$\mathbb{P}(G|h) = \frac{\sum_{j \neq h} f(d_{x_j;\, x_h}) \mathbb{P}(G|h, j) w(h, j)^{-1}}{\sum_{j \neq h} f(d_{x_j;\, x_h}) w(h, j)^{-1}} \tag{3.3}$$

où $d_{x_j;\, x_h}$ désigne la distance entre l'adulte j et l'adulte h, f le noyau de dispersion du pollen et $\mathbb{P}(G|h, j)$ la probabilité pour qu'un juvénile ayant pour mère h et pour père j

présente le génotype G. Les probabilités $\mathbb{P}(G|h,j)$ sont calculées sous l'hypothèse que la population considérée soit à l'équilibre d'Hardy-Weinberg (Falconer, 1974; Hartl et Clark, 1997). L'annexe B présente le calcul détaillé de $\mathbb{P}(G|h,j)$. La fonction f est une fonction positive telle que $\int_0^{+\infty} f(r)2\pi r\mathrm{d}r = 1$. La quantité $w(h,j)$ désigne la proportion de la circonférence du cercle de centre h et de rayon $d_{x_j;\,x_h}$ se situant à l'intérieur de B. Les coefficients $w(h,j)$ permettent de corriger les effets de bord et sont appelés coefficients de correction de Ripley. A moins que la zone B soit très vaste ou qu'elle soit très isolée, les arbres adultes peuvent être pollinisés par des arbres situés en dehors de B. L'équation 3.3 traduit le fait qu'il existe $w(h,j)^{-1}$ arbres ayant le même génotype que l'arbre j à l'extérieur de B et que le pollen ne peut pas être dispersé au delà de la distance $\max_j\{d_{x_j;\,x_h}\}$. Étant donné que $j \neq h$ dans l'expression de $\mathbb{P}(G|h)$, cela suppose qu'il n'y a pas de reproduction par autofécondation. Notons qu'il existe d'autres méthodes pour corriger les effets de bord ; pour plus de détails, on pourra se reporter à l'ouvrage de Cressie (1991).

Modèle avec apport extérieur de graines

Tous les juvéniles de la zone A n'ont pas obligatoirement une mère située dans la zone B. Certains peuvent être issus d'une graine produite par un adulte reproducteur situé en dehors de la zone B. On note $\overline{B}$ l'extérieur de B.

Le modèle de Shimatani a été étendu pour pouvoir prendre en compte l'apport extérieur de graines (Shimatani et al., 2006, 2007). La répartition spatiale des juvéniles est modélisée par un processus de Poisson hétérogène multivarié dont l'intensité associée à chaque processus de Poisson se décompose sous la forme d'une somme, le premier terme correspondant aux juvéniles dont les mères sont dans la zone B et le second aux juvéniles dont les mères sont dans $\overline{B}$:

$$\lambda(x,G) = \sum_{h:\,x_h\in B} U_h g(d_{x;\,x_h})\mathbb{P}(G|h) + \sum_{h:\,x_h\in\overline{B}} U_h g(d_{x;\,x_h})\mathbb{P}(G|\mathrm{ext}). \tag{3.4}$$

$\mathbb{P}(G|\mathrm{ext})$ désigne la probabilité qu'un juvénile ayant une mère hors de B présente le génotype G. Le calcul de $\mathbb{P}(G|\mathrm{ext})$ est présenté dans l'annexe B.

Le second terme de la somme n'est pas calculable explicitement puisque les positions des mères situées dans $\overline{B}$ sont inconnues. On distingue deux approches suivant la manière dont ce terme va être approché.

• Les juvéniles issus d'une graine produite par un adulte situé à l'extérieur de B sont répartis uniformément dans la zone A.

Cette approche consiste à poser $m = \sum_{h:\,x_h\in\overline{B}} U_h g(d_{x;\,x_h})$ (Shimatani et al., 2006). Le paramètre m correspond à l'ensemble des juvéniles dont la mère appartient à $\overline{B}$. La quantité $m\Delta x$ représente la probabilité qu'un juvénile ayant une mère dans $\overline{B}$ apparaisse dans l'aire infinitésimale Δx centrée en x. L'intensité de chaque processus de Poisson hétérogène s'écrit

alors :

$$\lambda(x, G) = \sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h}) \mathbb{P}(G|h) + m\mathbb{P}(G|\text{ext}).$$

Le second terme ne dépend plus de x, on considère que les juvéniles issus d'une graine produite par un adulte situé à l'extérieur de B sont répartis uniformément dans A. La probabilité $\mathbb{P}(G|h)$ est la probabilité d'obtenir un juvénile de génotype G sachant que sa mère est h et qu'elle a été fécondée par un pollen arbitraire ; le flux de dispersion du pollen n'est pas modélisé.

• Les juvéniles issus d'une graine produite par un adulte situé à l'extérieur de B ne sont pas répartis uniformément dans la zone A.

En général, nous nous attendons à observer plus de juvéniles ayant une mère située à l'extérieur de la zone B sur les bords de la zone A qu'en son centre. Les juvéniles issus de la migration des graines ne sont pas répartis uniformément sur la zone A. La seconde approche consiste donc à déterminer une approximation du second terme de la fonction d'intensité (équation 3.4) qui tienne compte de la localisation x (Shimatani et al., 2007). Le succès reproducteur de chacun des adultes situés en dehors de B est tout d'abord remplacé par un succès reproducteur commun à tous ces arbres, noté $\overline{U}$. Le second terme de la fonction d'intensité s'écrit donc :

$$\sum_{h:\, x_h \notin B} U_h g(d_{x;\, x_h}) \approx \overline{U} \sum_{h:\, x_h \notin B} g(d_{x;\, x_h}).$$

La somme $\displaystyle\sum_{h:\, x_h \notin B} g(d_{x;\, x_h})$ est approchée en utilisant la relation suivante :

$$\sum_{h:\, x_h \in B \cup \overline{B}} g(d_{x;\, x_h}) \approx \int_{B \cup \overline{B}} \varrho(x_h) g(d_{x;\, x_h}) \mathrm{d}x_h$$

où $\varrho(x)$ est l'intensité locale du processus ponctuel qui décrit la répartition spatiale des adultes. En faisant l'hypothèse que la répartition spatiale des adultes est homogène ($\varrho(x) \equiv \overline{d}$), l'approximation devient alors :

$$\sum_{h:\, x_h \in B \cup \overline{B}} g(d_{x;\, x_h}) \approx \overline{d} \int_{B \cup \overline{B}} g(d_{x;\, x_h}) \mathrm{d}x_h$$

où $\overline{d}$ désigne la densité d'adultes. En pratique, nous disposons seulement d'informations sur les adultes présents dans B, d'où :

$$\sum_{h:\, x_h \in B} g(d_{x;\, x_h}) \approx \overline{d} \int_{B} g(d_{x;\, x_h}) \mathrm{d}x_h.$$

On en déduit donc la quantité correspondant à la contribution des mères situées dans $\overline{B}$:

$$\sum_{h:\, x_h \notin B} g(d_{x;\, x_h}) \approx \overline{d} \left(1 - \int_{B} g(d_{x;\, x_h}) \mathrm{d}x_h\right).$$

Nous pouvons donc approcher le second terme de la fonction d'intensité par la relation :

$$\sum_{h:\, x_h \notin B} U_h g(d_{x;\, x_h}) = \overline{U}\,\overline{d}\beta_x \text{ avec } \beta_x = \left(1 - \int_B g(d_{x;\, x_h})\mathrm{d}x_h\right).$$

Finalement, l'intensité du processus s'écrit sous la forme :

$$\lambda(x, G) = \sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h})\mathbb{P}(G|h) + \overline{U}\,\overline{d}\beta_x \mathbb{P}(G|\text{ext}). \tag{3.5}$$

La probabilité $\mathbb{P}(G|h)$ peut être réécrite en tenant compte de l'apport extérieur de pollen. Par un raisonnement analogue à celui effectué pour approcher le second terme de l'intensité, on obtient :

$$\mathbb{P}(G|h) = \frac{\sum_{j:x_j \in B} f(d_{x_h;\, x_j})\mathbb{P}(G|h, j) + \eta_h \overline{d}\mathbb{P}(G|h, \text{ext})}{\sum_{j:x_j \in B} f(d_{x_h;\, x_j}) + \eta_h \overline{d}}$$

avec $\eta_h = 1 - \int_B f(d_{x_h;\, x_j})\mathrm{d}x_j$. $\mathbb{P}(G|h, \text{ext})$ désigne la probabilité qu'un juvénile ayant pour mère h et un père hors de B présente le génotype G. Le calcul de cette probabilité est présenté dans l'annexe B. Notons qu'ici le pollen peut provenir de n'importe quel adulte situé dans la zone B, autrement dit la reproduction par autofécondation est possible.

Cette approche offre l'avantage de permettre de modéliser le flux de pollen contrairement à la précédente.

Modèle avec survie

Il est possible d'ajouter une quantité $\mathcal{S}$ permettant de modéliser la survie des graines entre leur dispersion et l'installation des juvéniles dans chacun des modèles décrits précédemment (Shimatani et al., 2006). Étudier la survie sous-entend que l'on s'intéresse à une cohorte, c'est-à-dire à un ensemble d'individus de même âge, puisque, toutes choses étant égales par ailleurs, la probabilité de survie jusqu'à l'âge a décroît avec a. Pour les arbres, on s'intéresse, en général, à des cohortes de taille (juvéniles) plutôt que d'âge. Nous faisons l'hypothèse que la survie des juvéniles ne dépend que de l'environnement au point x ; il n'y a pas d'interaction génotype-environnement. Si l'on considère un modèle avec apport extérieur de graines, la fonction d'intensité de chacun des processus de Poisson hétérogènes indépendants s'écrit alors :

$$\lambda(x, G) = \mathcal{S}(x)\left(\sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h})\mathbb{P}(G|h) + \overline{U}\,\overline{d}\beta_x \mathbb{P}(G|\text{ext})\right). \tag{3.6}$$

Shimatani et al. (2006) proposent de faire dépendre la survie des juvéniles de leur accès à la lumière en écrivant la survie $S(x)$ sous la forme :

$$\mathcal{S}(x) = C_0 F(d_{x;\, x_p})$$

où p est l'index de l'adulte le plus proche de x, C_0 une constante et F un noyau. On choisit, en général, pour F une distribution asymétrique ayant un maximum proche de l'origine comme la distribution log-normale ou la distribution de Weibull.

Remarquons que, pour pouvoir estimer les paramètres relatifs à la survie, il est nécessaire de travailler sur une zone où les juvéniles ont été échantillonnés de manière exhaustive. En effet, sinon, les juvéniles non échantillonnés sont confondus avec les juvéniles morts avant leur installation dans le peuplement et la survie n'est pas estimable.

3.3 Modèle de recrutement avec prise en compte de variables environnementales

3.3.1 Description du modèle

Le modèle que nous proposons est une extension du modèle de Shimatani et al. (2006) pour lequel la survie dépend de variables environnementales. La répartition spatiale des juvéniles et leurs génotypes sont modélisés par un processus de Poisson hétérogène multivarié. La fonction d'intensité de chaque processus de Poisson est donnée par :

$$\lambda(x,G) = \mathcal{S}(x)\left(\sum_{h:\, x_h\in B} U_h g(d_{x;\,x_h})\mathbb{P}(G|h) + \overline{U}\,\overline{d}\beta_x \mathbb{P}(G|\text{ext})\right) \tag{3.7}$$

avec

$$\mathcal{S}(x) = \frac{\exp(\delta + \gamma'\mathbf{Y}(x))}{1+\exp(\delta+\gamma'\mathbf{Y}(x))}$$

où $\mathbf{Y}(x)$ est un vecteur à k composantes représentant l'environnement au point x, $\gamma \in \mathbb{R}^d$ et $\delta \in \mathbb{R}$.

Si l'on considère l'environnement comme connu, le processus est un processus de Poisson hétérogène multivarié. Si l'on considère l'environnement comme aléatoire, le processus est un processus de Cox hétérogène.

Le nombre de paramètres à estimer dans ce modèle est d'autant plus important que le nombre d'adultes dans B est élevé. Pour réduire le nombre de paramètres à estimer, il est possible de simplifier le modèle en considérant que tous les arbres adultes ont le même succès reproducteur U qu'ils soient ou non dans B. L'intensité du processus s'écrit alors :

$$\lambda(x,G) = \mathcal{S}(x)U\left(\sum_{h:\, x_h\in B} g(d_{x;\,x_h})\mathbb{P}(G|h) + \overline{d}\beta_x \mathbb{P}(G|\text{ext})\right).$$

Une autre solution consiste à écrire le succès reproducteur de chaque adulte en fonction de covariables. Par exemple, Goto et al. (2006) considère le succès reproducteur U_h de chaque adulte comme une fonction de son diamètre D_h et de l'intensité de la floraison. En écologie, les études menées sur la relation possible entre le diamètre de l'arbre et

le succès reproducteur conduisent à des résultats différents selon les espèces considérées (Burgos et al., 2008; Yasaka et al., 2008; Snook et al., 2005; Nabe-Nielsen et al., 2009). Snook et al. (2005) ont montré que la production de fruits (et donc de graines) augmente exponentiellement avec le diamètre chez le mahogany (*Swietenia macrophylla* King), une espèce tropicale. En ce qui concerne l'angélique, il existe une corrélation positive entre le diamètre et l'activité de reproduction (Jésel, 2005), c'est-à-dire que les arbres de grand diamètre fleurissent et produisent des graines plus fréquemment et plus régulièrement que les arbres de petit diamètre ; mais aucune relation n'a été établie entre le nombre de graines produites et le diamètre de l'arbre. Plusieurs types de relation peuvent être considérés entre succès reproducteur et diamètre :

- une relation exponentielle $U_h = \exp(bD_h)$. On définit alors $\overline{U}$ comme $\overline{U} = \exp(b\overline{D})$, où $\overline{D}$ le diamètre moyen des arbres adultes situés en dehors de B.
- une relation quadratique $U_h = bT_h = b\pi\left(\frac{D_h}{2}\right)^2$ (Sagnard et al., 2007; Schurr et al., 2008) où le succès reproducteur U_h de l'adulte h est le produit d'un coefficient de fécondité b à estimer et de la surface terrière[3] T_h de cet arbre exprimée en cm^2. On pose alors $\overline{U} = b\pi\left(\frac{\overline{D}}{2}\right)^2$ où $\overline{D}$ le diamètre moyen des arbres adultes situés en dehors de B.

3.3.2 Estimation des paramètres du modèle

L'environnement $\mathbf{Y}(x)$ est ici considéré comme connu. Autrement dit, le processus ponctuel qui décrit la répartition spatiale des juvéniles et leurs génotypes est un processus de Poisson hétérogène multivarié. L'estimation des paramètres du modèle est effectuée par maximum de vraisemblance.

Soient $x_1, x_2, \ldots, x_n$ les positions des juvéniles observés sur la zone A. Soit $\boldsymbol{\eta}$ le vecteur de tous les paramètres intervenant dans la définition de la fonction d'intensité $\lambda(x)$ du processus ponctuel.

Remarquons tout d'abord que

$$\lambda(x) = \sum_G \lambda(x, G) = \mathcal{S}(x)\left(\sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h}) + \overline{U}\,\overline{d}\beta_x\right) \tag{3.8}$$

car $\sum_G \mathbb{P}(G|h) = \sum_G \mathbb{P}(G|\text{ext}) = 1$.

Soit N_G le nombre de juvéniles présentant le génotype G. Soit x_i^G la position du $i^{\text{ème}}$ juvénile présentant le génotype G. Nous faisons l'hypothèse que la reproduction se produit indépendamment pour chaque génotype G. En utilisant la relation 3.8 et la définition de la log-vraisemblance d'un processus de Poisson hétérogène (équation 3.1), on montre que

[3]La surface terrière d'un arbre est la surface de la section transversale de son tronc à 1,30 m.

la log-vraisemblance des paramètres du processus de Poisson multivarié s'écrit :

$$\begin{aligned}\ell(\boldsymbol{\eta}; x_1, \ldots, x_n) &= \sum_{G} \sum_{i=1}^{N_G} \ln(\lambda(x_i^G, G; \boldsymbol{\eta})) - \int_A \lambda(x; \boldsymbol{\eta}) \mathrm{d}x \\ &= \sum_{i=1}^{n} \ln(\lambda(x_i, G_i; \boldsymbol{\eta})) - \int_A \lambda(x; \boldsymbol{\eta}) \mathrm{d}x \end{aligned} \tag{3.9}$$

où G_i désigne le génotype du $i^{\text{ème}}$ juvénile observé.

Le vecteur $\boldsymbol{\eta}$ est obtenu par maximisation de la log-vraisemblance (équation 3.9) à l'aide d'un algorithme itératif de type Newton. Cette maximisation s'effectue sous contraintes. Cette procédure d'optimisation est réalisée grâce au logiciel R (fonction *nlminb*) (Venables et al., 2009). La valeur initiale du vecteur de paramètres $\boldsymbol{\eta}$ est choisie aléatoirement. La procédure d'optimisation est répétée plusieurs fois avec différents points de départ pour s'assurer que l'algorithme d'optimisation n'est pas piégé dans un maximum local. Les intégrales intervenant dans le calcul de la log-vraisemblance ne peuvent pas être calculées explicitement : elles sont donc approchées numériquement par la méthodes des rectangles. Pour être rigoureux, les estimations des paramètres devraient être accompagnées de leur écart-type. Pour avoir une idée de la variabilité des estimations, il faudrait déterminer la matrice d'information de Fisher associée à la vraisemblance. Cela ne pose pas de problèmes d'un point de vue théorique. Cependant, les expressions des éléments de la matrice sont relativement complexes. Certaines font intervenir des intégrales dont le calcul n'est pas explicite ; ces intégrales doivent être approchées numériquement. D'un point de vue pratique, le calcul de la matrice d'information de Fisher nécessite donc d'être implémenté, ce qui n'a pas été fait ici. Notons que, dans les références bibliographiques précédemment citées, le calcul de la variabilité des estimations n'est pas effectué non plus et seuls les résultats d'estimation sont considérés.

Notons que certains modèles, comme les modèles sans survie, ne nécessitent pas d'avoir échantillonné tous les juvéniles. Dans ce cas, nous nous intéressons à la probabilité conditionnelle d'observer un juvénile de génotype G dans l'aire infinitésimale Δx centrée sur x sachant qu'il y a au moins un juvénile dans cette zone, ce qui s'exprime sous la forme :

$$\frac{\lambda(x, G)\Delta x}{\lambda(x)\Delta x} = \frac{\lambda(x, G)}{\lambda(x)}.$$

La vraisemblance s'écrit alors $\prod_i \frac{\lambda(x_i, G_i)}{\lambda(x_i)}$. Si l'estimation des paramètres d'un modèle avec survie était envisagée à partir d'un échantillonnage partiel des juvéniles, le terme de survie $\mathcal{S}(x)$ présent en facteur au numérateur et au dénominateur de la vraisemblance s'annulerait. Un échantillonnage partiel des juvéniles ne permet donc pas d'estimer les paramètres d'un modèle avec survie.

