Spécifier un processus caché non modélisé en déterminant le lien asymptotique entre résidus et processus caché. Application à l'analyse de la variabilité dans les expériences de propagation des rouilles du blé

Samuel S. Soubeyrand

UNIVERSITÉ MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ MONTPELLIER II

Discipline : Mathématiques Appliquées
Formation doctorale : Biostatistique
Ecole doctorale : Information, Structures, Systèmes

présentée et soutenue publiquement

par

Samuel SOUBEYRAND

le 1er décembre 2005

Titre :

SPÉCIFIER UN PROCESSUS CACHÉ NON MODÉLISÉ EN DÉTERMINANT LE LIEN ASYMPTOTIQUE ENTRE RÉSIDUS ET PROCESSUS CACHÉ

APPLICATION À L'ANALYSE DE LA VARIABILITÉ DANS LES EXPÉRIENCES DE PROPAGATION DES ROUILLES DU BLÉ

JURY

Jean-Noël Bacro, Université Montpellier II	Président
Liliane Bel, Université Paris-Sud	Examinatrice
Joël Chadœuf, INRA Avignon	Directeur de thèse
Ivan Sache, INRA Versailles-Grignon	Co-directeur de thèse
Bernard Tivoli, INRA Rennes	Rapporteur
Jian-Feng Yao, Université Rennes I	Rapporteur

Préface

Samuel m'a demandé "Ivan, fais moi une préface". C'était une commande de Samuel ; une commande de Samuel, ça ne se refuse pas.

Maintenant que tout est accompli, c'est plutôt une postface, en espérant que cela n'apparaisse pas comme une épitaphe ou une apothéose.
Des collègues bien intentionnés, un tantinet envieux, m'avaient dit : "Tu as de la chance, on t'a fourni un thésard pour interpréter tes données"[1]. D'abord les données, ce n'est pas donné ; ensuite, les doctorants non plus. Après avoir longtemps contemplé mes données, Samuel a dit que ce n'était pas bien de la part des épidémiologistes d'ajuster des gradients sur des moyennes, surtout avec cinq points. J'en fus fort aise, mais pas très longtemps, puisqu'il a ajouté : "D'ailleurs toi aussi, tu l'as fait, tu l'as même publié"[2]. Même si il faut rire de tout, c'était déjà moins drôle[3].
Contrairement à ce qu'il voudrait me faire dire, Samuel n'est pas impertinent, il est irrévérencieux. La Science gagnerait de plus d'irrévérence et de moins de respect des glorieux aînés, afin de perdre sa majuscule et son conformisme[4]. Avec son regard de mathématicien, Samuel a vu dans mes résultats expérimentaux des choses que je n'y avais jamais vues, malgré tout le temps passé à compter des lésions et à maugréer contre la variabilité biologique et les incertitudes climatiques[5]. En fait, je ne pouvais pas les voir, puisqu'il s'agissait de processus cachés. Les mathématiciens ne voient pas les choses comme nous, les biologistes, et la communication est souvent difficile faute d'une perspective commune. Non seulement Samuel a écouté mes questions de biologiste, mais en plus il y a répondu ; il s'est même assuré que j'avais bien compris les réponses, quitte à réécrire un peu les questions pour ne pas trop m'accabler quand je ne comprenais pas bien[6]. Tout irrévérencieux qu'il soit, Samuel reste prudent en toute circonstance.

[1] *Je pense donc tu suis*
[2] *Marguerite Duras n'a pas écrit que des conneries... Elle en a aussi filmées.*
[3] *Il faut rire de tout. C'est extrêmement important. C'est la seule façon de friser la lucidité sans tomber dedans.*
[4] *Plus le gradé a de barrettes, plus le salut doit être servile.*
[5] *Le travail, c'est pour les gens qui n'ont pas de talent.*
[6] *Quand un philosophe me répond, je ne comprends plus ma question.*

Tout ceci pour dire que ces trois années de parcours commun ont été une aventure humaine et scientifique des plus enrichissantes, avec une ouverture sur des domaines que je ne connaissais pas ou si peu et une expérience de coencadrement avec des personnes que je ne connaissais pas ou si peu. Finalement, je me suis même habitué aux stridulations des cigales comtadines. J'espère avoir bientôt l'occasion de les entendre à nouveau[7].

Thank you, Samuel !

Grignon, *Ivan Sache*
Décembre 2005

Comme le dit Samuel, une thèse est avant tout un accompagnement. Ceci dit, un accompagnement peut être très variable, pouvant aller jusqu'à "l'un traîne l'autre" ou "on marche ensemble mais on s'ignore". Samuel s'avère être un excellent compagon, sachant vous écouter, faire partager ses intérêts comme ses doutes, vous attendre si nécessaire.

Si la thèse peut être vue comme un accompagnement, le processus de recherche associé peut être considéré comme une promenade où le chemin est à tracer. Le plus grand plaisir que j'ai vécu ces trois ans, c'est finalement d'avoir vu le chemin à suivre être décidé ensemble, sans que jamais il ne refuse sa part de responsabilité.

Ainsi, je le remercie d'avoir accepté d'investir le domaine des écarts aux modèles, mon vieux dada. Formalisant une vague idée, il en tire des résultats, mais surtout ouvre des voies d'exploration qu'il nous reste à poursuivre. Je le remercie aussi de m'avoir entraîné dans des réflexions plus larges que la statistique, comme celle des changements d'échelle. Si nous sommes fortement sollicités à y investir, j'ai souvent éludé le problème. Samuel a pris le problème à bras le corps, en m'entraînant avec lui sur ce chemin nettement plus flou mais oh combien plus prospectif.

Avignon, *Joël Chadœuf*
Décembre 2005

[7] *La haine aveugle n'est pas sourde.*
Citations de Pierre Desproges, moraliste français de la fin du XXe siècle (1939-1988), inhumé au cimetière du Père-Lachaise au voisinage de Frédéric Chopin et Michel Petrucciani.

Cette thèse de doctorat, à la croisée de la statistique et de l'épidémiologie végétale, est notamment le fruit de réflexions en groupes et de discussions liées plus ou moins étroitement au sujet d'étude. Nombreux sont ceux qui ont marqué cette thèse d'une empreinte, que celle-ci soit visible ou non. Je leur en suis reconnaissant. En particulier, je remercie ma famille pour m'avoir toujours soutenu dans mes études ; Ivan et Joël pour m'avoir fait cheminer au cours des trois dernières années dans leurs univers respectifs ; les individus avec qui j'ai échangé dans les unités de biométrie d'Avignon et d'épidémiologie de Grignon et durant les séminaires, conférences et rencontres auxquels j'ai participé ; les individus qui m'ont conseillé dans l'écriture de ce rapport ; les membres du jury de thèse et du comité de pilotage (Michel Goulard, Etienne Klein, Christian Lannou, Christophe Lett) pour m'avoir consacré du temps et de l'attention ; certains professeurs et étudiants côtoyés par le passé qui ont contribué à mon parcours vers le doctorat ; l'INRA, son département SPE, son unité d'épidémiologie de Grignon et son unité de biométrie d'Avignon pour avoir financé mes recherches ; et presque tous les êtres de ce monde et des autres.

Avignon, *Samuel Soubeyrand*
Novembre 2005

Table des matières abrégée

Table des matières

1

Contexte, questions, approche

Ce chapitre donne un éclairage sur la propagation des maladies aériennes des végétaux, sur la variabilité de ce phénomène et sur la non-observation de certains processus impliqués. De plus, il présente les deux outils que nous avons développés pour conduire une analyse statistique pas-à-pas de ce phénomène variable et partiellement observé. Les chapitres 2 et 6, qui ouvrent les parties 1 et 2 de ce document, complètent cette introduction en présentant de manière plus ciblée, plus approfondie et plus technique la propagation des maladies aériennes et ces deux outils.

1.1 La propagation des maladies aériennes des végétaux

1.1.1 Fonctionnement de la propagation

Les acteurs de la propagation et leurs actions

Nous nous intéressons à des maladies aériennes des végétaux causées par des champignons (e.g. les rouilles). Un champignon se manifeste sur une plante (e.g. une plante de blé) par des symptômes appelés lésions où sont produites des propagules appelées spores. La spore est l'agent de la dissémination aérienne de la maladie : lorsqu'elle est libérée, transportée et déposée sur une plante sensible, elle peut éventuellement germer, pénétrer dans la plante et donner une nouvelle lésion. La séquence libération - transport - dépôt de la spore est appelée dispersion. La séquence germination - pénétration de la spore - apparition de la lésion est appelée infection. La propagation de la maladie est la répétition, à un temps donné et à des temps successifs, du cycle

1. production de spores dans les lésions,
2. dispersion des spores,
3. infection des plantes par les spores.