3.3.3 Identifiabilité des paramètres du modèle

Les paramètres du processus ponctuel proposé pour modéliser la répartition spatiale et le génotype des juvéniles, dont la fonction d'intensité est donnée par l'équation 3.6 ou par l'équation 3.7, ne sont pas identifiables. En effet, tous les autres paramètres étant égaux par ailleurs, les deux couples $(\overline{U}, \overline{d})$ et $(c\overline{U}, \frac{1}{c}\overline{d})$ avec $c > 0$ conduisent à la même fonction d'intensité et donc à la même vraisemblance. Pour assurer l'identifiabilité des paramètres du modèle, nous adoptons une des deux variantes du modèle présentées ci-dessus. Ces variantes diffèrent par le choix opéré concernant la modélisation du succès reproducteur. Soit on considère tous les succès reproducteurs égaux entre eux, c'est-à-dire, pour tout h tel que $x_h \in B$, $U_h = \overline{U}$; soit le succès reproducteur est considéré comme une fonction du diamètre D_h de l'arbre, c'est-à-dire que, pour tout h tel que $x_h \in B$, $U_h = f(D_h; b)$ et $\overline{U} = f(\overline{D}; b)$. C'est avec ces deux types de modèles que nous travaillons dans la suite.

Des problèmes d'identifiabilité peuvent également survenir pour les modèles où l'on introduit une survie qui dépend de l'environnement et où l'on considère tous les succès reproducteurs identiques égaux à U. L'intensité s'écrit sous la forme :

$$\lambda(x, G) = U \frac{\exp(\delta + \gamma' \mathbf{Y}(x))}{1 + \exp(\delta + \gamma' \mathbf{Y}(x))} \left(\sum_{h:\, x_h \in B} g(d_{x;\, x_h}) \mathbb{P}(G|h) + \overline{d} \beta_x \mathbb{P}(G|\text{ext}) \right).$$

Si les paramètres δ et γ et les variables environnementales $\mathbf{Y}(x)$ sont tels que $\exp(\delta + \gamma' \mathbf{Y}(x))$ est proche de 0 pour tout x de A, on a l'équivalence suivante :

$$\frac{1}{1 + \exp(\delta + \gamma' \mathbf{Y}(x))} \sim 1.$$

L'intensité peut alors être approchée par :

$$\lambda(x, G) \approx U \exp(\delta) \exp(\gamma' \mathbf{Y}(x)) \left(\sum_{h:\, x_h \in B} g(d_{x;\, x_h}) \mathbb{P}(G|h) + \overline{d} \beta_x \mathbb{P}(G|\text{ext}) \right).$$

Dans ce cas, les paramètres U et δ ne sont pas identifiables.

Notons que lorsque le modèle ne comprend qu'une seule variable environnementale et que cette variable $Y(x)$ est une variable ordinale à L modalités, une contrainte supplémentaire doit être ajoutée au modèle pour assurer l'identifiabilité des paramètres. En effet, le terme de survie s'écrit alors :

$$\mathcal{S}(x) = \frac{\exp(\delta + \sum_{l=1}^{L} \gamma_l \mathbb{1}_{\{Y(x)=l\}})}{1 + \exp(\delta + \sum_{l=1}^{L} \gamma_l \mathbb{1}_{\{Y(x)=l\}})}.$$

Comme $\sum_{l=1}^{L} \mathbb{1}_{\{Y(x)=l\}} = 1$, les jeux de paramètres (δ, γ_l) et $(\delta + c, \gamma_l - c)$, $c > 0$ conduisent à la même expression de la survie. Les paramètres ne sont donc pas identifiables. Une solution

consiste soit à fixer le paramètre δ, soit à fixer un des paramètres γ_l à condition que la $l^{\text{ème}}$ modalité soit présente dans les données. L'interprétation des résultats concernant la survie $\mathcal{S}$ est fonction de la contrainte choisie.

3.4 Simulations

L'efficience de la procédure d'estimation des paramètres a été évaluée à l'aide de simulations.

3.4.1 Simulation d'un jeu de données

Pour simuler un jeu de données, nous nous donnons d'abord trois zones rectangulaires A, B, C emboîtées telles ques $A \subseteq B \subseteq C$. La zone A correspond à une zone où tous les juvéniles ont été échantillonnés et la zone B à la zone où les adultes sont échantillonnés. L'ensemble fini $C \backslash B$ joue le rôle de l'ensemble $\overline{B}$ du modèle, qui, lui étant infini, ne peut pas être simulé. La zone C doit donc être choisie suffisamment grande pour ne pas introduire d'artéfact dans la simulation. Ici, nous prenons $A = [140; 440] \times [140; 440]$, $B = [120; 460] \times [120; 460]$ et $C = [0; 580] \times [0; 580]$. Nous nous donnons également le noyau de dispersion du pollen f et le noyau de dispersion des graines g. Dans l'exemple ci-dessous, les noyaux f et g sont choisis gaussiens et on note τ_1^2 et τ_2^2 leurs variances respectives. Autrement dit, la distance moyenne de dispersion du pollen est égale à $\sqrt{\frac{\pi}{2}}\tau_1$ et celle des graines à $\sqrt{\frac{\pi}{2}}\tau_2$ (Austerlitz et al., 2004). Nous nous donnons la densité $\overline{d}$ d'arbres adultes ainsi que les coefficients δ et γ intervenant dans la définition de la survie. Enfin, nous définissons les succès reproducteurs, U_h pour tout h tel que $x_h \in B$ et $\overline{U}$. Le cas le plus simple est celui où l'on considère tous les succès reproducteurs égaux à $\overline{U}$; il suffit alors de se donner la valeur de $\overline{U}$. Si le succès reproducteur est vu comme une fonction du diamètre des arbres, nous nous donnons les diamètres D_h pour tout h tel que $x_h \in B$, le diamètre moyen $\overline{D}$ des arbres situés dans $\overline{B}$ et le coefficient b et nous calculons les valeurs de U_h et $\overline{U}$ correspondantes. Pour pouvoir simuler les génotypes, nous devons connaître le nombre de locus étudiés, le nombre d'allèles par locus et les fréquences alléliques correspondantes. Ici, sauf mention du contraire, les jeux de données sont simulés en considérant 10 loci, avec six formes alléliques par locus et des fréquences alléliques équiréparties.

La simulation du jeu de données comprend cinq parties :

1. *Simulation de la position et du génotype des adultes*
   - Tirage de la position des adultes suivant un processus de Poisson homogène d'intensité $\overline{d}$ sur C. On note M le nombre d'adultes dans C.
   - Tirage du génotype des adultes.
2. *Simulation de la position et du génotype des juvéniles*
   - Tirage du nombre de graines dispersées par chaque adulte de C suivant une loi de Poisson de paramètre U_h si l'adulte est dans B ou de paramètre $\overline{U}$ si l'adulte est dans $C \backslash B$.

- Tirage de la position des graines produites par chaque adulte suivant une loi normale bivariée centrée sur l'adulte h, $\mathcal{N}\left(x_h, \tau_2^2 \mathrm{I}\right)$.
- Tirage du père de chaque juvénile potentiel (graine dispersée) ayant pour mère h suivant une loi multinomiale :
  - de paramètres $(p_1, \ldots, p_j, \ldots, p_M)$ avec $p_j = \dfrac{f(d_{x_h;x_j})}{\sum_{j:\, x_j \in C} f(d_{x_h;x_j})}$, si l'on considère que l'espèce peut se reproduire par autofécondation (le père j peut être le même arbre que la mère h).
  - de paramètres $(p_1, \ldots, p_{h-1}, p_{h+1}, \ldots, p_M)$ avec $p_j = \dfrac{f(d_{x_h;x_j})}{\sum_{j:\, x_j \in C,\, j \neq h} f(d_{x_h;x_j})}$, si l'on considère que l'espèce ne peut pas se reproduire par autofécondation.
- Tirage des génotypes des juvéniles.
- On ne conserve que les juvéniles situés dans A.

3. *Elimination des adultes situés dans $\overline{B}$*
   - On ne conserve que les adultes situés dans B. Le nombre moyen d'adultes présents dans le jeu de données est égal à $\overline{d}|B|$ où $|B|$ désigne l'aire de la zone B.
4. *Simulation de l'environnement*
   - Simulation d'un champ aléatoire représentant l'environnement sur une grille recouvrant A (l'environnement est connu sur toute la zone d'étude).
   - Simulation du champ aléatoire représentant l'environnement aux points où l'on a des juvéniles potentiels connaissant le champ sur la grille.
5. *Simulation de la survie des plantules*
   - Calcul de la survie $\mathcal{S}(x)$ aux points où l'on a des juvéniles potentiels.
   - Tirage des juvéniles qui survivent.

3.4.2 Résultats

Dans un premier temps, nous comparons deux modèles sans survie, l'un pour lequel les graines issues de l'extérieur de B sont réparties uniformément et l'autre pour lequel elles ne le sont pas. Le tableau 3.1 donne un exemple de résultats d'estimation obtenus pour ces deux modèles. L'estimation du paramètre associé au noyau de dispersion des graines pour le modèle avec un apport extérieur de graines réparties uniformément sur la zone A n'est, en général, pas très bonne. La variance τ_2^2 est sur-estimée et le paramètre m, représentant le nombre de graines issues de l'extérieur de B par unité de surface, est sous-estimé. De plus, ce modèle ne permet pas de déterminer la distance de dispersion du pollen. Le modèle considérant les graines issues de $\overline{B}$ comme non uniformément réparties conduit à une estimation correcte des paramètres et permet d'étudier l'ensemble des mécanismes de dispersion. Dans la suite, nous ne considérons donc plus que des modèles pour lesquels nous faisons l'hypothèse que les juvéniles ayant un ou des parents à l'extérieur de la zone B ne sont pas répartis uniformément dans A.

L'information génétique (nombre de loci, fréquences alléliques, etc) disponible joue un rôle important dans l'estimation des paramètres, notamment pour ceux associés aux noyaux

TAB. 3.1 – Estimation des paramètres pour des modèles avec prise en compte d'un apport extérieur de graines. Pour chaque modèle, la première ligne indique les vraies valeurs des paramètres et la seconde les valeurs estimées.

Paramètres	U	τ_1	τ_2	m	$\overline{d}$
Répartition uniforme	2	-	35	0,001	-
	2,63	-	63,20	0,0004	-
Répartition non uniforme	3	75	25	-	0,001
	2,77	74,38	25,96	-	0,0009

de dispersion des graines et du pollen. Dans un second temps, nous avons effectué des simulations pour déterminer l'impact du nombre de loci, du nombre d'allèles par locus et de la distribution des fréquences alléliques sur l'estimation des paramètres pour un modèle sans survie où tous les succès reproducteurs sont égaux à U. Les résultats obtenus sont présentés dans le tableau 3.2. Les estimations de la densité $\overline{d}$ et du paramètre τ_1 associé au

TAB. 3.2 – Comparaison des estimations obtenues pour un modèle sans survie suivant le nombre de loci disponibles, le nombre de formes alléliques par locus et la distribution des fréquences alléliques de chaque locus. Les valeurs des paramètres utilisées pour la simulation sont $U = 2$, $\tau_1 = 95$, $\tau_2 = 35$ et $\overline{d} = 0,001$.

Nombre de loci	**Nombre de formes alléliques**	**Fréquences alléliques équiréparties**				**Fréquences alléliques non équiréparties**			
		U	τ_1	τ_2	$\overline{d}$	U	τ_1	τ_2	$\overline{d}$
5	3	2,02	166,61	33,70	0,0001	2,13	200,00	39,24	0,0001
	6	2,16	128,62	36,59	0,0004	2,01	159,84	32,09	0,0003
10	3	1,98	152,36	37,92	0,0004	2,06	166,53	35,05	0,0002
	6	1,94	111,38	37,34	0,0008	1,91	123,66	31,18	0,0005
15	3	2,18	114,06	37,65	0,0008	2,23	200,00	35,54	0,0002
	6	1,90	90,85	37,55	0,0011	2,21	104,07	34,41	0,0008

noyau de dispersion du pollen sont sensibles à la quantité d'information génétique disponible, alors que les estimations du paramètre U et de la variance τ_2^2 du noyau de dispersion des graines sont plus robustes. Les estimations des paramètres sont d'autant moins biaisées que le nombre de loci et le nombre de formes alléliques par locus sont grands. A nombre de loci et de formes alléliques fixés, les estimations sont d'autant moins biaisées que la dis-

tribution des fréquences alléliques à chaque locus est proche de l'équiprobabilité. En effet, plus les formes alléliques présentes sont variées, plus il y a de chance que la probabilité de déterminer la filiation de chaque juvénile soit élevée. Au vu des résultats présentés dans le tableau 3.2, il ne semble pas possible d'obtenir des estimations correctes des paramètres en utilisant moins de dix loci. Les simulations réalisées par la suite, le sont donc avec dix loci, six formes alléliques par locus et une distribution des fréquences alléliques proches de l'équiprobabilité.

Le modèle proposé est basé sur l'hypothèse que la répartition spatiale des adultes est homogène (cf. page 70). La robustesse du modèle au non respect de cette hypothèse a été testée par des simulations. Le tableau 3.3 présente les estimations des paramètres obtenues pour différents types de répartition spatiale des adultes (homogène ou hétérogène).

TAB. 3.3 – Comparaison des estimations des paramètres obtenues pour différents types de répartition spatiale des adultes (homogène ou hétérogène). La répartition spatiale des adultes est la réalisation d'un processus ponctuel de Poisson (PPP) de fonction d'intensité $\lambda(x, y)$. Le modèle est un modèle sans survie pour lequel tous les succès reproducteurs sont égaux.

Paramètres		U	τ_1	τ_2	$\overline{d}$
Vraies valeurs		2	95	35	0,001
PPP homogène	$\lambda(x,y) = 0,001$	2,22	91,98	34,74	0,0010
PPP homogène	$\lambda(x,y) = 0,001$	2,02	119,13	36,17	0,0006
PPP hétérogène	$\lambda(x,y) = \frac{0,001}{580}(x+y)$	2,04	108,09	40,06	0,0006
PPP hétérogène	$\lambda(x,y) = 0,001 \exp\left(-\frac{3(x-260)}{580}\right)$	2,27	118,95	34,90	0,0007
PPP hétérogène	$\lambda(x,y) = \begin{cases} 0,002 & \text{si } x < 290 \\ 0 & \text{sinon} \end{cases}$	2,16	113,04	34,33	0,0009

Les paramètres estimés à partir d'un jeu de données où les adultes présentent une répartition spatiale hétérogène sont dans la même gamme de valeurs que ceux obtenus à partir d'un jeu de données où les adultes présentent une répartition spatiale homogène. Les paramètres les plus sensibles au non respect de l'hypothèse d'homogénéité sont la densité $\overline{d}$ qui est souvent sous-estimée et la variance du noyau de dispersion du pollen τ_1^2 qui, elle, est sur-estimée.

Dans un troisième temps, des simulations ont été réalisées pour valider la procédure d'estimation des paramètres d'un modèle qui prend en compte les variables environnementales. Les différentes variantes du modèle selon le choix effectué pour la modélisation du succès reproducteur ont été testées. Une seule variable environnementale est incluse dans le modèle. Les résultats sont donnés dans le tableau 3.4. Les estimations des paramètres

TAB. 3.4 – Estimation des paramètres pour des modèles avec prise en compte des variables environnementales. Pour chaque modèle, la première ligne indique les vraies valeurs des paramètres et la seconde les valeurs estimées.

Paramètres	U ou b	τ_1	τ_2	$\overline{d}$	γ	δ
$U = \overline{U}$	50	95	35	0,0005	-5	0,5
	57,51	97,01	36,19	0,00053	$-5,01$	0,386
$U = \exp(bD)$	0,05	95	35	0,001	$-2,5$	7
	0,0495	89,23	38,60	0,0010	$-2,45$	6,76
$U = b\pi(D/2)^2$	0,04	95	35	0,0005	-5	3
	0,039	96,23	31,94	0,00064	$-4,679$	2,784

sont globalement cohérentes avec les valeurs utilisées pour les simulations, quelle que soit la modélisation choisie pour le succès reproducteur. Seul le paramètre τ_1 est sous-estimé lorsque la relation entre le succès reproducteur et le diamètre est exponentielle. Les paramètres δ et γ associés aux variables environnementales sont correctement estimés.

Nous avons ensuite testé les performances de la procédure d'estimation sur un modèle permettant de prendre en compte plusieurs variables environnementales. Dans ces exemples, le succès reproducteur augmente exponentiellement avec le diamètre de l'arbre.

TAB. 3.5 – Estimation des paramètres pour des modèles avec prise en compte de plusieurs variables environnementales. Le nombre de composantes k du vecteur $\mathbf{Y}(x)$ représentant l'environnement est donné à gauche. Pour chaque modèle, la première ligne indique les vraies valeurs des paramètres et la seconde les valeurs estimées.

Paramètres	b	τ_1	τ_2	$\overline{d}$	γ_1	γ_2	γ_3	γ_4	δ
$\mathbf{k = 2}$	0,05	95	35	0,001	-2.5	1	-	-	3
	0,045	93,74	34,72	0,0015	$-2,57$	1,17	-	-	3,11
$\mathbf{k = 3}$	0,1	75	25	0,001	$-2,5$	1	$-0,2$	-	2
	0,088	72,29	25,38	0,0014	$-3,63$	1,36	$-0,25$	-	3,31
$\mathbf{k = 4}$	0,05	95	35	0,001	-2.5	1	$-0,25$	4	3
	0,050	88,14	32,59	0,0008	$-2,294$	0,93	$-0,21$	3,39	2,62

Comme le montre le tableau 3.5, il est possible de traiter un modèle comportant plusieurs variables environnementales grâce à la procédure d'estimation proposée. Cependant, au delà d'un certain nombre de paramètres à estimer, la fonction d'optimisation *nlminb*

peut s'avérer inefficace pour maximiser la vraisemblance. Si le nombre de paramètres du modèle est trop important, il faut alors envisager la mise en œuvre de procédures d'optimisation différentes comme les procédures d'optimisation stochastique.

3.5 Variabilité des estimations liée à la prédiction de l'environnement

Les variables environnementales interviennent dans la modélisation du recrutement en tant que variables explicatives de la survie (équation 3.6). La valeur de ces variables aux points non échantillonnés est prédite. Classiquement, les prédictions sont considérées comme les vraies valeurs de ces variables. Pourtant, celles-ci sont entachées d'erreur. Après avoir mis en évidence l'impact de ces erreurs sur l'estimation des paramètres du processus ponctuel modélisant le recrutement, nous proposons plusieurs approches pour prendre en compte et mesurer cette variabilité due à la prédiction de l'environnement.

3.5.1 Mise en évidence de l'impact des erreurs de prédiction de l'environnement sur l'estimation des paramètres du processus ponctuel

Des simulations ont été effectuées pour quantifier l'impact des erreurs de prédiction sur l'estimation des paramètres du processus ponctuel. Le modèle de recrutement considéré ici est un modèle avec survie, avec apport extérieur de graines « non uniforme » et où le succès reproducteur est une fonction exponentielle du diamètre de l'arbre $U = \exp(bD)$. Les paramètres sont estimés une première fois avec un environnement connu, puis une seconde fois avec un environnement prédit à partir de 100 points tirés aléatoirement dans la zone d'étude. Les prédictions sont réalisées à l'aide d'un krigeage ordinaire. Le tableau 3.6 présente les résultats obtenus pour des variables environnementales ayant différentes structures spatiales. La prédiction de l'environnement a une répercussion sur l'estimation des paramètres du processus ponctuel, en particulier sur les paramètres γ et δ associés à la survie et sur le paramètre b associé au succès reproducteur. La variabilité engendrée par la prédiction de l'environnement doit donc être considérée.