Un exemple de scénario

La figure 1.1 représente le suivi de la propagation de la rouille jaune du blé dans une parcelle expérimentale. A la date $t = 1$, cela fait déjà 53 jours que des plantes de blé malades

ont été disposées au centre de la parcelle. Ensuite, en 21 jours (de $t = 1$ à $t = 4$), la maladie s'est répandue presque partout dans la parcelle (à un niveau d'infection dépassant un seuil de détection donné). Mais l'intensité de la maladie n'est pas uniforme. Autour de la source (centre de la parcelle) s'est développé un foyer, dit primaire, de maladie ; et sur les bords de ce foyer primaire, on observe des foyers secondaires.

On peut également décrire le développement épidémique en terme de tendance moyenne, de "creux" et de "pics" de maladie. Quelque soit la date d'observation, les plus fortes concentrations de lésions sont à proximité de la source initiale de maladie, et quand on s'éloigne de la source initiale, la concentration a tendance à décroître. Cette tendance moyenne s'explique ainsi : une spore dispersée est plus probablement déposée à proximité de la lésion qui l'a produite que loin d'elle. Cependant, la décroissance n'est pas régulière : il y a ici et là des creux et des pics de maladie isolés. Les creux peuvent être justifiés ainsi : aucune spore n'a été déposée au site considéré, ou encore le site considéré n'était pas réceptif à la maladie. Les pics de maladie peuvent être justifiés ainsi : le site considéré présentait des conditions particulièrement favorables à la maladie, ou encore des spores venues d'une autre parcelle ont généré en ce site une source de maladie autre que la source volontairement introduite par l'expérimentateur et cette source exogène a répandu localement la maladie.

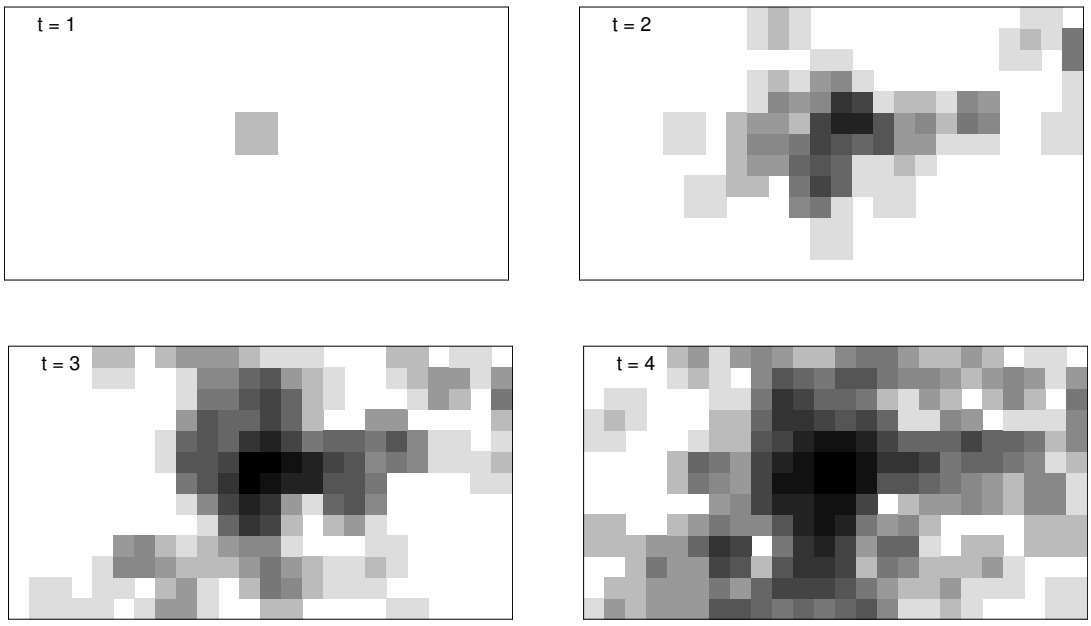


Fig. 1.1. Suivi de la propagation, dans l'espace (25 m × 14 m) et le temps (4 semaines), d'une épidémie de rouille jaune. Noir : concentration élevée de lésions. Source : I. Sache et N. Schermesser.

Variabilité de la propagation

La variabilité de la propagation d'une maladie aérienne est causée par divers processus biologiques et physiques intervenant à différents niveaux.

- Variabilité au niveau de la production des spores : le vieillissement des lésions, le climat, le micro-climat à l'échelle de la feuille, ou encore la densité de lésions font que la production des spores varie dans le temps pour une même lésion et varie d'une lésion à une autre.
- Variabilité au niveau du transport des spores : le vent, la pluie, la structure géométrique du couvert font du transport des spores un mécanisme variable dans l'espace et le temps (même si une spore est déposée plus probablement à proximité de la source que loin d'elle).
- Variabilité au niveau de l'infection : la présence d'une autre maladie ou l'existence de zones où les plantes subissent des carences font du couvert végétal un support hétérogène en terme de réceptivité (i.e. les feuilles ont des propensions variables à être infectées).

La description de la figure 1.1 et le paragraphe précédent illustrent le fait que la propagation est un phénomène qui est caractérisé par des tendances moyennes, mais qui est également caractérisé par une variabilité importante. Comme nous le verrons dans la première partie de ce document, les tendances moyennes du phénomène 'propagation' ont été relativement bien étudiées ; en revanche, les écarts à ces tendances sont moins bien compris et quantifiés. On peut donc se demander

Question 1.1. Quelle approche permettrait d'élucider et quantifier les raisons pour lesquelles la propagation de la maladie s'écarte de ses tendances pour aboutir, par exemple, à la carte en bas à droite de la figure 1.1 ? Autrement dit, comment analyser la variabilité de la propagation ?

Une telle approche aiderait (i) à mieux comprendre la propagation de la maladie, et (ii) à voir si les écarts aux tendances jouent un rôle en terme de ralentissement et d'accélération de la propagation. Remplir les objectifs (i) et (ii) permettrait de mieux appréhender la lutte contre la maladie (cf. section 2.1.4).

1.1.2 Une approche combinant expérimentation et statistique pour analyser la variabilité de la propagation

Les processus causant la variabilité de la propagation se sont combinés pour aboutir aux cartes de maladie de la figure 1.1. Ces cartes contiennent donc de l'information sur certains de ces processus. Par conséquent, réaliser des expériences de propagation et analyser les données qui en sont issues devraient permettre (i) de détecter des processus jouant un rôle dans la propagation et (ii) de quantifier leurs effets. Détecter un processus peut être effectué

en émettant l'hypothèse que ce processus existe et en procédant à un test statistique de cette hypothèse. Quantifier l'effet d'un processus peut être effectué en modélisant ce processus et en ajustant le modèle aux données par des techniques d'estimation statistique.

L'approche esquissée ci-dessus, qui combine expérimentation et statistique, est une réponse possible à la question 1.1. Cependant, des difficultés existent. Elles sont présentées dans les deux paragraphes suivants.

Inférence limitée par l'échelle d'observation

La possibilité de détecter un processus et de quantifier son effet à partir d'une carte de maladie dépend de l'échelle d'observation. Par exemple, les cartes de la figure 1.1 (échelle de la parcelle) ne permettent pas de tester et quantifier le processus d'auto-infection[1] à l'échelle de la feuille, ou encore le processus de transport des spores dans les courants atmosphériques qui cause la propagation de la maladie à l'échelle continentale. En terme statistique, un test portant sur ces processus et réalisé à partir de données à l'échelle de la parcelle manquerait de puissance.

Réaliser des expériences de propagation à différentes échelles doit permettre d'inférer sur des processus intervenant à différentes échelles.

Non-observation de certains processus

Un problème de non-observation rend difficile la quantification des effets de certains processus : dans une expérience telle que celle illustrée par la figure 1.1, la plupart des processus ne sont pas observés directement. Les données récoltées sont généralement

- la position de la source introduite par l'expérimentateur,
- des mesures de concentration de la maladie en différents points de l'espace et du temps, et éventuellement
- des mesures de variables climatiques le plus souvent faites à l'échelle du champ.

Ainsi, les raisons pour lesquelles la maladie ne se propage pas en cercles concentriques éventuellement déformés par l'effet du vent, les raisons pour lesquelles la maladie ne forme pas à chaque pas de temps une bosse lisse sans perturbation locale, ces raisons là ne sont généralement pas observées.