3.5.2 Impact de la prédiction de l'environnement sur la prédiction de la régénération

Les simulations ont montré que la prédiction de l'environnement a un impact sur l'estimation des paramètres du processus ponctuel, et donc sur la prédiction de la régénération. Les erreurs de prédiction de l'environnement entraînent notamment une modification des estimations des paramètres associés à la survie des juvéniles. Nous nous attendons également à voir augmenter la variabilité des estimations des paramètres du processus ponctuel lorsque l'environnement est prédit. Pour confirmer ou infirmer cette hypothèse et mesurer la variabilité des estimations, nous envisageons d'estimer simultanément les paramètres associés à l'environnement et ceux associés au recrutement grâce à une approche hiérarchique. Le modèle spatial multivarié proposé au chapitre 1 pour prédire l'environnement est complété : des niveaux hiérarchiques supplémentaires sont ajoutés pour pouvoir estimer les

TAB. 3.6 – Comparaison des estimations des paramètres du processus ponctuel obtenues avec un environnement connu (env. connu) et un environnement prédit (env. prédit) pour différentes variables environnementales. Les variables environnementales utilisées sont des variables gaussiennes ayant un effet de pépite égal à 1 et une tendance égale à 5. Leurs fonctions de covariance sont indiquées à gauche dans le tableau. Pour chaque variable environnementale, la première ligne donne les valeurs réelles des paramètres, la seconde les estimations à environnement connu et la troisième les estimations à environnement prédit.

$C(\mathbf{h})$	b	τ_1	τ_2	$\overline{d}$	γ	δ
$8\exp\left(-\left(\frac{\mathbf{h}}{12}\right)^2\right)$	$0,14$	95	35	$0,001$	$-0,5$	$-3,5$
	$0,160$	$97,61$	$33,94$	$0,00089$	$-0,48$	$-4,48$
	$0,168$	$96,67$	$34,08$	$0,00094$	$-0,27$	$-4,76$
$20\exp\left(-\left(\frac{\mathbf{h}}{50}\right)^2\right)$	$0,14$	95	35	$0,001$	$-0,5$	-5
	$0,165$	$93,35$	$34,61$	$0,0085$	$-0,48$	$-6,22$
	$0,082$	$92,39$	$34,35$	$0,0094$	$-0,08$	$-2,36$
$10\exp\left(-\left(\frac{\mathbf{h}}{25}\right)\right)$	$0,14$	95	35	$0,001$	$-0,5$	-4
	$0,144$	$90,86$	$38,77$	$0,0010$	$-0,47$	$-4,06$
	$0,126$	$90,32$	$39,15$	$0,0011$	$-0,42$	$-2,32$

paramètres du processus ponctuel modélisant le recrutement. Pour simplifier, on considère ici un modèle ne faisant intervenir qu'une seule variable environnementale Y gaussienne. Soient μ et ν^2 la moyenne et l'effet de pépite de $Y(\mathbf{s})$. Soit X le processus représentant la répartition spatiale des juvéniles, $\boldsymbol{\eta}$ le vecteur des paramètres qui lui sont associés et λ sa fonction d'intensité. Soit $\mathbf{Y}$ le vecteur de la variable environnementale aux n points de mesure. On note $\mathbf{Y}_0$ le vecteur de la variable environnementale considérée aux n_0 où l'on souhaite effectuer des prédictions. Ces n_0 points sont ceux constituant une grille fine recouvrant l'ensemble de la zone d'étude. La prédiction de l'environnement sur cette grille nous permet de considérer que ce dernier est connu sur toute la zone. Les différents niveaux de l'approche hiérarchique sont les suivants :

$X_{\mathcal{B}}|N(\mathcal{B}), \mathbf{Y}_0, \boldsymbol{\eta} \sim \text{Binomial}(E, N(\mathcal{B}), p)$ avec $p(\xi) = \lambda(\xi)/\Lambda(\mathcal{B})$
où A désigne la zone d'étude, $\mathcal{B}$ un borélien de A et où $\xi \in A$,

$N(\mathcal{B})|\mathbf{Y}_0, \boldsymbol{\eta} \sim \mathcal{P}(\Lambda(\mathcal{B}))$, $\forall \mathcal{B} \subseteq E$
où $\Lambda(\mathcal{B}) = \displaystyle\int_{\mathcal{B}} \lambda(\xi; \boldsymbol{\eta}) \mathrm{d}\xi$,

$$\mathbf{Y}, \mathbf{Y}_0|\mathbf{S}, \mathbf{S}_0, \mu, \nu \sim \mathcal{N}_{n+n_0}\left(\mu\mathbf{1} + \begin{pmatrix} \mathbf{S} \\ \mathbf{S}_0 \end{pmatrix}, \nu^2\mathbf{I}_{n+n_0}\right),$$

$$\mathbf{S}, \mathbf{S}_0|\boldsymbol{\theta} \sim \mathcal{N}_{n+n_0}(\mathbf{0}, \mathbf{C}),$$

$\boldsymbol{\theta}, \mu, \nu, \boldsymbol{\eta} \sim$ lois *a priori*.

Les deux premiers niveaux de la hiérarchie concernent la modélisation de la répartition spatiale des juvéniles. Le second niveau indique que le nombre de juvéniles attendus dans $\mathcal{B}$ suit une loi de Poisson. Les trois derniers niveaux de la hiérarchie concernent la modélisation de l'environnement selon l'approche hiérarchique décrite au paragraphe 1.3.2 du chapitre 1.

La procédure d'inférence d'un tel modèle peut être gourmande en ressources informatiques. Le principal avantage de cette méthode est que les erreurs commises lors de la prédiction de l'environnement sont directement prises en compte dans la procédure d'estimation des paramètres du processus ponctuel. Il est ainsi possible de déterminer l'écart-type des estimations des paramètres relatifs au recrutement sous un environnement prédit.

La modélisation du recrutement doit permettre de déterminer le nombre de juvéniles dans les zones non échantillonnées et de caractériser leur structure spatiale connaissant l'environnement. Soit l'environnement est connu de manière exacte, soit il est prédit (sur une grille fine) à partir des m points d'échantillonnage. Soit A une zone où les juvéniles n'ont pas été échantillonnés. Nous souhaitons comparer le nombre moyen de juvéniles attendus dans la zone A à environnement connu, $\mathbb{E}[N(A)]$, au nombre moyen de juvéniles attendus à environnement prédit, $\mathbb{E}[N(A)]_p$. Soit $\boldsymbol{\eta}$ le vecteur des paramètres du processus ponctuel modélisant la répartition spatiale des juvéniles. On note $\hat{\boldsymbol{\eta}}$ l'estimateur de $\boldsymbol{\eta}$ obtenu quand la variable environnementale Y est connue, c'est-à-dire quand $Y(x)$ est déterministe, et $\hat{\boldsymbol{\eta}}_p$ l'estimateur de $\boldsymbol{\eta}$ obtenu quand la variable environnementale Y est prédite, c'est-à-dire quand $Y(x)$ est considérée comme la réalisation d'un champ aléatoire. Nous nous intéressons donc à la différence

$$\widehat{\mathbb{E}[N(A)]}_p - \widehat{\mathbb{E}[N(A)]} = \int_A \lambda(x; \hat{\boldsymbol{\eta}}_p) - \lambda(x; \hat{\boldsymbol{\eta}}) \mathrm{d}x.$$

Il s'agit de savoir si l'estimation du nombre de juvéniles attendus dans la zone A obtenue à partir d'une prédiction de l'environnement basée un échantillonnage de m points, est correcte. Cette question peut être reformulée sous la forme d'un problème d'échantillonnage qui consisterait à déterminer le nombre de points m qu'il faut échantillonner pour assurer une prédiction « correcte » du nombre de juvéniles et de leur organisation spatiale sur la zone A. D'un point de vue mathématique, cela conduit à étudier le comportement asymptotique de la différence $\widehat{\mathbb{E}[N(A)]}_p - \widehat{\mathbb{E}[N(A)]}$ quand le nombre de points échantillonnés m tend vers l'infini. Cette piste reste à explorer. Nous pourrons nous appuyer sur les travaux réalisés par Rathbun (1996), Rathbun et al. (2007) et Waagepetersen (2008) portant sur l'étude des effets de covariables partiellement observées sur l'intensité d'un processus ponctuel dont le logarithme de l'intensité s'écrit comme une fonction linéaire des covariables.

3.6 Conclusion

Prédire la répartition spatiale et le génotype des juvéniles nécessite d'utiliser des modèles de régénération incluant de l'information génétique. Ces données génétiques permettent, entre autre, d'obtenir des estimations plus fiables des distances de dispersion des graines et du pollen. Le modèle proposé ici s'appuie sur la théorie des processus ponctuels. Ce modèle est d'autant plus intéressant qu'il est interprétable biologiquement. Un terme modélisant la survie des juvéniles en fonction des variables environnementales est intégré au modèle. Notons qu'une extension du modèle pourrait être envisagée en écrivant la survie comme une fonction de l'environnement et du génotype des juvéniles afin d'étudier les effets des interactions génotype-environnement sur la régénération. Le succès reproducteur U doit être modélisé en tenant compte des connaissances dont on dispose sur l'espèce étudiée. Il faudra cependant s'assurer de l'identifiabilité des paramètres du modèle.

Alors que nous traitons le processus ponctuel modélisant la régénération comme un processus de Poisson hétérogène dans les simulations en considérant les variables environnementales comme connues, ce processus est, en réalité, un processus de Cox car les variables environnementales sont prédites. Cette prédiction de l'environnement introduit une variabilité dans l'estimation des paramètres du processus ponctuel. Estimer simultanément les paramètres liés à l'environnement et à la régénération grâce à un modèle hiérarchique permettrait de déterminer la variabilité des paramètres associés à la régénération lorsque l'environnement est prédit. L'impact de la prédiction de l'environnement sur la prédiction du nombre de juvéniles attendus dans les zones non échantillonnées et sur leur organisation spatiale doit également être étudié. Ces questions restent ouvertes pour de futures recherches.

Chapitre 4

Prédiction de la régénération de l'angélique (*Dicorynia guianensis* Amshoff) en Guyane française

Sommaire

Les modèles statistiques développés aux chapitres 1 et 3 sont appliqués à la prédiction de la régénération de l'angélique (*Dicorynia guianensis* Amshoff) à partir de données collectées sur le dispositif expérimental de Paracou en Guyane française.

4.1 Le dispositif expérimental de Paracou

Les données dont nous disposons ont été recueillies sur le dispositif expérimental de Paracou. Ce site est dédié à l'étude des effets des interventions sylvicoles sur la reconstitution du peuplement forestier.

4.1.1 Situation géographique du dispositif

Le dispositif de Paracou est implanté sur la commune de Sinnamary, à une soixantaine de kilomètres à l'ouest de Kourou et à une quinzaine de kilomètres de la côte (5°15'N, 52°55'O) (Gourlet-Fleury et al., 2004). Le climat est de type tropical humide, marqué par une alternance entre saison sèche et saison des pluies. La pluviométrie annuelle moyenne est de 3 081 mm et la température annuelle moyenne de 26°C. Le relief est composé d'une succession de petites collines (100-300 m de diamètre) pouvant présenter de fortes pentes, séparées les unes des autres par des zones humides appelées bas-fonds (Epron et al., 2006). L'altitude du site est comprise entre 0 et 50 m. Les sols du dispositif expérimental sont essentiellement des acrisols[1] (FAO-ISRIC-ISSS, 1998). Le sous-sol correspond à une formation métamorphique du Précambrien. Le sol est composé de schistes, de grès et peut présenter des veines de pegmatite[2], d'aplite[3] et de quartz.

4.1.2 Description du dispositif expérimental

Installé en 1984, le dispositif expérimental comprenait à l'origine 12 parcelles de 6,25 ha[4]. Entre 1986 et 1987, 9 des 12 parcelles ont subi des traitements sylvicoles (exploitation, exploitation + éclaircie) d'intensité croissante. Trois traitements ont été appliqués, chacun répété trois fois. Trois parcelles sont restées intactes constituant des témoins (figure 4.1). Au début des années 90, le dispositif a été complété par l'installation de 3 parcelles de 6,25 ha et une de 25 ha non perturbées.

4.1.3 Inventaire du peuplement

Dans chaque parcelle, tous les arbres de plus de 10 cm dbh ont été identifiés et géoréférencés. Les trois composantes de la dynamique forestière (accroissement, mortalité et recrutement) sont suivies depuis 1984 pour les 12 premières parcelles et depuis 1991 pour les autres. Cet inventaire a été complété par un échantillonnage exhaustif des individus de 1 à 10 cm de diamètre d'une quinzaine d'espèces dont l'angélique, sur la partie sud du

[1] Sol riche en argile associé au climat tropical humide.

[2] Roche magmatique à grands cristaux de taille supérieure à 20 mm. La plupart des pegmatites ont une composition granitique.

[3] Roche magmatique granitique composée de grains très fins.

[4] Source : http ://www.cirad.fr/guyane

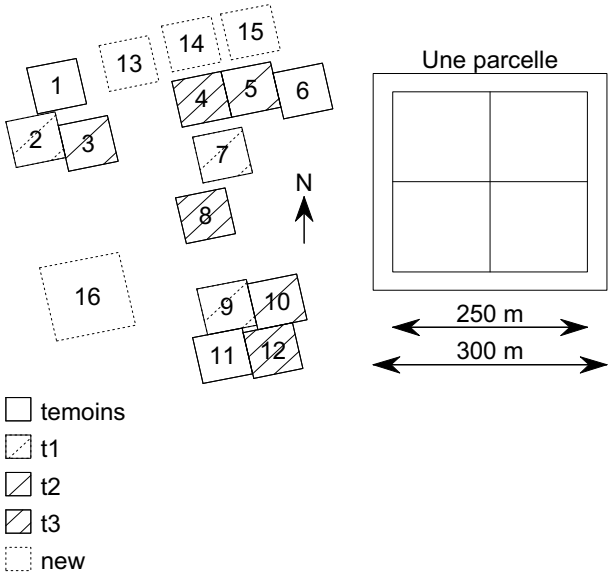


FIG. 4.1 – Schéma des parcelles permanentes de suivi du dispositif expérimental de Paracou en Guyane française

dispositif (parcelles 9 à 12), dans l'optique d'étudier le phénomène de régénération (Flores, 2005). Notre zone d'étude pour la régénération de l'angélique se limite donc aux quatre parcelles constituant le bloc sud du dispositif.

Les études menées à partir de ces différents inventaires ont permis de mieux comprendre les écosystèmes forestiers. Elles ont notamment permis d'améliorer la connaissance des vitesses de reconstitution du stock des principales espèces exploitées, et ainsi de raisonner les durées de rotation dans les forêts aménagées du nord de la Guyane.

4.1.4 Données pédologiques

Outre les données sur le peuplement, plusieurs types de données caractérisant l'environnement sont disponibles. Des relevés topographiques (altitude, pente) ont été effectués lors de la mise en place des parcelles permanentes. Les conditions édaphiques ont aussi fait l'objet de mesures. La texture du sol, sa couleur, la présence ou l'absence d'éléments grossiers et de tâches de couleur (humidité) ont été relevés pour distinguer les différents types de régime hydrique qui sont au nombre de six (Sabatier et al., 1997; Ferry et al., 2003). La teneur du sol en éléments chimiques (C,N,P,Al,Ca,Mg,K) a été mesurée sur quelques parcelles. Environ 70 relevés ont été effectués par parcelle ; ces points de mesure sont répartis aléatoirement sur chaque parcelle.

Nous nous intéressons plus particulièrement aux données environnementales recueillies sur le bloc sud (parcelles 9, 10, 11 et 12) du dispositif de Paracou, zone d'étude de la régénération. Les quatre parcelles du bloc sud n'ont pas fait l'objet de mesures chimiques.

La topographie du bloc sud a été étendue à une partie des zones tampons comprises entre les parcelles (Figure 4.2). Un jeu de données trivarié composé de la pente, de l'altitude et

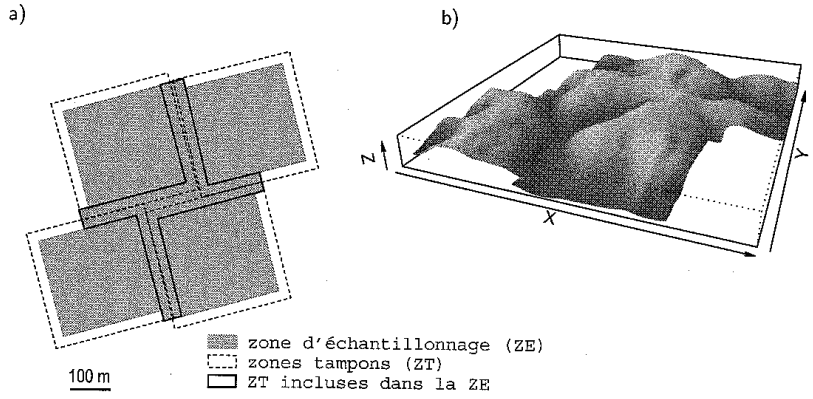


FIG. 4.2 – Figure extraite de la thèse de Flores (2005). a) Carte du bloc sud. La zone d'échantillonnage des populations (ZE) couvre quatre parcelles, mais n'inclut pas toutes les zones tampons (ZT). b) Représentation de la topographie du bloc sud à partir d'un modèle numérique de terrain. L'axe Y est orienté selon une direction N-S.

du drainage et auquel nous allons appliquer le modèle hiérarchique spatial multivarié, a été constitué à partir de ces données. La pente est une variable gaussienne. L'altitude n'est pas considérée ici comme une variable continue, mais comme une variable ordinale à trois modalités (altitude inférieure à 20 m, altitude comprise entre 20 et 30 m, altitude supérieure à 30 m). Comme nous l'avons évoqué précédemment, il existe six types de drainage classant les sols des moins hydromorphes aux plus hydromorphes. La variable drainage est donc une variable ordinale. Certaines de ces catégories de drainage sont peu représentées au sein du bloc sud ; elles sont donc regroupées avec une catégorie voisine. La variable drainage de notre jeu de données comporte finalement quatre modalités. Nous disposons de 327 observations sur le bloc sud. Deux cents d'entre elles sont choisies aléatoirement pour estimer les paramètres du modèle hiérarchique spatial multivarié. Les 127 restantes sont utilisées pour mesurer la qualité des prédictions.

4.2 L'angélique (*Dicorynia guianensis* Amshoff)

Les propriétés technologiques de son bois et son abondance locale font de l'angélique une espèce d'intérêt économique majeure en Guyane. L'angélique représente à elle seule

34 % de la production de bois d'œuvre de la Guyane (Barret, 2001) ; c'est pourquoi elle a été choisie comme espèce modèle dans cette étude.

4.2.1 Caractérisation botanique de l'angélique

L'angélique est une espèce endémique du plateau des Guyanes qui s'étend du bassin versant de l'Orénoque au Venezuela à celui de l'Amazone au Brésil. C'est un grand arbre de la canopée pouvant atteindre 45 m de haut et dépasser 100 cm de diamètre (Jésel, 2005). Son tronc est cylindrique et bien conformé (Figure 4.3). L'angélique est capable de fleurir à partir de 18 cm de diamètre et de produire des graines viables à partir de 22 cm (diamètre minimum) (Caron et al., 1998). En moyenne, les arbres au-dessus de 25 cm de diamètre peuvent être considérés comme reproducteurs actifs produisant des graines viables (Jésel, 2005). La production de graines peut être très abondante (10 à 25 000 par arbre), mais une forte proportion de gousses (44 %) est stérile (Loubry, 1993). *Dicorynia guianensis* est une espèce anémochore ; les graines sont disséminées par le vent à de faibles distances des pieds-mères. A petite échelle, *Dicorynia guianensis* présente donc une distribution spatiale agrégée. Les agrégats regroupent en général une à plusieurs dizaines d'individus de diamètre supérieur à 10 cm. La distribution diamétrique des arbres composant chaque agrégat est variable (Jésel, 2005). Les agrégats, d'environ 50 m de diamètre, sont distants les uns des autres d'une centaine de mètres (Cabrera-Gaillard et Gignoux, 1989; Collinet, 1997). Des échanges de pollen peuvent s'opérer à longue distance entre les agrégats. L'espèce est fortement allogame[5] (taux d'allofécondation pour la population $\approx 91\,\%$) (Jésel, 2005).

4.2.2 Données d'inventaire du bloc sud

Depuis la mise en place des parcelles permanentes en 1984, tous les individus de diamètre supérieur à 10 cm dbh font l'objet d'un suivi. Les différents traitements sylvicoles ont été appliqués sur les parcelles entre 1986 et 1988. Les données d'inventaire dont nous disposons sur l'angélique ont été recueillies en 2002, après l'application des traitements. Les zones tampons et la zone intersticielle (Figure 4.2), non échantillonnées auparavant, ont également fait l'objet d'un inventaire. 337 arbres de diamètre supérieur à 10 cm ont été répertoriés sur le bloc sud (Figure 4.4).

Sur l'ensemble du dispositif, la densité moyenne des arbres de diamètre supérieur à 10 cm est de 6,9 arbres/ha (Caron et al., 1998). La figure 4.5 présente la distribution diamétrique des 337 arbres de dbh supérieur à 10 cm répertoriés sur le bloc sud. Le déficit d'individus dans les classes de grand diamètre (diamètre supérieur à 50 cm) est en partie dû aux traitements sylvicoles appliqués sur les parcelles 9, 10 et 12. Sur les 337 arbres répertoriés, seuls 155 ont un dbh supérieur ou égal à 25 cm (classes grisées de la figure 4.5) et sont considérés comme reproducteurs. Les jeunes arbres ayant un dbh inférieur à 10 cm ont également été inventoriés sur la zone d'étude, soit 1 028 individus. Pour définir le stade de juvéniles, nous nous basons sur les recommandations de Flores (2005). Le stade juvénile rassemble les individus des stades jeunes dont l'apparition est le plus susceptible d'avoir eu lieu après les traitements sylvicoles. La taille limite supérieure de ce stade correspond

[5]Se dit d'une plante dont la fécondation se fait par le pollen issu d'une autre plante.

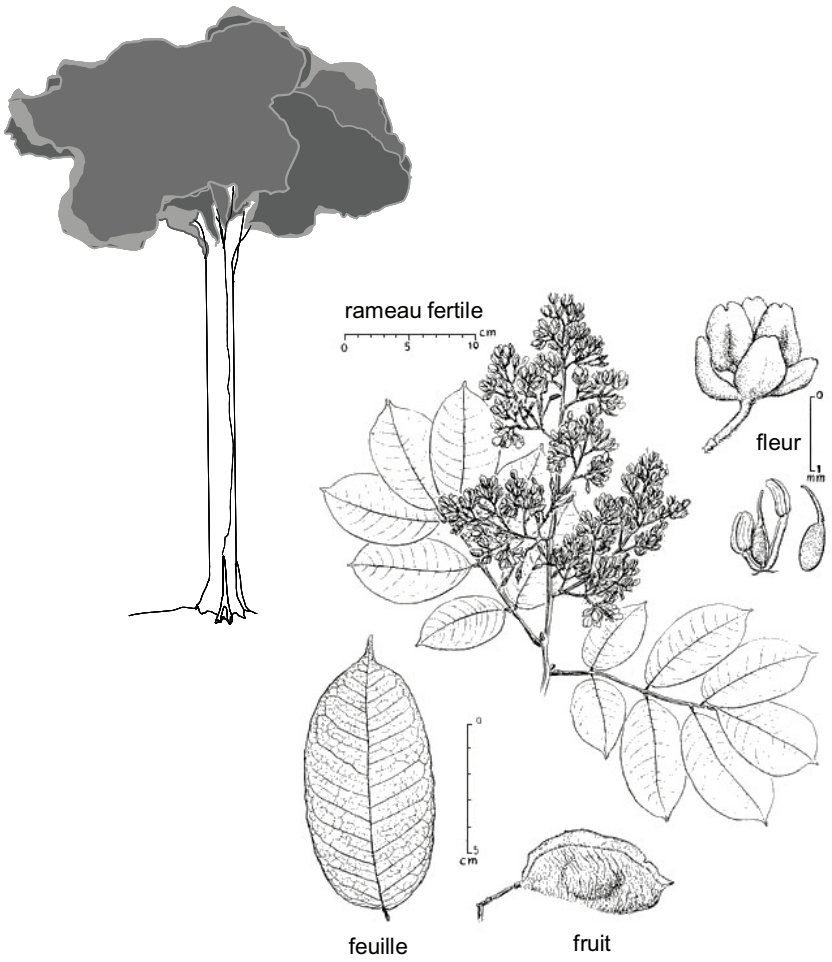


FIG. 4.3 – Schéma extrait de la thèse de Jésel (2005). Morphologie schématique de l'arbre et éléments de description botanique de *Dicorynia guianensis* Amshoff (Caesalpiniaceae).

à la taille qu'aurait atteint en 2002, au moment de l'inventaire, un individu moyen de dbh 1 cm en 1989 après le traitement. Pour l'angélique, cette taille limite est de 5 cm. Nous considérons donc comme juvéniles les jeunes arbres ayant un diamètre inférieur ou égal à

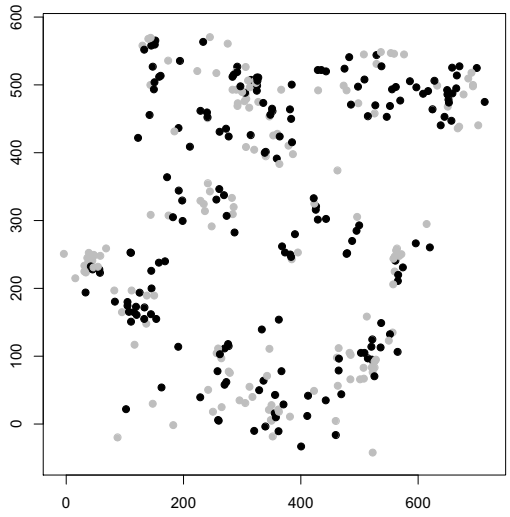


FIG. 4.4 – Carte représentant les positions des 337 arbres de diamètre supérieur à 10 cm inventoriés sur le bloc sud. Les arbres ayant un diamètre compris entre 10 et 25 cm sont représentés en noir et les arbres ayant un diamètre supérieur à 25 cm en gris.

5 cm.

4.2.3 Données génétiques

Plusieurs études (Caron et al., 2000; Latouche-Hallé et al., 2003) menées sur l'angélique ont contribué au recueil d'information génétique sur les individus du bloc sud. Les données génétiques dont nous disposons sont constituées de six marqueurs moléculaires. Cinq d'entre eux sont des marqueurs microsatellites. Un marqueur moléculaire est un fragment polymorphe de l'ADN qui renseigne sur le génotype de celui qui le porte. Un microsatellite est une proportion de la chaîne d'ADN constituée de répétitions de motifs composés de 1 à 4 nucléotides, de type AAAAA = $(A)_5$, GAGAGAGAGAGAGA = $(GA)_7$, $(CTT)_8$ où $(ACGT)_n$ où n varie de quelques unités à plusieurs dizaines. Si la mise en évidence de la variabilité repose sur la détection d'une telle suite de motifs, le marqueur est appelé marqueur microsatellite. Les microsatellites étant des régions très polymorphes de l'ADN, ils constituent donc des marqueurs génétiques très puissants (ONF, 2004). Le sixième marqueur disponible est un marqueur chloroplastique. Il s'agit d'un marqueur moléculaire qui met en évidence la variabilité au niveau des molécules d'ADN situées dans les chloroplastes. Le génome choloroplastique a la particularité d'être transmis par le parent femelle chez les feuillus et par le parent mâle chez les résineux. Chez l'angélique, le marqueur chloroplastique nous apporte donc de l'information sur la filiation maternelle. Les 155 adultes reproducteurs et environ 350 juvéniles ont été génotypés. Les problèmes liés au génotypage (ADN dégradé, mauvaise amplification, etc) font que seuls 144 arbres adultes et 270 juvéniles (Figure 4.6) présentent de l'information pour les six loci et peuvent être conservés

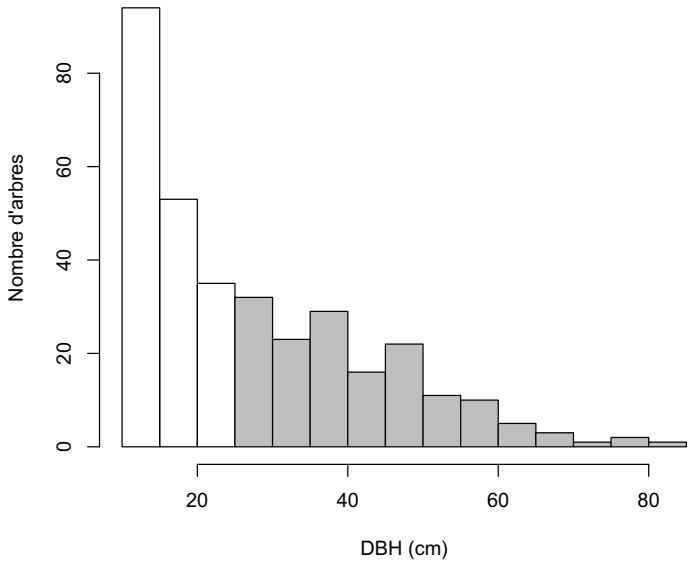


FIG. 4.5 – Histogramme de la distribution diamétrique des angéliques de dbh supérieur à 10 cm répertoriés sur le bloc sud du dispositif de Paracou. Les classes grisées correspondant aux arbres de plus de 25 cm de dbh considérés comme reproducteurs.

pour l'étude.

Le nombre d'allèles et le taux d'hétérozygotie observés par locus sont donnés dans le tableau 4.1. L'hétérozygotie observée est du même ordre de grandeur chez les adultes (0,36-0,78) que chez les juvéniles (0,39-0,72). Le nombre moyen de formes alléliques par locus est d'environ 9. Toutes les plantules n'ayant pas été génotypées, nous ne disposons que d'un échantillonnage partiel des juvéniles, excepté sur une zone de 1,8 ha au centre du bloc sud (Figure 4.6) où tous les juvéniles, soit 123 individus, ont été génotypés (Latouche-Hallé et al., 2003).

4.3 Résultats

4.3.1 Prédiction de l'environnement

Chaque couple de variables constituant le jeu de données trivarié est analysé séparément. Nous ne présentons ici que les résultats obtenus pour le jeu de données gaussien-ordinal composé de la pente et du drainage et le jeu de données ordinal-ordinal composé du drainage et de l'altitude. Comme pour les simulations, les fonctions moyennes mobiles utilisées sont proportionnelles au noyau gaussien, mais leur paramétrisation est différente.

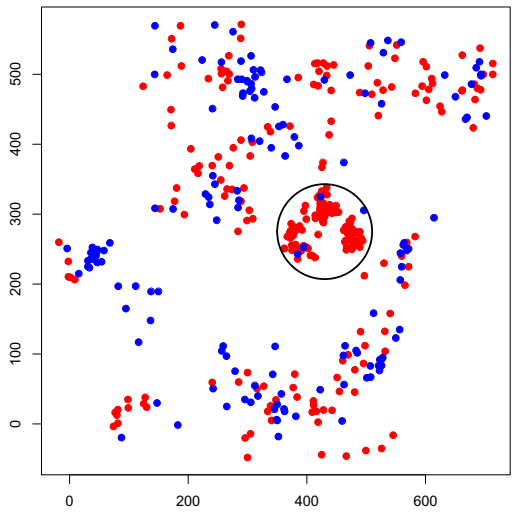


FIG. 4.6 – Carte des positions des 144 adultes (en bleu) et des 270 juvéniles (en rouge) génotypés. La zone d'échantillonnage exhaustif des juvéniles est matérialisée par un cercle.

TAB. 4.1 – Nombre d'allèles et taux d'hétérozygotie par locus. Le locus L_6 correspond au marqueur chloroplastique.

	Locus	**Nombre d'allèles par locus**	**Taux d'hétérozygotie observé**
Adultes	L_1	6	0,359
	L_2	10	0,690
	L_3	5	0,768
	L_4	13	0,782
	L_5	6	0,676
	L_6	3	-
Juvéniles	L_1	7	0,393
	L_2	9	0,660
	L_3	8	0,718
	L_4	14	0,698
	L_5	5	0,653
	L_6	3	-

Les paramètres ϕ_k et σ_k sont remplacés respectivement par $\phi_k^2/2$ et $\phi_k^{-1}\sqrt{4\sigma_k/\pi}$. Cette paramétrisation est préférée à celle des fonctions moyennes mobiles utilisées pour les si-

mulations pour conserver des paramètres ayant le même ordre de grandeur que ceux des simulations. La matrice de covariance correspondante a la forme analytique suivante :

$$\begin{aligned}\mathrm{Cov}[S_k(\mathbf{s}_i), S_k(\mathbf{s}_j)] &= \sigma_k \exp\left(-\frac{\|\mathbf{s}_j - \mathbf{s}_j\|^2}{\phi_k^2}\right),\\ \mathrm{Cov}[S_k(\mathbf{s}_i), S_m(\mathbf{s}_j)] &= \frac{2\rho_k\rho_m\sqrt{\sigma_k\sigma_m}\phi_k\phi_m}{\phi_k^2+\phi_m^2}\exp\left(-\frac{2\|\mathbf{s}_j - \mathbf{s}_j\|^2}{\phi_k^2+\phi_m^2}\right).\end{aligned}$$

Les estimations des paramètres obtenues pour le jeu de données constitué par la pente et le drainage sont présentées dans le tableau 4.2. Le tableau 4.3 présente celles obtenues pour le jeu de données composé du drainage et de l'altitude.

TAB. 4.2 – Estimations des paramètres obtenues à partir du jeu de données gaussien-ordinal composé de la pente Y_1 et du drainage Y_2.

Paramètres	**Estimations bivariées**	**Écart-type**
σ_1	17,41	(3,38)
ϕ_1	30,44	(4,24)
ν_1	2,59	(0,42)
μ_1	10,58	(0,49)
σ_2	3,08	(1,12)
ϕ_2	76,73	(11,73)
$\alpha_{2,2}$	1,67	(0,22)
$\alpha_{2,3}$	4,39	(0,41)
μ_2	2,88	(0.37)
ρ_{12}	0,16	(0,21)

Comme nous l'avions constaté pour les simulations, la vitesse de convergence des paramètres associés à la variable gaussienne est plus élevée que celle associée aux variables ordinales. Les estimations des paramètres pour la variable gaussienne sont cohérentes avec la portée, le palier et l'effet de pépite observés sur le variogramme empirique. Pour le premier jeu de données, l'estimation du paramètre μ_1 est proche de la moyenne du vecteur $\mathbf{Y}_1$. Les variables pente et drainage ne sont pas corrélées. Les estimations des paramètres relatifs au drainage obtenues à partir du second jeu de données (tableau 4.3, variable 1) sont cohérentes avec celles obtenus à partir du premier jeu de données (tableau 4.2, variable 2). Les écarts-types des estimations des paramètres σ_2 et ρ_{12} obtenues à partir du jeu de données gaussien-ordinal sont élevées. Il en est de même pour celui de l'estimation du paramètre σ_1 obtenue avec le jeu de données ordinal-ordinal. Comme nous pouvions nous y attendre, la corrélation spatiale entre le drainage du sol et l'altitude est négative. De manière générale, les estimations obtenues à partir des données réelles sont moins précises que celles obtenues à partir des jeux de données simulés. Les estimations pourraient être améliorées en augmentant la taille du jeu de données utilisé pour l'inférence. Rappelons

TAB. 4.3 – Estimations des paramètres obtenues à partir du jeu de données ordinal-ordinal composé du drainage Y_1 et de l'altitude Y_2.

Paramètres	**Estimations bivariées**	**Écart-type**
σ_1	4,44	(1,89)
ϕ_1	57,82	(9,50)
$\alpha_{1,2}$	1,87	(0,29)
$\alpha_{1,3}$	5,26	(0,71)
μ_1	3,40	(0,53)
σ_2	221,87	(59,51)
ϕ_2	76,04	(4,38)
$\alpha_{2,2}$	19,97	(3,58)
μ_2	10,11	(2,85)
ρ_{12}	$-0,80$	(0,07)

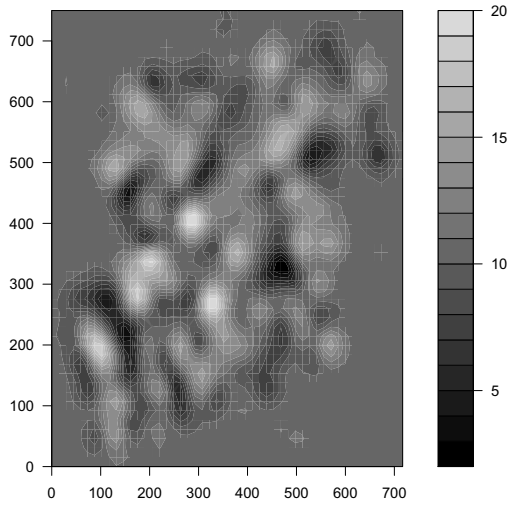
qu'ici seuls les deux tiers des données disponibles sont utilisées pour l'estimation ; le reste étant utilisé pour la validation.