Les processus qui causent la variabilité sont donc des processus sous-jacents qui ne sont observés qu'indirectement au travers des mesures de concentration de la maladie. On parle de processus caché.

Compte tenu des éléments avancés dans cette section, la question 1.1 devient

Question 1.2. Quels outils statistiques permettraient, à partir de données expérimentales de propagation réalisées à différentes échelles,

[1] L'auto-infection advient lorsqu'un spore émise par une lésion sur une feuille génère une lésion fille sur la même feuille. Lorsque la lésion fille est créée sur une autre feuille, on parle d'allo-infection.

– de détecter des processus cachés impliqués dans la propagation d'une maladie, et
– de quantifier les effets de ces processus cachés ?

1.2 Deux outils pour progresser dans la compréhension de la propagation

1.2.1 Une analyse pas-à-pas de la propagation

Ce document expose des outils voués à la détection des processus cachés impliqués dans la propagation et à la quantification de leurs effets à partir de données expérimentales (question 1.2). Ces outils sont

1. un cadre multi-échelle de modélisation-estimation qui permet de construitre des modèles résumant ce que l'on connait de la propagation et de quantifier cette connaissance (partie 1), et
2. une méthode d'analyse de résidus qui aide à caractériser les processus cachés non modélisés (partie 2).

Ces outils sont des composantes du processus de l'analyse statistique illustré par la figure 1.2. En effet, si l'on dispose d'un jeu de données caractéristique du phénomène étudié, le cadre de modélisation-estimation permet de construire un modèle sensé approcher le comportement des variables observées, et d'estimer les paramètres du modèle. Si le modèle est jugé comme étant une bonne approximation au vu de la question posée, alors les conclusions de l'inférence peuvent être tirées. Dans le cas contraire, le modèle doit être modifié. C'est ce que signifie la flèche qui va de droite à gauche dans la figure 1.2. Si l'on connaît mal les raisons pour lesquelles le modèle n'est pas une bonne approximation, alors la méthode d'analyse de résidus permet de guider la modification, ou plus précisément la complexification, du modèle.

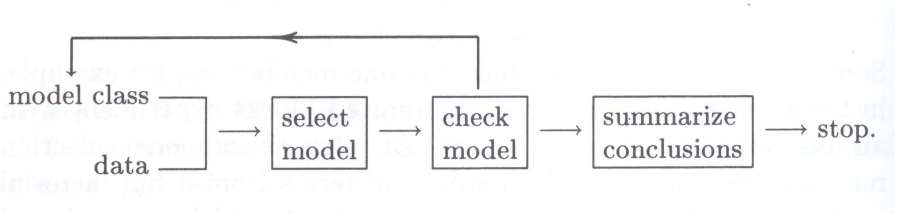


Fig. 1.2. Schéma du processus de l'analyse statistique. La flèche de droite à gauche signifie que le modèle doit être modifié si l'ajustement du modèle aux données (*model check*) n'est pas satisfaisant. Source : McCullagh and Nelder (1989).

En itérant le processus d'analyse statistique exposé ci-dessus, on doit pouvoir appréhender petit à petit la complexité d'un phénomène tel que la propagation des maladies aériennes des végétaux :

Etape 1. analyse des données avec un modèle,

Etape 2. analyse des résidus et complexification du modèle,

Etape 3. analyse des données avec le nouveau modèle, et retour à l'étape 2.
Chaque modèle défini sert de base à partir de laquelle on va chercher de l'information sur les processus cachés non observés. Mais à quel moment cette analyse pas-à-pas doit-elle prendre fin ? A quel moment doit-on s'arrêter de complexifier le modèle ? Ce moment dépend du problème que l'on veut résoudre mais aussi des données. Rien ne sert d'introduire dans le modèle des éléments qui sont sans lien avec le problème posé ou sur lesquels les données disponibles ne contiennent pas d'information.

1.2.2 Premier outil : un cadre multi-échelle de modélisation-estimation

Propagation spatiale

Le cadre multi-échelle de modélisation-estimation sert à décrire, à diverses échelles, la propagation spatiale d'une maladie aérienne depuis une source ponctuelle. Nous définissons la propagation spatiale depuis une source comme le résultat

- de la production de spores dans les lésions mères présentes sur la source,
- de la dispersion des spores, et
- de l'infection des plantes par les spores.

Nous nous sommes limités à l'étude de la propagation spatiale parce que c'est le moteur d'une épidémie, c'est ce qui fait qu'une maladie se propage.

La première partie de ce document a pour objet le cadre de modélisation-estimation. Le chapitre 2 présente les rouilles du blé et leur mode de propagation, expose des modèles de propagation existants et leurs limites, et énonce des principes de modélisation que nous avons appliqués pour construire des modèles de propagation spatiale.

Modèle pour la propagation spatiale à courte distance

Nous avons développé un modèle de fragilité pour décrire la propagation spatiale à courte distance[2] de la rouille brune, i.e. de 0 à 50 centimètres (chapitre 3). Ce modèle a été construit pour analyser des mesures de maladie faites à l'échelle de la feuille de blé (nombre de lésions par feuille). De telles données laissent apparaître une diminution de la concentration de lésions quand la distance à la source augmente, mais également une hétérogénéité des feuilles face à l'infection. La diminution de la concentration de lésions quand la distance à la source augmente a été modélisée par une fonction de dispersion. L'hétérogénéité des feuilles, dont les causes ne sont pas observées, a elle été modélisée en introduisant des effets aléatoires (ou fragilités) dans le modèle.

[2] Evaluer la propagation spatiale à courte distance doit servir à comprendre le développement d'un foyer de maladie.

Modèle pour la propagation spatiale à longue distance

Un autre modèle a été développé pour décrire la propagation à longue distance[3] de la rouille jaune du blé, i.e. de 10 à 250 mètres (chapitre 4). Ce modèle a été construit pour analyser des mesures de maladie faites à l'échelle du mètre carré (nombre de feuilles infectées dans des placettes d'un mètre carré). De telles données laissent apparaître une diminution de la concentration de feuilles malades quand la distance à la source augmente, mais également une anisotropie dans la dispersion des spores. La diminution de la concentration de lésions quand la distance à la source augmente a été modélisée, comme pour la courte distance, par une fonction de dispersion. L'anisotropie de la dispersion a été prise en compte en utilisant des fonctions angulaires déformant la fonction de dispersion.

Cadre commun de modélisation

Les modèles à courte et longue distance ont un élément commun :

– la diminution de la concentration de maladie quand la distance à la source augmente,

et des éléments propres à l'échelle d'observation :

– l'hétérogénéité des feuilles face à l'infection (courte distance) et

– l'anisotropie dans la dispersion des spores (longue distance).

Le chapitre 5 rassemble ces modèles développés à des échelles différentes au sein d'un même cadre de modélisation. De ce cadre peuvent être dérivés différents modèles adaptés à diverses échelles. Ces modèles dérivés partagent des éléments communs, et diffèrent par des éléments propres. Ainsi, on peut étudier l'évolution des éléments communs quand l'échelle change, et on peut étudier quels éléments jouent un rôle majeur ou un rôle mineur selon l'échelle d'intérêt (thème développé au chapitre 10).

Le chapitre 5 discute également d'éléments qui ne sont pas pris en compte dans les modèles (processus cachés non modélisés). Une analyse de résidus classique permet de détecter de tels éléments. Mais une fois que ceux-ci sont détectés, comment complexifier le modèle pour les prendre en compte ? La deuxième outil présenté dans ce document est une réponse à cette question.

1.2.3 Deuxième outil : une méthode d'analyse de résidus

Un modèle peut être vu comme un résumé de la connaissance, et l'ignorance peut être définie comme ce qui n'est pas intégré au modèle. Complexifier un modèle revient alors à réduire la part d'ignorance. L'analyse de résidus exposée dans la partie 2 aide à cette

[3] Evaluer la propagation spatiale à longue distance doit sevir à mesurer les risques d'infection entre parcelles agricoles distantes et à comprendre la formation de foyers secondaires de maladie éloignés du foyer primaire.

réduction. C'est, en termes plus imagés, une méthode qui met l'analyste "*on a road to the unknown*"[4].