La figure 4.7 présente les cartes de prédiction obtenues à partir des deux jeux de données. Les cartes du drainage et de l'altitude sont cohérentes avec les connaissances que l'on a de la zone étudiée. La corrélation négative entre le drainage et l'altitude se traduit sur les cartes de prédiction par une correspondance entre les zones de faible altitude et les bas-fonds. Le tableau 4.4 résume les différents critères utilisés pour évaluer la qualité des prédictions. Les prédictions de la variable gaussienne (pente) sont peu biaisées. La RM-

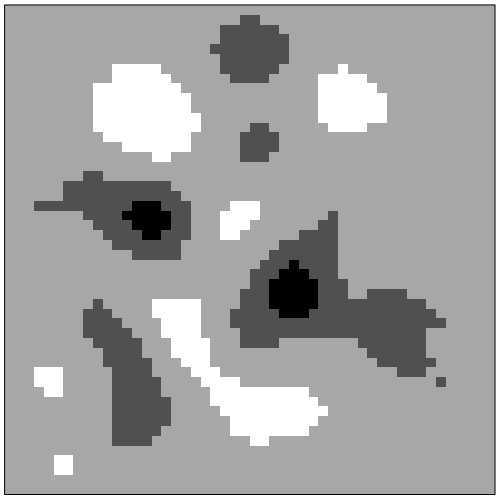
TAB. 4.4 – Critères de validation permettant de mesurer la qualité des prédictions obtenues à partir du jeu de données composé de la pente et du drainage et du jeu de données composé du drainage et de l'altitude. Le biais, la RMSPE, la RMEV et l'intervalle de couverture 80%PI sont donnés pour les variables gaussiennes. Le pourcentage de valeurs correctement prédites est indiqué pour chaque variable ordinale.

Jeu de données	**Variable 1**		**Variable 2**	
gaussien-ordinal	biais	$-0,68$		
	RMSPE	4,91	%CP	67,7
	RMEV	4,27		
	80%PI	0,80		
ordinal-ordinal	%CP	66,9	%CP	78,7

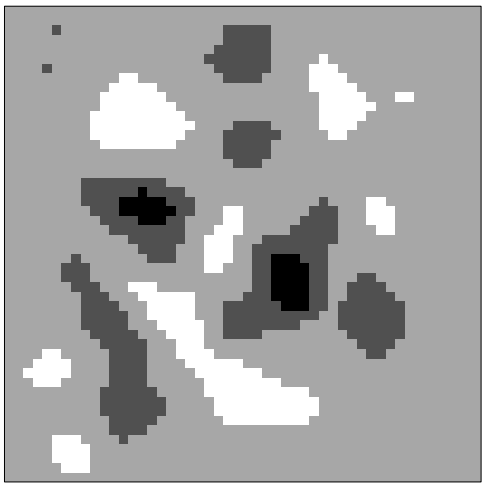
SEP est un peu plus élevée que la RMEV ; la variance de prédiction est donc estimée assez



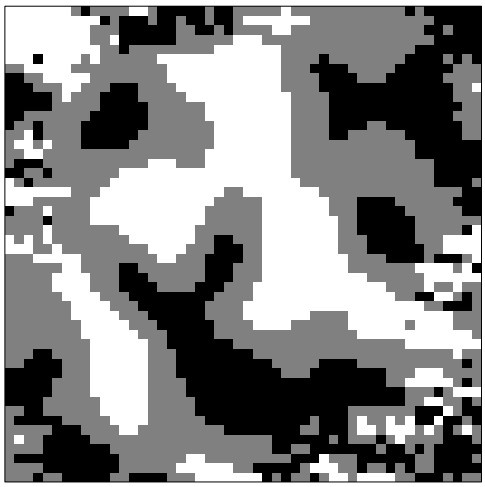
Pente



Drainage



Drainage



Altitude

FIG. 4.7 – Cartes de prédiction obtenues à partir du jeu de données composé de la pente et du drainage (en haut) et du jeu de données composé du drainage et de l'altitude (en bas). La zone d'étude est un rectangle de 717 mètres sur 750. La pente est mesurée en degrés. Les modalités du drainage sont représentées suivant une échelle de gris allant du noir correspondant aux sols bien drainés au blanc correspondant aux sols hydromorphes. Les modalités de l'altitude sont représentées suivant une échelle de gris allant du noir correspondant aux altitudes faibles au blanc correspondant aux altitudes élevées.

précisément. Le pourcentage de valeurs correctement prédites pour la variable ordinale du premier jeu de données est d'environ 68 %. Les points pour lesquels la prédiction n'est pas correcte sont la plupart du temps situés sur les bords des parcelles ou dans des zones où

peu de mesures ont été effectuées (Figure 4.8). Ces points correspondent aux points où les variances de prédiction des estimations des paramètres relatifs à la variable gaussienne sont élevées. Le nombre d'observations par modalité pour la variable ordinale est déséquilibré. Le manque d'information concernant certaines modalités peut expliquer certaines erreurs dans les prédictions. On constate le même phénomène pour les prédictions concernant le drainage obtenues à partir du second jeu de données, où seulement 67 % des valeurs sont correctement prédites. Pour l'altitude, 79 % des prédictions effectuées sont correctes. Pour le jeu de données composé de la pente et du drainage, les résultats de validation pour la variable ordinale sont assez décevants par rapport à ceux obtenus avec le jeu de données simulées gaussien-ordinal où les points étaient choisis aléatoirement. Une des raisons pouvant expliquer la diminution de la qualité des prédictions pourrait être la disposition spécifique des points du jeu de données réelles (Figure 4.8). Les mesures ayant été effectuées sur les parcelles permanentes du dispositif, les points présentent une structure agrégée et des simulations ont montré que la qualité des prédictions diminue quand les points de mesure sont agrégés. Il est, en général, préférable d'avoir échantillonné des points aléatoirement sur l'ensemble de la zone d'étude pour réduire le biais et la variance des prédictions. Une faible structuration spatiale des données pourrait aussi expliquer la moindre qualité des prédictions sur le jeu de données réelles par rapport à celles obtenues sur les simulations.

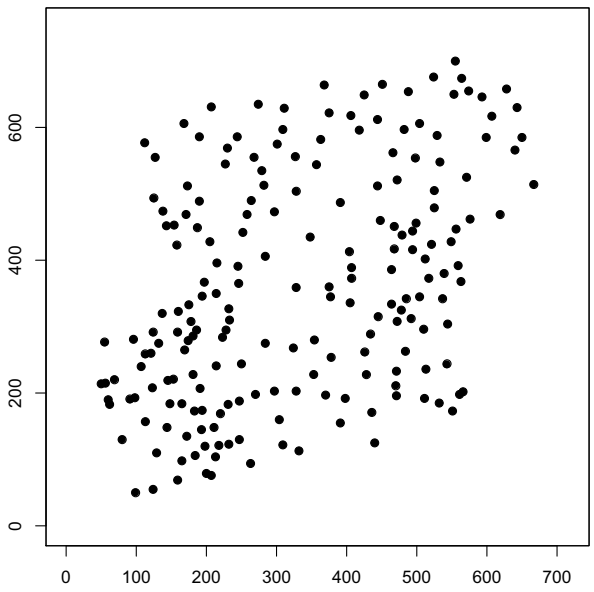


FIG. 4.8 – Localisation des 200 points d'observation utilisés pour l'estimation des paramètres

4.3.2 Modélisation de la régénération

Plusieurs modèles de régénération (avec ou sans survie) ont été testés sur les données concernant l'angélique. L'information génétique est constituée des cinq marqueurs micro-

satellites. Le succès reproducteur de chaque arbre adulte est modélisé comme une fonction exponentielle du diamètre. La répartition spatiale des adultes peut être considérée comme la réalisation d'un processus ponctuel. Ne disposant que d'une réalisation de ce processus, nous ne sommes pas en mesure de savoir si ce processus est hétérogène ou non (Goreaud, 2000).

• Modèle sans survie

Nous présentons d'abord les résultats obtenus avec le modèle de Shimatani sans survie (équation 3.5). Le modèle sans survie ne nécessite pas d'avoir échantillonné les juvéniles de manière exhaustive. Nous utilisons donc comme données l'ensemble des juvéniles génotypés, soit 270 individus. Autrement dit, les zones A et B sont considérées comme identiques.

Les paramètres à estimer sont le paramètre b intervenant dans la définition du succès reproducteur, les variances τ_1^2 et τ_2^2 des noyaux de dispersion du pollen et des graines et la densité $\overline{d}$. Lors de la procédure d'optimisation de la vraisemblance, la densité d'adultes reproducteurs $\overline{d}$ décroît jusqu'à atteindre la borne inférieure fixée par l'utilisateur (cette borne est fixée à moins d'un individu adulte à l'hectare) entraînant une augmentation du paramètre τ_1. Les données dont nous disposons ne permettent pas d'estimer de manière satisfaisante l'ensemble des paramètres. Les estimations obtenues laissent présager un manque d'information génétique (voir paragraphe 3.4.2). Des résultats similaires sont obtenus si l'information génétique est constituée des cinq marqueurs microsatellites et du marqueur chloroplastique. Pour pouvoir estimer les paramètres, nous allons fixer l'intensité $\overline{d}$. Pour cela, nous faisons l'hypothèse que la population d'angéliques observée sur le bloc sud est représentative de la population d'angéliques toute entière. Environ 46 % des arbres du bloc sud ayant un diamètre supérieur à 10 cm dbh peuvent être considérés comme reproducteurs (dbh supérieur à 25 cm) (Figure 4.5) et la densité moyenne d'arbres dont le dbh est supérieur à 10 cm, à Paracou, est de 6,9 individus par hectare. Nous pouvons donc considérer que, pour l'angélique, la densité d'adultes reproducteurs est d'environ de 3 individus à l'hectare, valeur que nous utilisons dans la suite pour les procédures d'estimation. Les estimations obtenues sont alors les suivantes :

$$\begin{aligned}\hat{b} &= 0,045\,(\mathrm{cm}^{-1}),\\ \hat{\tau_1} &= 998,44\,(\mathrm{m}),\\ \hat{\tau_2} &= 100,82\,(\mathrm{m}).\end{aligned}$$

La distance de dispersion moyenne du pollen est de 1 250 m et celle des graines de 126 m. Chaque adulte reproducteur de 42 cm de diamètre[6] est à l'origine de l'installation d'environ 7,5 juvéniles dans le peuplement.

Si on se réfère à l'article de Latouche-Hallé et al. (2004), les flux de pollen à longue distance pour l'angélique peuvent atteindre le kilomètre. Ici, la distance moyenne de dispersion du pollen estimée (1 250 m) semble beaucoup trop importante. De même, la distance de dispersion moyenne des graines estimée est supérieure à celle normalement admise (distance de dispersion moyenne de 30 m, distance de dispersion maximale de 60 m (Jésel, 2005)). Plusieurs hypothèses peuvent être envisagées pour expliquer l'incohérence entre les

[6]Cette valeur correspond au diamètre moyen des adultes reproducteurs situés dans la zone B.

estimations et les connaissances dont nous disposons sur l'angélique. En premier lieu vient le manque d'information génétique. Comme nous l'avons montré grâce aux simulations, l'utilisation de cinq marqueurs génétiques s'avère en général insuffisante pour estimer les paramètres de manière satisfaisante. Le nombre de couples potentiels pouvant expliquer l'apparition d'un juvénile est alors plus important que lorsque l'information génétique est suffisamment discriminante. Ce manque d'information entraîne donc une sur-estimation de τ_1 et de τ_2. De plus, des traitements sylvicoles ont été appliqués à trois des quatre parcelles. Une partie des adultes reproducteurs a été exploitée lors des traitements. Des graines issues de ces adultes ont pu donner naissance à des juvéniles après les traitements. Des juvéniles ayant des parents déjà morts ont donc pu être échantillonnés. L'absence de ces adultes entraîne une sur-estimation des paramètres de dispersion. En effet, si aucun adulte de la zone B n'explique le génotype du juvénile, on considère que ce dernier est issu de parents situés hors de B, augmentant ainsi les distances de dispersion des graines et du pollen. Enfin, la zone A étant identique à la zone B, la proportion de juvéniles ne trouvant pas de parents compatibles dans B a tendance à être supérieure à celle obtenue quand la zone A est strictement incluse dans la zone B, ce qui entraîne également une augmentation des distances moyennes de dispersion.

• **Modèle avec survie**

Bien que le manque d'information génétique et les perturbations liées aux traitements sylvicoles ne permettent pas d'estimer les paramètres de dispersion de manière satisfaisante, nous avons appliqué le modèle avec survie aux données sur l'angélique. La zone B est identique à celle utilisée dans le modèle sans survie ; elle comprend 144 adultes reproducteurs. La zone A correspond ici à la zone d'échantillonnage exhaustive des juvéniles matérialisée par un cercle sur la figure 4.6. Cette zone compte 123 juvéniles. Les variables environnementales choisies sont la pente, l'altitude et le drainage. L'altitude est considérée non pas comme une variable ordinale comme dans le paragraphe 4.3.1, mais comme une variable gaussienne. Les résultats obtenus sont présentés dans le tableau 4.5. L'environnement est prédit à partir d'un échantillon de 200 points.

Environnement	b (cm^{-1})	τ_1 (m)	τ_2 (m)	$\overline{d}$ ($\mathrm{ind.m}^{-2}$)	δ	$\boldsymbol{\gamma}$
Pente	0,069	638,99	152,72	fixée	$-4,366$	0,740
Altitude	0,063	643,21	147,31	fixée	$-4,493$	0,988
Drainage	0,080	625,40	164,82	fixée	fixée	$\gamma_1 = -1,772$ $\gamma_2 = 0,432$ $\gamma_3 = -0,861$ $\gamma_4 = -16,202$

TAB. 4.5 – Estimation des paramètres du processus ponctuel modélisant la répartition spatiale des juvéniles. La variable environnementale utilisée dans le terme de survie est indiquée à gauche.

Quelle que soit la variable environnementale utilisée, l'estimation des paramètres b, τ_1 et τ_2 est stable. La distance moyenne de dispersion estimée du pollen est comprise entre 783 et 806 m et celle des graines entre 184 et 206 m. Le nombre de juvéniles installés issus d'un adulte de 42 cm varie entre 14 et 28. Ces estimations sont cohérentes avec les estimations obtenues pour un modèle sans survie appliqué sur les mêmes données. La distance moyenne de dispersion du pollen est plus faible dans le modèle avec survie où l'on ne considère que les juvéniles situés dans la zone d'échantillonnage exhaustif que dans le modèle sans survie où l'on considère l'ensemble des juvéniles génotypés. Cela est dû au fait que la plupart des juvéniles appartenant à la zone d'échantillonnage exhaustif ont un père potentiel dans B. La distance de dispersion des graines estimée dans le modèle avec survie est, quant à elle, légèrement plus élevée que celle estimée dans le modèle précédent. Pour le modèle où la pente est choisie comme variable environnementale, les paramètres δ et $\boldsymbol{\gamma}$ sont stables. Le taux de survie des juvéniles augmente fortement lorsque la pente dépasse 5 degrés et se stabilise à un palier maximal lorsque la pente atteint 10 degrés. Lorsque le modèle fait intervenir l'altitude comme variable explicative de la survie, les paramètres δ et $\boldsymbol{\gamma}$ sont instables. Cela semble dû au fait que l'altitude est peu contrastée sur la zone d'échantillonnage exhaustif. L'altitude minimale sur la zone est de 20 m ; les altitudes faibles ne sont pas représentées. 50 % de la zone a une altitude comprise entre 27 et 35 m. Le taux de survie des juvéniles est quasiment constant pour toutes les altitudes de la zone. Une mise en évidence d'un effet de l'altitude sur la régénération des juvéniles n'est donc pas possible à partir des données dont nous disposons. En ce qui concerne le drainage, la modalité 2 est la plus favorable au développement des juvéniles, les sols étant classés des sols bien drainés aux sols hydromorphes. La probabilité de survie d'un juvéniles sur un sol hydromorphe, c'est-à-dire dans les bas-fonds, est quasi nulle. Ces résultats sont en accord avec la littérature (Kokou, 1994; Leroy, 2000).

4.4 Discussion

Les prédictions effectuées à partir des données environnementales recueillies sur la partie sud du dispositif de Paracou en utilisant le modèle hiérarchique spatial multivarié s'avèrent cohérentes avec les connaissances dont nous disposons sur le site.

Les prédictions obtenues à partir du jeu de données réelles sont de moins bonne qualité que celles obtenues à partir des jeux de données simulées. Le nombre d'observations utilisées pour estimer les paramètres du modèle est sans doute insuffisant. La différence entre les deux applications pourrait être due à la disposition spécifique des points de mesure du jeu de données réelles. Comme les mesures ont été effectuées sur les parcelles permanentes du dispositif, les points sont agrégés alors qu'ils ont été choisis aléatoirement pour les jeux de données simulés. Il en résulte une faible représentation de la structure spatiale et des effets de bord plus importants. La différence de qualité des résultats entre jeux de données simulées et jeux de données réelles peut aussi s'expliquer par le choix des fonctions moyennes mobiles. Contrairement aux simulations où la forme des fonctions moyennes mobiles est connue, la forme des fonctions moyennes mobiles est ici choisie par l'utilisateur. Ce choix est crucial puisque la forme et la souplesse du variogramme qui traduit la dé-

pendance entre les variables en dépendent. Le choix que nous avons opéré qui consiste à travailler avec des fonctions moyennes mobiles gaussiennes n'est peut être pas le plus approprié. Il aurait été souhaitable de tester d'autres formes de fonctions moyennes mobiles pour voir si cela améliorait la qualité des prédictions. Cela n'a pas été fait par manque de temps.

Le modèle de régénération proposé permet de prendre en compte les données génétiques ainsi que les effets de l'environnement sur la survie des juvéniles. Son application aux données de régénération sur l'angélique n'a pas abouti à une estimation satisfaisante des paramètres traduisant les mécanismes de dispersion, les distances de dispersion des graines et du pollen estimées étant supérieures aux valeurs communément admises. Les deux hypothèses les plus vraisemblables pour expliquer cette sur-estimation sont le manque d'information génétique et l'effet des traitements sylvicoles ayant entraîné l'exploitation ou la dévitalisation d'un certain nombre d'adultes reproducteurs pouvant potentiellement être les parents des juvéniles échantillonnés. Pour confirmer ces hypothèses, le modèle de régénération a été appliqué à une seconde espèce, le wacapou (*Vouacapoua americana*), pour laquelle nous disposons de données de régénération sur une zone non perturbée (parcelle 16 du dispositif de Paracou) et de plus d'information génétique (neuf marqueurs microstallites). Les paramètres de dispersion estimés se sont alors révélés être plus cohérents avec les valeurs de la littérature que dans le cas de l'angélique. Notons que cette espèce n'avait pas été initialement choisie, car la zone où elle a été échantillonnée n'a pas fait l'objet de mesures environnementales aussi complètes que les quatre parcelles du bloc sud. La mise en œuvre du modèle avec survie nécessite donc non seulement d'avoir échantillonné des juvéniles de manière exhaustive sur une zone non perturbée ayant un environnement suffisamment hétérogène pour mettre en évidence les effets des différentes variables environnementales, mais aussi d'avoir génotypé ces individus sur un nombre suffisant de loci pour avoir une probabilité plus élevée de déterminer leurs véritables parents. La présence parmi les données génétiques de marqueurs, comme les marqueurs choloroplastiques, permettant de discriminer le père et la mère dans un couple potentiel de parents peut contribuer à l'amélioration des estimations des distances de dispersion.

L'estimation de la densité d'adultes reproducteurs $\overline{d}$ n'a pas été possible sur les données concernant l'angélique étant donné le manque d'information génétique, ou sur le jeu de données sur le wacapou. Ce problème reste en suspens. La mise en œuvre du modèle est donc, pour le moment, conditionnée à la connaissance de la densité des adultes reproducteurs de l'espèce étudiée.

La forme de la fonction de survie utilisée n'est pas toujours appropriée. Par exemple, lorsque le modèle ne comprend qu'une variable environnementale Y continue, la fonction de survie ne permet pas de mettre en évidence un effet favorable sur la régénération des valeurs intermédiaires de Y par rapport aux petites et aux grandes valeurs de Y, la fonction de survie étant monotone en Y. Une réflexion plus approfondie sur la fonction de survie doit donc être envisagée.

Le principal avantage du modèle de régénération proposé est qu'il permet d'estimer en une fois et d'une façon cohérente et intégrée un ensemble de paramètres écologiques (distance de dispersion, dépendance de la survie des juvéniles à l'environnement, patrons de répartition spatiale) qui sont difficiles à estimer et que les écologues estiment en général

séparément. Le prix à payer pour pouvoir estimer simultanément et de manière satisfaisante tous ces paramètres est de disposer d'une quantité suffisante de données.

Conclusions et perspectives

L'objectif de ce travail de thèse était de prédire la répartition spatiale et la diversité génétique des juvéniles à grande échelle à partir d'un échantillonnage raisonnable des adultes, des juvéniles et de l'environnement. Ce travail a été divisé en deux parties : la prédiction de l'environnement et la modélisation de la régénération.

La principale difficulté soulevée par la prédiction de l'environnement était de savoir comment prédire un champ aléatoire composé de variable de nature différente. Le problème a en partie été résolu par la mise en œuvre d'un modèle hiérarchique spatiale multivarié permettant de prédire simultanément des variables gaussiennes, des variables de Poisson et des variables ordinales. Ce modèle est une généralisation au cas multivarié des modèles linéaires généralisés spatiaux. L'écriture d'un modèle linéaire généralisé spatial pour chaque variable environnementale permet d'associer à chacune d'elle sa composante spatiale. Les variables étant de nature différente, il n'est pas possible de modéliser leur dépendance de manière directe. Leur dépendance est donc modélisée au travers de leurs composantes spatiales qui, elles, sont toutes continues. De plus, la structure hiérarchique du modèle facilite la mise en œuvre de la procédure d'estimation des paramètres par les méthodes MCMC.

Bien que les modèles de covariance classiques puissent être employés, le choix opéré pour modéliser la dépendance spatiale entre les variables s'est finalement porté sur une matrice de covariance construite à partir de la méthode moyenne mobile. Cette méthode offre l'avantage d'être très souple grâce au large choix de fonctions moyennes mobiles possibles. Cette approche laisse entrevoir la possibilité de modéliser des structures de dépendance plus complexes (en modifiant le support des fonctions moyennes mobiles ou le processus sous-jacent) et de prendre en compte la dépendance à différentes échelles (par sommation de plusieurs intégrales construites à partir de différentes fonctions moyennes mobiles). Cette grande variété de fonctions moyennes mobiles peut se révéler être un inconvénient étant donné qu'aucune règle n'est donnée pour choisir ces fonctions. Le choix est principalement guidé par l'observation du variogramme empirique, par la présence ou non d'anisotropie dans les données et par le principe de parcimonie. Le choix de fonctions moyennes mobiles proportionnelles au noyau gaussien pour prédire l'environnement du bloc sud peut d'ailleurs être critiqué. D'autres fonctions moyennes mobiles devraient être testées pour voir si elles améliorent la qualité des prédictions. Une étude plus approfondie des correspondances entre forme des fonctions moyennes mobiles et forme des variogrammes permettrait d'envisager d'édicter des règles concernant le choix des fonctions moyennes mobiles.

D'un point de vue pratique, la procédure d'estimation des paramètres du modèle hiérarchique spatial multivarié est longue bien qu'elle soit implémentée en langage C. Le temps de calcul augmente avec le nombre de variables étudiées et le nombre de sites échantillonnés. Le traitement de variables nominales, bien que tout à fait envisageable d'un point de vue théorique, n'est pas opérationnel d'un point de vue pratique car il entraîne une augmentation importante de la taille des matrices manipulées dans la procédure d'estimation, et donc une augmentation du temps de calcul. Un travail algorithmique serait nécessaire pour optimiser le code et rendre la procédure d'estimation plus rapide quelle que soit la nature des variables composant le champ aléatoire représentant l'environnement.

Le modèle a montré une bonne capacité à prédire l'environnement sur un jeu de données réelles. Il faudrait envisager de comparer les résultats obtenus grâce au modèle hiérarchique spatial multivarié avec ceux obtenus par d'autres méthodes comme le cokrigeage disjonctif ou l'approche BME. La mise en œuvre du modèle hiérarchique spatial multivarié n'offre un réel intérêt que si les variables composant le champ aléatoire multivarié sont spatialement corrélées. Dans ce cas, les simulations ont montré que la qualité des prédictions obtenues par une procédure d'estimation bivariée où la dépendance entre les variables est prise en compte est supérieure ou égale à celle des prédictions obtenues par une procédure d'estimation univariée. Travailler dans un cadre bivarié améliore notamment la qualité des prédictions des variables ordinales. Effectuer une prédiction sur un plus grand nombre de variables ne garantit pas forcément d'améliorer significativement la qualité des prédictions et augmente le temps de calcul.

Un des objectifs de ce travail de thèse était de prédire l'environnement à grande échelle. Ici, l'environnement a été prédit sur une superficie d'une quarantaine d'hectares. Étendre la prédiction de l'environnement à une zone plus importante, par exemple à l'échelle du dispositif tout entier, nécessiterait d'effectuer des mesures des variables environnementales entre les parcelles pour assurer la qualité des prédictions, et donc d'avoir des moyens informatiques plus importants pour pouvoir mettre en œuvre l'algorithme d'estimation. Il serait intéressant de pourvoir déterminer à quelle vitesse la qualité des prédictions se dégrade en fonction du nombre et de la répartition spatiale des points échantillonnés afin de savoir combien d'observations il est nécessaire d'effectuer sur de nouvelles zones pour assurer une prédiction correcte de l'environnement.

La seconde partie de ce travail de thèse a été consacrée à la modélisation de la régénération, l'objectif étant de prédire la répartition spatiale et la diversité génétique des juvéniles connaissant l'environnement. Le modèle que nous avons proposé est novateur dans le sens où il permet à la fois d'intégrer de l'information génétique et d'étudier l'effet de l'environnement sur la régénération. Ce modèle est un processus de Poisson hétérogène multivarié dont la fonction d'intensité comporte un terme modélisant la survie des juvéniles en fonction de l'environnement. La forme de la fonction de survie actuellement utilisée n'est pas entièrement satisfaisante. Cette fonction étant monotone, certains effets des variables environnementales sur la régénération ne peuvent pas être mis en évidence. Il faudra donc envisager de modifier cette fonction de survie, en la remplaçant, par exemple, par une fonction sinusoïdale ou par une fonction comprise entre 0 et 1 qui présente un optimum. Dans le modèle proposé, la survie ne dépend que de l'environnement. Or l'environnement n'est

sûrement pas le seul à avoir une influence sur la survie des juvéniles. Nous pourrions, par exemple, considérer une fonction de survie qui dépende non seulement de l'environnement, mais aussi du génotype, voire des interactions génotype-environnement. Les phénomènes de compétition intra et inter spécifique ont été occultés alors qu'ils jouent pourtant un rôle important dans la régénération. Nous pourrions envisager de traduire la compétition sous forme d'une variable qui interviendrait dans la fonction de survie au même titre qu'une variable environnementale ou de modifier la nature du processus ponctuel pour qu'il permette de modéliser une certaine régularité entre les individus.

Bien que la procédure d'estimation ait été validée par des simulations, nous n'avons pas obtenu les résultats escomptés sur les données de régénération concernant l'angélique : difficulté pour estimer la densité d'adultes reproducteurs, sur-estimation des paramètres de dispersion. Plusieurs explications peuvent être avancées pour expliquer les problèmes rencontrés. La forme des noyaux de dispersion utilisés est peut être mal appropriée. Il faudrait tester d'autres noyaux pour voir si l'on obtient des résultats similaires. Les résultats d'estimation couplés aux diverses études de simulation réalisées laissent plutôt présager un manque d'information génétique. Les données de génotypage disponibles ne sont pas suffisamment discriminantes pour déterminer avec une probabilité élevée les parents potentiels de chaque juvénile observé. De plus, l'absence de certains adultes reproducteurs due aux traitements sylvicoles peut elle aussi être à l'origine de la sur-estimation des paramètres de dispersion. Ces deux dernières hypothèses semblent être confirmées par l'étude complémentaire effectuée sur la régénération du wacapou. Si de nouvelles études sur la régénération doivent être envisagées, il faudra échantillonner des juvéniles et des adultes reproducteurs sur une zone non perturbée, et disposer de données de génotypage suffisamment riches (nombre de loci, nombre d'allèles par locus) pour assurer des résultats d'estimation satisfaisants. La zone d'échantillonnage exhaustif des juvéniles devra également présenter un environnement contrasté si l'objectif est de mettre en évidence l'influence du milieu sur la régénération. Si la zone d'échantillonnage s'avère perturbée, le modèle de régénération devra être adapté pour prendre en compte les adultes déjà morts. Notons que pour estimer la survie de manière plus exacte, il faudrait pouvoir intégrer aux données les juvéniles situés dans la zone d'échantillonnage exhaustif qui présentent des données manquantes dans leur génotypage et qui pour le moment ont été écartés de l'étude. Dans le cas particulier de l'angélique, un échantillonnage exhaustif de tous les arbres ayant un diamètre compris entre 1 et 10 cm, soit environ un millier d'individus, a été réalisé. Seules quelques centaines d'entre eux ont été génotypés, dont une moitié environ est située sur la zone d'échantillonnage exhaustif A. Seuls ces derniers ont été pris en compte pour l'estimation des paramètres du modèle. Les autres juvéniles géoréférencés situés hors de A pourraient, dans ce cas, apporter une information complémentaire pour estimer la survie et mettre en évidence les effets de l'environnement. Étant donné le temps nécessaire pour effectuer un tel échantillonnage, il serait souhaitable de réfléchir à une éventuelle prise en compte de cette information dans le modèle.

Le problème de la prise en compte de l'impact des erreurs liées à la prédiction de l'environnement sur l'estimation des paramètres caractérisant la régénération n'a été qu'effleuré et fera l'objet de futurs travaux. Plusieurs pistes de travail sont envisagées. La première

consisterait à traiter globalement la prédiction de l'environnement et la modélisation de la régénération grâce à un modèle hiérarchique. Cette structure permettrait de prendre directement en compte l'erreur de prédiction dans l'estimation des paramètres du processus ponctuel. Une seconde piste consisterait à déterminer l'effort d'échantillonnage nécessaire pour que la qualité des prédictions de l'environnement nous assure de prédire la régénération de manière satisfaisante.

Les données dont nous disposons pour cette étude ont été recueillie à une date t ; nous ne bénéficions pas de mesures répétées dans le temps. Travailler à une date fixée ne pose pas de problème en ce qui concerne la prédiction de l'environnement. Ce dernier variant très lentement, il peut être considéré comme constant dans le temps. Cela est beaucoup plus problématique pour la modélisation de la régénération qui, elle, est constituée d'une succession de phénomènes biologiques qui s'étalent dans le temps. Nous l'avons, par exemple, constaté avec l'absence des adultes liés aux traitements sylvicoles. Il faudrait donc à terme introduire une dimension temporelle à l'étude afin d'améliorer la prédiction de la régénération, mais cela nécessite d'effectuer de nouvelles campagnes de mesures. Il serait alors possible d'envisager de coupler modèle de régénération et modèle de dynamique des populations.

Annexe A

Prédiction de variables nominales

Le modèle tel qu'il est décrit au paragraphe 1.3.2 ne permet de traiter que des variables gaussiennes, des variables de Poisson et des variables ordinales. Il est possible d'envisager de traiter des variables nominales en généralisant le modèle probit multinomial (Daganzo, 1979; Natarajan et al., 2000) au cas spatial, de la même manière que nous avons généralisé le modèle probit ordinal multivarié pour traiter des champs spatiaux ordinaux.

Définition 14 Modèle probit multinomial
On considère une variable nominale Y à L modalités. Soit Y_i la $i^{ème}$ observation. Cette $i^{ème}$ observation peut être représentée par un vecteur $\mathbf{w}_i = (w_{i1}, \ldots, w_{iL})'$. Chaque composante de $\mathbf{w}_i$ est binaire : $w_{il} = 1$ si la $i^{ème}$ observation est égale à l et $w_{il} = 0$ sinon. De plus, les composantes du vecteur $\mathbf{w}_i$ vérifient $\sum_{l=1}^{L} w_{il} = 1$. Il existe un vecteur de variables latentes continues qui engendrent les $\mathbf{w}_i$:

$$w_{il} = \mathbb{1}_{\{u_{il} = \max_{k=1,\ldots,L}(u_{ik})\}} \tag{A.1}$$

et

$$\mathbf{u}_i \sim \mathcal{N}_L(\mathbf{X}_i^* \boldsymbol{\beta}^*, \boldsymbol{\Omega}^*) \tag{A.2}$$

où $\mathbf{X}_i^*$ est une matrice de variables explicatives de dimension $L \times p$, $\boldsymbol{\beta}^*$ est un vecteur de dimension $p \times 1$ et $\boldsymbol{\Omega}^*$ est une matrice de variance-covariance de dimension $L \times L$.

Le modèle ainsi décrit (équations A.1 et A.2) n'est pas identifiable. Pour assurer l'identifiabilité, le modèle doit être reformulé en terme de $L-1$ différences relatives $z_{il} = u_{il} - u_{iL}$. Le vecteur $\mathbf{z}_i = (z_{i1}, \ldots, z_{iL-1})$ a pour distribution :

$$\mathbf{z}_i \sim \mathcal{N}_{L-1}(\mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Omega})$$

où $\mathbf{X}_i$, $\boldsymbol{\beta}$ et $\boldsymbol{\Omega}$ sont les transformations appropriées de $\mathbf{X}_i^*$, $\boldsymbol{\beta}^*$ et $\boldsymbol{\Omega}^*$. L'équation A.1 s'exprime alors sous la forme :

$$\begin{aligned} w_{il} &= \mathbb{1}_{\{z_{il} \geq 0,\, z_{il} = \max_{k=1,\ldots,L}(z_{ik})\}}\, l = 1, \ldots, L-1, \\ w_{iL} &= \mathbb{1}_{\{z_{il} < 0, \forall l\}}. \end{aligned}$$

Puisque l'échelle de la variable auxiliaire z_{il} est indéterminée, on fixe le premier élément de la diagonale de $\mathbf{\Omega}$ à 1 (Geweke et al., 1997).

Une généralisation de ce modèle au cas spatial, dans un cadre univarié, a été proposée par Wall et Liu (2009). Soit $Y(\mathbf{s}_i)$ une variable nominale à L modalités au point $\mathbf{s}_i$. Pour chaque variable $Y(\mathbf{s}_i)$, on considère un vecteur sous-jacent $\mathbf{u}(\mathbf{s}_i) = (u_1(\mathbf{s}_i), \ldots, u_L(\mathbf{s}_i))'$ de loi normale tel que la modalité l est observée si la $l^{\text{ème}}$ composante de $\mathbf{u}(\mathbf{s}_i)$ est plus grande que les autres, i.e.

$$Y(\mathbf{s}_i) = l \text{ si } u_l(\mathbf{s}_i) = \max_{k=1,\ldots,L}(u_k(\mathbf{s}_i))$$

et

$$\mathbf{u}(\mathbf{s}_i) = \boldsymbol{\mu}_{\mathbf{u}} + \mathbf{u}_0(\mathbf{s}_i)$$

où $\boldsymbol{\mu}_{\mathbf{w}}$ est un vecteur de paramètres de taille L (ce vecteur peut être remplacé par des covariables) et $\mathbf{u}_0(\mathbf{s}_i)$ est un vecteur aléatoire de loi normale centrée et de matrice de covariance inconnue de dimension $L \times L$.

Le modèle tel qu'il est présenté ici n'est pas identifiable et doit être reformulé en terme de différences. La dernière catégorie L est choisie comme référence. Un nouveau vecteur $\mathbf{z}(\mathbf{s}_i) = (z_1(\mathbf{s}_i), \ldots, z_{L-1}(\mathbf{s}_i))'$ où $z_l(\mathbf{s}_i) = u_l(\mathbf{s}_i) - u_L(\mathbf{s}_i)$ est utilisé. Ce vecteur $\mathbf{z}(\mathbf{s}_i)$ de taille $L-1$ a une distribution gaussienne et peut s'écrire sous la forme :

$$\mathbf{z}(\mathbf{s}_i) = \boldsymbol{\mu}_{\mathbf{z}} + \mathbf{S}(\mathbf{s}_i)$$

où $\mathbf{S}(\mathbf{s}_i)$ est un vecteur gaussien de dimension $L-1$ de moyenne nulle et de matrice de covariance inconnue. On a alors :

$$\begin{aligned} Y(\mathbf{s}_i) = l \quad &\text{si} \quad \max_{k=1,\ldots,L-1}(z_k(\mathbf{s}_i)) = z_l(\mathbf{s}_i) \text{ et } z_l(\mathbf{s}_i) \geq 0,\ l = 1, \ldots, L-1, \\ Y(\mathbf{s}_i) = L \quad &\text{si} \quad \max_{k=1,\ldots,L-1}(z_k(\mathbf{s}_i)) < 0. \end{aligned}$$

Comme dans le cas non spatial, le premier terme de la diagonale de la matrice de covariance de $\mathbf{S}(\mathbf{s}_i)$ doit être fixé à 1 pour assurer l'identifiabilité du modèle. Chaque composante spatiale $S_l(\mathbf{s}_i)$ du vecteur $\mathbf{S}(\mathbf{s}_i)$ est obtenue suivant la construction moyenne mobile :

$$\forall l = 1, \ldots, L,\ S_l(\mathbf{s}) = \int_{\mathbb{R}^2} f_l(\mathbf{x} - \mathbf{s}) V(\mathbf{x}) \mathrm{d}\mathbf{x}$$

où f_l est une fonction moyenne mobile et $V(.)$ un mélange de bruits blancs. La forme de la matrice de covariance est alors connue. Le traitement de la variable nominale rentre donc lui aussi dans le cadre du modèle hiérarchique spatial multivarié proposé ci-dessus. La principale difficulté liée au traitement des variables nominales est la taille des matrices à manipuler. Pour un modèle univarié ne comprenant qu'une variable nominale à L modalités, la matrice de covariance du vecteur $(\mathbf{S}(\mathbf{s}_1)', \ldots, \mathbf{S}(\mathbf{s}_n)')$ est de taille $n(L-1) \times n(L-1)$. Si le modèle comprend K variables nominales chacune ayant L_k modalités, les matrices à manipuler sont de taille $n(\sum_{k=1}^{K} L_k - K) \times n(\sum_{k=1}^{K} L_k - K)$. Le traitement des variables nominales s'avère donc difficile en pratique, excepté si le nombre de variables nominales, le

nombre de points échantillonnés et le nombre de modalités de chaque variable nominales sont faibles. C'est pourquoi nous nous sommes essentiellement intéressés aux trois types de variables considérées dans la description initiale du modèle.

Annexe B

Calcul des probabilités d'observer le génotype G

On considère que le génotypage des arbres a été effectué sur L loci. Le génotype G du juvénile auquel nous nous intéressons peut donc s'écrire sous la forme :

$$a_{11}^G, a_{12}^G \quad a_{21}^G, a_{22}^G \quad \dots \quad a_{l1}^G, a_{l2}^G \quad \dots \quad a_{L1}^G, a_{L2}^G.$$

où a_{li}^G désigne le $i^{\text{ème}}$ allèle du $l^{\text{ème}}$ locus. De même, le génotype de l'adulte h peut s'écrire :

$$a_{11}^h, a_{12}^h \quad a_{21}^h, a_{22}^h \quad \dots \quad a_{l1}^h, a_{l2}^h \quad \dots \quad a_{L1}^h, a_{L2}^h.$$

On désigne par f_a la fréquence de l'allèle a dans la population.

B.1 Calcul de $\mathbb{P}(G|h,j)$

La probabilité $\mathbb{P}(G|h,j)$ est la probabilité pour qu'un juvénile ayant pour mère h et pour père j présente le génotype G. On considère que les loci sont indépendants :

$$\mathbb{P}(G|h,j) = \prod_{l=1}^{L} \mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \{a_{l1}^j, a_{l2}^j\}).$$

Le calcul de la probabilité $\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \{a_{l1}^j, a_{l2}^j\})$ s'effectue différemment suivant que le juvénile est homozygote (a_{l1}^G et a_{l2}^G sont identiques) ou hétérozygote (a_{l1}^G et a_{l2}^G sont différents) au locus l.

- Si le juvénile est homozygote au locus l, la probabilité $\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \{a_{l1}^j, a_{l2}^j\})$ est égale au produit de la probabilité d'observer l'allèle a_{l1}^G chez la mère et de la probabilité d'observer cet allèle chez le père :

$$\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \{a_{l1}^j, a_{l2}^j\}) = \frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l1}^G = a_{li}^j\}}}{2} + \frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l1}^G = a_{li}^j\}}}{2}.$$

- Si le juvénile est hétérozygote au locus l, la probabilité $\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \{a_{l1}^j, a_{l2}^j\})$ se décompose en une somme :

$$\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \{a_{l1}^j, a_{l2}^j\}) = \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l2}^G = a_{li}^j\}}}{2} + \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l2}^G = a_{li}^h\}}}{2} \times \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^j\}}}{2}.$$

  Le premier terme de la somme traduit la possibilité que l'allèle a_{l1}^G ait été transmis par la mère et l'allèle a_{l2}^G par le père. Le second terme traduit la possibilité que l'allèle a_{l2}^G ait été transmis par la mère et l'allèle a_{l1}^G par le père.

Finalement, la probabilité $\mathbb{P}(G|h, j)$ s'écrit :

$$\mathbb{P}(G|h,j) = \prod_{l=1}^{L} \left[\mathbb{1}_{\{a_{l1}^G = a_{l2}^G\}} \left(\frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^j\}}}{2} \right) + \mathbb{1}_{\{a_{l1}^G \neq a_{l2}^G\}} \left(\frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l2}^G = a_{li}^j\}}}{2} + \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l2}^G = a_{li}^h\}}}{2} \times \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^j\}}}{2} \right) \right].$$

B.2 Calcul de $\mathbb{P}(G|h, \text{ext})$

$\mathbb{P}(G|h, \text{ext})$ désigne la probabilité pour qu'un juvénile ayant pour mère h et un père hors de B présente le génotype G. On considère que les loci sont indépendants :

$$\mathbb{P}(G|h, \text{ext}) = \prod_{l=1}^{L} \mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \text{ext}).$$

Le calcul de la probabilité $\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \text{ext})$ s'effectue différemment suivant que le juvénile est homozygote ou hétérozygote au locus l.

- Si le juvénile est homozygote au locus l, la probabilité $\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \text{ext})$ est égale au produit de la probabilité d'observer l'allèle a_{l1}^G chez la mère et de la probabilité d'observer cet allèle chez le père. Ici, le père n'est pas connu. La probabilité d'observer cet allèle chez le père est donc égale à la probabilité d'observer cet allèle sur un individu quelconque de la population. Cette probabilité est donnée par la fréquence allélique $f_{a_{l1}^G}$:

$$\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \text{ext}) = \frac{\sum_{i=1}^2 \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times f_{a_{l1}^G}.$$

- Si le juvénile est hétérozygote au locus l, la probabilité $\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \text{ext})$ se décompose en une somme :

$$\mathbb{P}(\{a_{l1}^G, a_{l2}^G\}|\{a_{l1}^h, a_{l2}^h\}, \text{ext}) = \frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times f_{a_{l2}^G} + \frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l2}^G = a_{li}^h\}}}{2} \times f_{a_{l1}^G}.$$

  Le premier terme de la somme traduit la possibilité que l'allèle a_{l1}^G ait été transmis par la mère et l'allèle a_{l2}^G par un père inconnu. Le second terme traduit la possibilité que l'allèle a_{l2}^G ait été transmis par la mère et l'allèle a_{l1}^G par le père inconnu.

Finalement, la probabilité $\mathbb{P}(G|h, \text{ext})$ s'écrit :

$$\mathbb{P}(G|h, \text{ext}) = \prod_{l=1}^{L} \left[\mathbb{1}_{\{a_{l1}^G = a_{l2}^G\}} \left(\frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times f_{a_{l1}^G} \right) + \mathbb{1}_{\{a_{l1}^G \neq a_{l2}^G\}} \left(\frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l1}^G = a_{li}^h\}}}{2} \times f_{a_{l2}^G} + \frac{\sum_{i=1}^{2} \mathbb{1}_{\{a_{l2}^G = a_{li}^h\}}}{2} \times f_{a_{l1}^G} \right) \right].$$

B.3 Calcul de $\mathbb{P}(G|\text{ext})$

$\mathbb{P}(G|\text{ext})$ désigne la probabilité qu'un juvénile ayant une mère hors de B présente le génotype G. Autrement dit, les parents du juvénile ne sont pas connus. Cela revient à calculer la probabilité pour qu'un juvénile issu d'un accouplement aléatoire présente le génotype G. Les locis sont considérés comme indépendants :

$$\begin{aligned} \mathbb{P}(G|\text{ext}) &= \prod_{l=1}^{L} \mathbb{P}(\{a_{l1}^G, a_{l2}^G\}) \\ \mathbb{P}(G|\text{ext}) &= \prod_{l=1}^{L} \mathbb{P}(\{a_{l1}^G\}) \mathbb{P}(\{a_{l2}^G\}). \end{aligned}$$

La probabilité d'observer un allèle particulier dans la population est donnée par sa fréquence allélique, d'où :

$$\mathbb{P}(G|\text{ext}) = \prod_{l=1}^{L} f_{a_{l1}^G} f_{a_{l2}^G}.$$

Bibliographie

Achard, F., H. D. Eva, H.-J. Stibig, P. Mayaux, J. Gallego, T. Richards, et J.-P. Malingreau (2002). Determination of deforestation rates of the world's humid tropical forests. *Science 297*(5583), 999–1002.

Adams, W. T. et D. Birkes (1989). Mating patterns in seed orchards. In *Proceedings of 20th Southern Forest Tree Improvement Conference*, Charleston, South Carolina.

Adams, W. T. et D. S. Birkes (1991). Estimating mating patterns in forest tree populations. In M. E. Malvolti, F. Cannata, et H. H. Hattemer (Eds.), *Biochemical markers in the population genetics of forest trees*, The Hague, The Netherlands. SPB Academic Publishing.

Albert, J. et S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc. 88*(422), 669–679.

Andrieu, C. et E. Moulines (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab. 16*(3), 1462–1505.

Armstrong, M. et G. Matheron (1986a). Disjunctive kriging revisited : Part I. *Math. Geol. 18*(8), 711–728.

Armstrong, M. et G. Matheron (1986b). Disjunctive kriging revisited : Part II. *Math. Geol. 18*(8), 729–742.

Ashford, S. et R. Swoden (1970). Multivariate probit analysis. *Biometrics 26*, 535–546.

Atchade, Y. (2006). An adaptive version for the Metropolis Adjusted Langevin Algorithm with a truncated drift. *Methodol. Comput. Appl. Probab. 8*(2), 235–254.

Atchade, Y. F. et J. S. Rosenthal (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli 11*(5), 815–828.

Atkinson, P. et C. Lloyd (2001). Ordinary and indicator kriging of monthly mean nitrogene dioxide concentrations in the United Kingdom. In P. Monestiez, D. Allard, et R. Froidevaux (Eds.), *geoENV III - Geostatistics for Environmental Applications*. Kluwer academic publishers.

Austerlitz, F., C. Dick, C. Dutech, E. Klein, S. Oddou-Muratorio, P. Smouse, et V. Sork (2004). Using genetic markers to estimate the pollen dispersal curve. *Mol. Ecol. 13*(4), 937–954.

Baillargeon, S. (2005). Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations. Mémoire de maîtrise, Université Laval, Québec.

Banerjee, S., B. Carlin, et A. Gelfand (2004). *Hierarchical modeling and analysis for spatial data*. Number 101 in Monogr. Statist. Appl. Probab. Boca Raton, FL : Chapman & Hall/CRC.

Baraloto, C. (2001). *Tradeoffs between neotropical tree seedling traits and performance in contrasting environments*. Ph. D. thesis, University of Michigan, Ann Arbor, Michigan.

Baraloto, C. (2003). Régénération forestière naturelle : De la graine à la jeune tige. *Rev. For. Fr. 55*(NS), 179–187.

Barret, J. (Ed.) (2001). *Atlas illustré de la Guyane*, Paris, France. IRD (Institut de Recherche pour le Développement).

Barry, R. et J. Ver Hoef (1996). Blackbox kriging : spatial prediction without specifying variogram models. *J. Agric. Biol. Environ. Stat. 1*(3), 297–322.

Beckage, B. et J. S. Clark (2003). Seedling survival and growth of three forest tree species : the role of spatial heterogeneity. *Ecology 84*(7), 1849–1861.

Besag, J. E. (1994). Discussion of paper by U. Grenander and M. I. Miller. *J. Roy. Statist. Soc. Ser. B 56*, 591–592.

Bogaert, P. (2002). Spatial prediction of categorical variables : the Bayesian maximum entropy approach. *Stoch. Environ. Res. Risk Assess. 16*, 425–448.

Bogaert, P. et D. D'Or (2002). Estimating soil properties from thematic soil maps : The Bayesian Maximum Entropy approach. *Soil Sci. Soc. Am. J. 66*, 1492–1500.

Bourrennane, H., S. Salvador-Blanes, S. Cornu, et D. King (2003). Scale of spatial dependence between chemical properties of topsoil and subsoil over a geologically contrasted area (Massif Central, France). *Geoderma 112*, 235–251.

Bracewell, R. (1965). *The Fourier transform and its applications*. New York, NY : McGraw-Hill.

Breyer, L. A. et G. O. Roberts (2000). From Metropolis to diffusions : Gibbs states and optimal scaling. *Stochastic. Process. Appl. 90*, 181–206.

Brigham, E. (1974). *The fast Fourier transform*. Englewood Cliffs, New Jersey : Prentice-Hall, Inc.

Burczyk, J., W. Adams, D. Birkes, et I. Chybicki (2006). Using genetic markers to directly estimate gene flow and reproductive success parameters in plants on the basis of naturally regenerated seedlings. *Genetics 173*, 363–372.

Burczyk, J., W. T. Adams, F. Moran, et A. R. Griffin (2002). Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. *Mol. Ecol. 11*, 2379–2391.

Burgos, A., A. A. Grez, et R. O. Bustamante (2008). Seed production, pre-dispersal seed predation and germination of *Nothofagus glauca* (Nothofagaceae) in a temperate fragmented forest in Chile. *Forest. Ecol. Manag. 255*(3-4), 1226–1233.

Burley, J. (2002). La diversité biologique forestière : tour d'horizon. *Unasylva 209*(53), 3–9.

Cabrera-Gaillard, C. et J. Gignoux (1989). Répartitions spatiales et sylviculture en forêt guyanaise. Technical report, CIRAD-Forêt, Kourou.

Caron, H., S. Dumas, G. Marque, C. Messier, E. Bandou, R. J. Petit, et A. Kremer (2000). Spatial and temporal distribution of chloroplast DNA polymorphism in a tropical tree species. *Mol. Ecol. 9*, 1089–1098.

Caron, H., C. Dutech, et E. Bandou (1998). Variations spatiotemporelles du régime de reproduction de *Dicorynia guianensis* Amshoff (Caesalpiniaceae) en forêt guyanaise. *Genet. Sel. Evol. 30*(Suppl. 1), 153–166.

Caswell, H. (2001). *Matrix population models : Construction, analysis and interpretation* (Second ed.). Sunderland, Massachusetts : Sinauer Associates, Inc. Publishers.

Chagneau, P., F. Mortier, et N. Picard (2009). Designing permanent sample plots by using a spatially hierarchical matrix population model. *J. Roy. Statist. Soc. Ser. C 58*(3), 345–367.

Chen, M.-H. et Q.-M. Shao (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *J. Multivariate Anal. 71*(2), 277–296.

Chib, S. et E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika 85*(2), 347–361.

Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method : cokriging. *J. Real Estate Res. 29*(1), 92–114.

Chilès, J. et P. Delfiner (1999). *Geostatistics. Modeling spatial uncertainty.* Wiley Ser. Probab. Stat. New York : John Wiley & sons.

Christakos, G. (1990). A Bayesian maximum-entropy view to the spatial estimation problem. *Math. Geol. 22*(7), 763–777.

Christakos, G. (1998). Bayesian Maximum Entropy analysis and mapping : a farewell to kriging estimators ? *Math. Geol. 30*(4), 435–462.

Christensen, O. et R. Waagepetersen (2002). Bayesian prediction of spatial count data using Generalized Linear Mixed Models. *Biometrics 58*, 280–286.

Christensen, O. F., J. Møller, et R. P. Waagepetersen (2001). Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalized linear mixed models. *Methodol. Comput. Appl. Probab. 3*, 309–327.

Clark, D. A. et D. B. Clark (1984). Spacing dynamics of a tropical rain forest tree : evaluation of the Janzen-Connell model. *Am. Nat. 124*(6), 769–788.

Clark, J., B. Beckage, P. Camill, B. Cleveland, J. HilleRisLambers, J. Lichter, J. McLachlan, J. Mohan, et P. Wyckoff (1999). Interpreting recruitment limitation in forests. *Am. J. Bot. 86*(1), 1–16.

Clark, J., M. Silman, R. Kern, E. Macklin, et J. HilleRisLambers (1999). Seed dispersal near and far : patterns accross temperate and tropical forests. *Ecology 80*(5), 1475–1494.

Collinet, F. (1997). *Essai de regroupements des principales espèces structurantes d'une forêt dense humide d'après l'analyse de leur répartition spatiale.* Thèse de doctorat, Université Lyon I, Lyon, France.

Cowles, M. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statist. Comput. 6*, 101–111.

Cressie, N. (1991). *Statistics for spatial Data.* John Wiley & Sons.

Dacunha-Castelle, D. et M. Duflo (1982). *Probabilités et statistiques. Tome 1 : Problèmes à temps fixe.* Paris : Masson.

Daganzo, C. (1979). *Multinomial probit. The theory and its application to demand forecasting.* New York, NY : Academic Press, A Subsidiary of Harcourt Brace Jovanovich, Publishers.

Daley, D. et D. Vere-Jones (1988). *An introduction to the theory of point processes.* Springer Ser. Statist. Springer-Verlag.

de Coligny, F., P. Ancelin, G. Cornu, B. Courbaud, P. Dreyfus, F. Goreaud, S. Gourlet-Fleury, C. Meredieu, C. Orazio, et L. Saint-André (2004). Capsis : Computer-Aided Projection for Strategies In Sylviculture : Open architecture for a shared forest-modelling plateform. In *Fourth Workshop IUFRO S5.01.04 conference-September 8-15 2002*, Harrison, British Columbia, Canada, pp. 371–380.

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Comput. Statist. Data Anal. 34*(3), 299 – 314.

DeAngelis, D. et L. Gross (1992). *Individual-based models and approaches in ecology - Populations, communities and ecosystems.* New York, NY : Chapman & Hall.

Diggle, P., J. Tawn, et R. Moyeed (1998). Model-based geostatistics. (With discussion). *J. Roy. Statist. Soc. Ser. C 47*(3), 299–350.

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

D'Or, D. et P. Bogaert (2004). Spatial prediction of categorical variables with the Bayesian Maximum Entropy approach : the Ooypolder case study. *Eur. J. Soil Sci. 55*, 763–775.

Eerikäinen, K., J. Miina, et S. Valkonen (2007). Models for the regeneration establishment and the development of established seedlings in uneven-aged, Norway spruce forest stands of southern Finland. *Forest. Ecol. Manag. 242*, 444–461.

Eidsvik, J., S. Martino, et H. Rue (2009). Approximate Bayesian inference in spatial generalized linear mixed models. *Scand. J. Statist. 36*, 1–22.

Epron, D., A. Bosc, D. Bonal, et V. Freycon (2006). Spatial variation of soil respiration across a topographic gradient in a tropical rain forest in French Guiana. *J. Trop. Ecol. 22*, 565–574.

Falconer, D. S. (1974). *Introduction à la génétique quantitative.* Masson.

FAO-ISRIC-ISSS (1998). World reference base for soil resources. In *World Soil Resources Report 84*, pp. 109.

Favrichon, V. (1997). Réaction de peuplements forestiers tropicaux à des interventions sylvicoles. *Bois For. Trop. 254*(4), 5–24.

Ferry, B., V. Freycon, et D. Paget (2003). Genèse et fonctionnement hydrique des sols sur socle cristallin en Guyane. *Rev. For. Fr. 55*(spéc.), 37–59.

Flores, O. (2005). *Déterminisme de la régénération chez quinze espèces d'arbres tropicaux en forêt guyanaise : les effets de l'environnement et de la limitation par la dispersion.* Thèse de doctorat, Université de Montpellier II.

Flores, O., V. Rossi, et F. Mortier (2009). Autocorrelation offsets zero-inflation in models of tropical saplings density. *Ecol. Model. 220*, 1797–1809.

Franc, A., S. Gourlet-Fleury, et N. Picard (2000). *Une introduction à la modélisation des forêts hétérogènes.* ENGREF.

Gaetan, C. et X. Guyon (2008). *Modélisation et statistique spatiales.* Springer.

Gelfand, A. E., S. Banerjee, C. F. Sirmans, Y. Tu, et S. E. Ong (2007). Multilevel modeling using spatial processes : Application to the Singapore housing market. *Comput. Statist. Data Anal. 51*(7), 3567–3579.

Gelman, A., J. B. Carlin, H. S. Stern, et D. B. Rubin (2004). *Bayesian data analysis* (Second ed.). Boca Raton, FL : Chapman & Hall/CRC.

Gelman, A., G. Roberts, et W. Gilks (1996). Efficient Metropolis jumping rules. In J. Bernardo, J. Berger, A. David, et A. Smith (Eds.), *Bayesian Statistics 5*. Oxford University Press.

Gelman, A. et D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci. 7*(4), 457–472.

Gerber, S., Latouche-Hallé, M. Lourmas, M. Morand-Prieur, S. Oddou-Muratorio, L. Schiber, E. Bandou, H. Caron, B. Degen, N. Frascaria-Lacoste, A. Kremer, F. Lefèvre, et B. Musch (2004). Flux de gènes par pollen et par graines chez quelques espèces forestières : exemples des chênes, de l'alisier, du cèdre et du frêne. *RDV techniques de l'ONF Hors-série*(1), 16–23.

Getzin, S., T. Wiegand, K. Wiegand, et F. He (2008). Heterogeneity influences spatial patterns and demographics in forest stands. *J. Ecol. 96*(4), 807–820.

Geweke, J. F., M. P. Keane, et D. E. Runkle (1997). Statistical inference in the multinomial multiperiod probit model. *J. Econometrics 80*(1), 125–165.

Gilks, W. R., G. O. Roberts, et S. K. Sahu (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc. 93*, 1045–1054.

Golam Kibria, B., L. Sun, J. V. Zidek, et N. D. Le (2002). Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. *J. Amer. Statist. Assoc. 97*(457), 112–124.

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

Goreaud, F., F. de Coligny, B. Courbaud, P. Dreyfus, et T. Pérot (2005). La modélisation : un outil pour la gestion et l'aménagement en forêt. *VertigO - la revue électronique en sciences de l'environnement 6*(2).

Goto, S., K. Shimatani, H. Yoshimaru, et Y. Takahashi (2006). Fat-tailed gene flow in the dioecious canopy tree species *Fraxinus mandshurica* var. *japonica* revealed by microsatellites. *Mol. Ecol. 15*, 2985–2996.

Goulard, M. et M. Voltz (1992). Linear coregionalization model : tools for estimation and choice of cross-variogram matrix. *Math. Geol. 24*(3), 269–286.

Gourlet-Fleury, S., J. Guehl, et O. Laroussinie (2004). *Ecology and management of neotropical rainforest lessons drawn from Paracou, a long-term experimental research site in French Guiana.* Paris : Elsevier.

Grzebyk, M. et H. Wackernagel (1994). Multivariate analysis and spatial/temporal scales : Real and complex models. In *Proceedings of the XVIIth International Biometric Conference, Hamilton, Ontario.*

Haario, H., E. Saksman, et J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli 7*, 223–242.

Hartl, D. et G. Clark (1997). *Principles of Population Genetics* (3rd ed.). Sinauer Associates, Inc.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Herrera, C. (2002). Topsoil properties and seedling recruitment in *Lavandula latifolia* : stage-dependence and spatial decoupling of influential parameters. *OIKOS 97*(2), 260–270.

Higdon, D. (2001). Space and space-time modeling using process convolutions. Technical report, Institute of statistical and decisison sciences, Duke University.

HilleRisLambers, J., J. S. Clark, et B. Beckage (2002). Density-dependent mortality and the latitudinal gradient in species diversity. *Nature 417*(6890), 732–735.

Houle, G. (1995). Seed dispersal and seedling recruitment : the missing link(s). *Ecoscience 2*(3), 238–244.

Howe, H. et J. Smallwood (1982). Ecology of seed dispersal. *Annu. Rev. Ecol. Syst. 13*, 201–228.

Jésel, S. (2005). *Écologie et dynamique de la régénération de Dicorynia guianensis (Caesalpiniaceae) dans une forêt guyanaise.* Thèse de doctorat, Institut National Agronomique Paris-Grignon, Paris.

Joe, H. (1997). *Multivariate models and dependence concepts.* Number 73 in Monogr. Statist. Appl. Probab. London : Chapman & Hall.

Jones, F. et H. Muller-Landau (2008). Mesuring long-distance seed dispersal in complex natural environments : an evaluation and integration of classical and genetic methods. *J. Ecol. 96*, 642–652.

Journel, A. G. et C. J. Huijbregts (1978). *Mining Geostatistics.* London : Academic press.

Kern, J. (2000). *Bayesian process-convolution approaches to specifying spatial dependence structure.* Ph. D. thesis, Institute of Statistics and Decision Sciences, Duke University.

Kokou, K. (1994). Évolution spatiale des agrégats d'angélique de Guyane (*Dicorynia guianensis*, Caesalpiniaceae) sur le dispositif d'étude "Forêt naturelle" de Paracou en Guyane française. *Acta Bot. Gallica 141*(3), 351–359.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chem. Met. Min. Soc. S. Africa 52*, 119–139.

Kuo, H. (2001). White noise theory. In *Kannan, D. (ed.) et al., Handbook of stochastic analysis and applications.*, Volume 163. New York, NY : Marcel Dekker.

Lark, R. M. et R. B. Ferguson (2004). Mapping risk of soil nutrient deficiency or excess by disjunctive and indicator kriging. *Geoderma 118*, 39–53.

Latouche-Hallé, C., A. Ramboer, E. Bandou, H. Caron, et A. Kremer (2003). Nuclear and chloroplast genetic structure indicate fine-scale spatial dynamics in a neotropical tree population. *Heredity 91*, 181–190.

Latouche-Hallé, C., A. Ramboer, E. Bandou, H. Caron, et A. Kremer (2004). Long-distance pollen flow and tolerance to selfing in a neotropical tree species. *Mol. Ecol. 13*, 1055–1064.

Leroy, C. (2000). Caractérisation dendrométrique, architecturale et spatiale de la structure de deux agrégats d'angélique (*Dicorynia guianensis* Amshoff, Caesalpiniaceae). Mémoire de dea, Université Nancy 1, Nancy.

Levrel, H. (2007). *Quels indicateurs pour la gestion de la biodiversité ?* Les cahiers de l'IFB. Institut Français de la Biodiversité.

Lexerød, N. L. (2005). Recruitment models for different tree species in Norway. *Forest. Ecol. Manag. 206*, 91–108.

Loubry, D. (1993). Les paradoxes de l'angélique (*Dicorynia guianensis* Amshoff) : dissémination et parasitisme des graines avant dispersion chez un arbre anémochore de la forêt guyanaise. *Revue d'écologie 48*(4), 353–363.

Lourmas, M. (2003). Diversité génétique et aménagement : utilité d'une modélisation intégrée. *Bois For. Trop. 276*(2), 85–87.

Madelaine, C., R. Pélissier, G. Vincent, J. F. Molino, D. Sabatier, M. F. Prévost, et C. de Namur (2007). Mortality and recruitment in a lowland tropical rain forest of French Guiana : effects of soil type and species guild. *J. Trop. Ecol. 23*(3), 277–287.

Marin, J.-M. et C. P. Robert (2007). *Bayesian core : A pratical approach to computational Bayesian statistics.* Springer text in statistics. New York, NY : Springer-Verlag.

Matheron, G. (1963). *Traité de géostatistique appliquée. Tome II : le krigeage.* Mémoires du BRGM, 24. Paris : Éditions Bureau de Recherches Géologiques et Minières.

Matheron, G. (1973). Le krigeage disjonctif. Technical Report N-360, Centre de géostatistique, École des mines de Paris, Fontainebleau.

McBratney, A. B., I. O. Odeh, T. F. Bishop, M. S. Dunbar, et T. M. Shatar (2000). An overview of pedometric techniques for use in soil survey. *Geoderma 97*(3-4), 293–327.

Messaoud, Y. et G. Houle (2006). Spatial patterns of tree seedling establishment and their relationship to environmental variables in a cold-temperate deciduous forest of eastern North America. *Plant Ecol. 185*, 319–331.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, et E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–1092.

Miller, J. et J. Franklin (2002). Modeling the distribution of four vegetation alliances using GLM and classification trees with spatial dependence. *Ecol. Model. 157*, 227–247.

Møller, J. et R. P. Waagepetersen (2004). *Statistical inference and simulation for spatial point processes.* Number 100 in Monog. Statist. Appl. Probab. Boca Raton, FL : Chapman & Hall/CRC.

Nabe-Nielsen, J., J. Kollmann, et M. Peña-Claros (2009). Effects of liana load, tree diameter and distances between conspecifics on seed production in tropical timber trees. *Forest. Ecol. Manag. 257*(3), 987–993.

Natarajan, R., C. McCulloch, et N. Kiefer (2000). A Monte Carlo EM method for estimating multinomial probit models. *Comput. Statist. Data Anal. 34*, 33–50.

Nathan, R. et H. Muller-Landau (2000). Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends Ecol. Evol. 15*(7), 278–285.

Ochi, Y. et R. L. Prentice (1984). Likelihood inference in a correlated probit regression model. *Biometrika 71*, 531–543.

ONF (2004). Diversité génétique des arbres forestiers : un enjeu de gestion ordinaire. *Rendez-vous techniques de l'ONF Hors-série*(1).

Parent, E. et J. Bernier (2007). *Le raisonnement bayésien : Modélisation et inférence.* Springer.

Pavé, A. (1994). *Modélisation en biologie et en écologie.* Aléas.

Press, W., S. Teukolsky, W. Vetterling, et B. Flannery (1992). *Numerical Recipes in C : The Art of Scientific Computing* (Second ed.). Cambridge : Cambridge University Press.

Rathbun, S., S. Shiffman, et C. Gwaltney (2007). Modelling the effects of partially observed covariates on Poisson process intensity. *Biometrika 94*(1), 153–165.

Rathbun, S. L. (1996). Estimation of Poisson intensity using partially observed concomitant variables. *Biometrics 52*(1), 226–242.

Rathbun, S. L. et B. Black (2006). Modeling and spatial prediction of pre-settlement patterns of forest distribution using witness tree data. *Environ. Ecol. Stat. 13*, 427–448.

Rathbun, S. L. et S. Fei (2006). A spatial zero-inflated Poisson regression model for oak regeneration. *Environ. Ecol. Stat. 13*, 409–426.

Ribbens, E., A. Silander, John, et S. W. Pacala (1994). Seedling recruitment in forests : calibrating models to predict patterns of tree seedling dispersion. *Ecology 75*(6), 1794–1806.

Rivoirard, J. (1991). *Introduction au krigeage disjonctif et à la géostatistique non linéaire* (Seconde ed.). Fontainebleau : ENSMP.

Robert, C. P. (1992). *L'analyse statistique bayésienne*. Paris : Éditions Économica.

Robert, C. P. (1996). *Méthodes de Monte Carlo par chaînes de Markov*. Paris : Éditions Économica.

Robert, C. P. (2006). *Le choix bayésien : Principes et pratiques*. Paris : Springer.

Roberts, G. O. et J. S. Rosenthal (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. Roy. Statist. Soc. Ser. B 60*, 255–268.

Roberts, G. O. et R. L. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli 2*, 341–363.

Robledo-Arnuncio, J. J. et C. Garcia (2007). Estimation of the seed dispersal kernel from exact identification of source plants. *Mol. Ecol. 16*(23), 5098–5109.

Rollet, B. (1969). La régénération naturelle en forêt dense sempervirente de plaine de la Guyane Vénézuélienne. *Bois For. Trop. 124*, 19–38.

Rue, H., S. Martino, et N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Statist. Soc. B 71*(2), 319–392.

Sabatier, D., M. Grimaldi, M. Prevost, J. Guillaume, M. Godron, M. Dosso, et P. Curmi (1997). The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. *Plant Ecol. 131*, 81–108.

Sagnard, F., C. Pichot, P. Dreyfus, P. Jordano, et B. Fady (2007). Modelling seed dispersal to predict seedling recruitment : recolonization dynamics in a plantation forest. *Ecol. Model. 203*, 464–474.

Savelieva, E., V. Demyanov, M. Kanevski, M. Serre, et G. Christakos (2005). Bme-based uncertainty assessement of the chernobyl fallout. *Geoderma 128*, 312–324.

Schurr, F., O. Steinitz, et R. Nathan (2008). Plant fecundity and seed dispersal in spatially heterogeneous environments : models, mechanisms and estimation. *J. Ecol. 96*, 628–641.

Shimatani, K. (2004). Spatial molecular ecological models for genotyped adults and offspring. *Ecol. Model. 174*, 401–410.

Shimatani, K., M. Kimura, K. Kitamura, Y. Suyama, Y. Isagi, et H. Sugita (2007). Determining the location of deceased mother tree and estimating forest regeneration variables by use of microsatellites and spatial genetic models. *Popul. Ecol. 49*, 317–330.

Shimatani, K., K. Kitamura, T. Kanazashi, et H. Sugita (2006). Genetic inhomogeneous Poisson processes describing the roles of an isolated mature tree in forest regeneration. *Popul. Ecol. 48*, 203–214.

Snook, L., L. Cámara-cabrales, et M. Kelty (2005). Six years of fuit production by mahogany trees (*Swietenia macrophylla* King) : patterns of variation and implications for sustainability. *Forest. Ecol. Manag. 206*, 221–235.

Stoyan, D. et S. Wagner (2001). Estimating the fruit dispersion of anemochorous forest trees. *Ecol. Model. 145*, 35–47.

Tu, C. (2006). *Bayesian nonparametric modeling using Lévy process priors with application for function estimation, time series modeling and spatio-temporal modeling.* Ph. D. thesis, Institute of Statistics and Decision Sciences, Duke University.

Tufto, J., S. Engen, et K. Hindar (1997). Stochastic dispersal processes in plant populations. *Theor. Popul. Biol. 52*, 16–26.

Van Lieshout, M. (2000). *Markov point processes and their applications.* Imperial College Press.

Vanclay, J. (1992). Modelling regeneration and recruitment in a tropical moist forest. *Can. J. For. Res. 22*(9), 1235–1248.

Vanclay, J. (1995). Growth models for tropical forests : A synthesis of models and methods. *For. Sci. 41*(1), 7–42.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Stat. Anal. 92*(1), 1–28.

Venables, W. N., D. M. Smith, et the R Development Core Team (2009). An introduction to R. ISBN 3-900051-12-7.

Ver Hoef, J. et R. Barry (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *J. Statist. Plann. Inference 69*(2), 275–294.

Ver Hoef, J., N. Cressie, et R. P. Barry (2004). Flexible spatial models for kriging and cokriging using moving averages and the Fast Fourier Transform (FFT). *J. Comput. Graph. Statist. 13*, 265–289.

von Steiger, B., R. Webster, R. Schulin, et R. Lehmann (1996). Mapping heavy metals in polluted soil by disjunctive kriging. *Environ. Pollut. 94*(2), 205–215.

Waagepetersen, R. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika 95*(2), 351–363.

Wackernagel, H. (2003). *Multivariate geostatistics. An introduction with applications.* (3rd completely revised ed.). Berlin : Springer.

Wall, M. M. et X. Liu (2009). Spatial latent class analysis model for spatially distributed multivariate binary data. *Comput. Statist. Data Anal. 53*(8), 3057 – 3069.

Wang, B. et T. Smith (2002). Closing the seed dispersal loop. *Trends Ecol. Evol. 177*(8), 379–385.

Webster, R. et M. Oliver (1989). Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional porbability. *Eur. J. Soil Sci. 40*(3), 497–512.

Wibrin, M., P. Bogaert, et D. Fashbender (2006). Combining categorical and continuous spatial information within the Bayesian maximum entropy paradigm. *Stoch. Environ. Res. Risk Assess. 20*, 423–433.

Wikle, C. (2003). Hierarchical Bayesian models for predicting the spread of ecology processes. *Ecology 84*(6), 1382–1394.

Wolpert, R. L. et K. Ickstadt (1998). Poisson/gamma random field models for spatial statistics. *Biometrika 85*(2), 251–267.

Yaglom, A. (1987). *Correlation theory of stationary and related random functions. Volume I : Basic results.* New York : Springer-Verlag.

Yasaka, M., M. Takiya, I. Watanabe, Y. Oono, et N. Mizui (2008). Variation in seed production among years and among individuals in 11 broadleaf tree species in northern Japan. *J. For. Res. 13*(2), 83–88.

Table des figures

Liste des tableaux

Modélisation bayésienne hiérarchique pour la prédiction multivariée de processus spatiaux non gaussiens et processus ponctuels hétérogènes d'intensité liée à une variable prédite. Application à la prédiction de la régénération en forêt tropicale humide.

Résumé

Un des points faibles des modèles de dynamique forestière spatialement explicites est la modélisation de la régénération. Un inventaire détaillé du peuplement et des conditions environnementales a permis de mettre en évidence les effets de ces deux facteurs sur la densité locale de juvéniles. Mais en pratique, la collecte de telles données est coûteuse et ne peut être réalisée à grande échelle : seule une partie des juvéniles est échantillonnée et l'environnement n'est connu que partiellement. L'objectif est ici de proposer une approche pour prédire la répartition spatiale et le génotype des juvéniles sur la base d'un échantillonnage raisonnable des juvéniles, des adultes et de l'environnement. La position des juvéniles est considérée comme la réalisation d'un processus ponctuel marqué, les marques étant constituées par les génotypes. L'intensité du processus traduit les mécanismes de dispersion à l'origine de l'organisation spatiale et de la diversité génétique des juvéniles. L'intensité dépend de la survie des graines, qui dépend elle-même des conditions environnementales. Il est donc nécessaire de prédire l'environnement sur toute la zone d'étude. L'environnement, représenté par un champ aléatoire multivarié, est prédit grâce à un modèle hiérarchique spatial capable de traiter simultanément des variables de nature différente. Contrairement aux modèles existants où les variables environnementales sont considérées comme connues, le modèle de régénération proposé doit prendre en compte les erreurs liées à la prédiction de l'environnement. La méthode est appliquée à la prédiction de la régénération des juvéniles en forêt tropicale (Guyane française).

Mots-clés : Régénération, Modèle hiérarchique spatial multivarié, Processus ponctuel marqué

Hierarchical Bayesian modelling for multivariate prediction of non Gaussian spatial processes and inhomogeneous spatial point processes with intensity related to a predicted variable. Application to the regeneration prediction in tropical rainforest.

Abstract

One of the weak points of forest dynamics models is the recruitment. Classically, ecologists make the assumption that recruitment mainly depends on both spatial pattern of mature trees and environment. A detailed inventory of the stand and the environmental conditions enabled them to show the effects of these two factors on the local density of seedlings. In practice, such information is not available : only a part of seedlings is sampled and the environment is partially observed. The aim of the paper is to propose an approach in order to predict the spatial distribution and the seedlings genotype on the basis of a reasonable sampling of seedling, mature trees and environmental conditions. The spatial pattern of the seedlings is assumed to be a realization of a marked point process. The intensity of the process is not only related to the seed and pollen dispersal but also to the sapling survival. The sapling survival depends on the environment ; so the environment must be predicted on the whole study area. The environment is characterized through spatial variables of different nature and predictions are obtained using a spatial hierarchical model. Unlike the existing models which assume the environmental covariables as exactly known, the recruitment model we propose takes into account the error related to the prediction of the environment. The prediction of seedling recruitment in tropical rainforest in French Guiana illustrates our approach.

Keywords : Hierarchical spatial model, Marked point process, Regeneration

Discipline : Biostatistique

Laboratoire d'accueil
CIRAD
UR Dynamique des forêts naturelles
Campus international de Baillarguet
TA C-37|D
34 398 Montpellier CEDEX 5

Laboratoire de rattachement
I3M
Université de Montpellier II
Case Courrier 051
Place Eugène Bataillon
34 095 Montpellier CEDEX