Introduction d'effets aléatoires pour complexifier un modèle

Expliquons ce que 'complexifier un modèle' signifie dans ce document. Plaçons-nous dans le cas où les éléments qui ne sont pas intégrés au modèle correspondent à des processus cachés sous-jacents. Prendre en compte un de ces éléments dans un nouveau modèle peut être réalisé en utilisant des effets aléatoires car ce qui n'est pas observé peut être vu comme aléatoire. Un modèle dans lequel il y a des effets aléatoires est un modèle hiérarchique à deux niveaux, le deuxième niveau étant composé des effets aléatoires. La place dans le modèle et la distribution des effets aléatoires dépendent de l'élément en question.

Si l'on sait la plupart du temps où placer les effets aléatoires dans le modèle, on ne sait généralement pas quelle spécification choisir pour leur distribution, c'est-à-dire quelle distribution leur attribuer. La première section du chapitre 6 qui introduit la partie 2 présente les modèles hiérarchiques, en donne des exemples, et décrit le problème de la spécification de la distribution d'effets aléatoires.

Exploitation des résidus d'un modèle de base pour spécifier la distribution des effets aléatoires d'un modèle hiérarchique

En statistique, l'analyse des résidus permet classiquement de tester la validité d'un modèle. L'idée qui sous-tend cette approche est que les résidus sont fonction de ce qui n'a pas été intégré au modèle, c'est-à-dire fonction de l'ignorance. Si ce qui n'a pas été intégré au modèle joue un rôle important dans les données, alors les résidus vont le laisser apparaître et le modèle sera jugé comme non valide. Dans certains cas, les résidus peuvent ensuite être exploités pour guider l'analyste vers une modification du modèle. C'est ce qu'expose la deuxième section du chapitre 6.

Gardons à l'esprit ce à quoi peut servir l'analyse de résidus et revenons au problème de spécification de la distribution des effets aléatoires d'un modèle hiérarchique. Considérons la version de ce modèle dans laquelle les effets aléatoires sont supposés tous égaux. Cette version est appelé 'modèle de base'. Dans le modèle de base, les variations des effets aléatoires ne sont pas intégrés, une analyse des résidus de ce modèle doit donc permettre d'obtenir de l'information sur les effets aléatoires. L'information obtenue doit pouvoir ensuite être exploitée pour spécifier leur distribution. La troisième section du chapitre 6 détaille le principe de cette méthode. Ici, nous illustrons ce principe au travers d'un exemple et d'une analogie.

Exemple du flipper

On dit qu'un flipper "claque" lorsque le joueur atteint un score fixé. Un flipper qui claque fait gagner au joueur une nouvelle balle ou une nouvelle partie. Pour des raisons

[4] Expression extraite d'un texte introduisant à la peinture surréaliste à l'Ateneum Art Museum d'Helsinki. http ://www.ateneum.fi/.

économiques, un propriétaire de flipper souhaite savoir si le score auquel le flipper claque est assez élevé, c'est-à-dire si la probabilité que le flipper claque quand un joueur quelconque joue est assez basse. Si ce n'est pas le cas, il élèvera le score.

Pour étudier cette question, dix joueurs font dix parties. Notons Y_{ij} la variable indiquant si le joueur i a fait claquer le flipper à sa j-ème partie. Supposons, dans un premier temps, que les Y_{ij} sont indépendantes et identiquement distribuées selon une loi de Bernouilli de probabilité p

$$Y_{ij} \sim \text{Bernouilli}(p), \quad i = 1, \ldots, 10, j = 1, \ldots, 10. \tag{1.1}$$

Après l'estimation de p et une analyse des résidus pour voir si le modèle (1.1) est satisfaisant, on voit que les données reflètent un effet joueur : les joueurs ont des aptitudes inégales. Le modèle doit donc être complexifier. On propose le modèle hiérarchique suivant : les Y_{ij} sont indépendantes et distribuées selon des lois de Bernouilli de probabilité p_i

$$Y_{ij} \sim \text{Bernouilli}(p_i), \quad i = 1, \ldots, 10, j = 1, \ldots, 10. \tag{1.2}$$

En l'absence de connaissance sur les joueurs, les probabilités p_i sont vues comme des effets aléatoires. Pour aller plus loin dans l'analyse, il faut spécifier la distribution de ces effets. On sait que cette distribution doit avoir pour support $[0, 1]$ puisque les p_i sont des probabilités mais, à part cela, aucune connaissance ne permet de décider quel type de distribution leur attribuer.

Pour traiter ce problème de spécification de la distribution des effets aléatoires, les résidus calculés sous le modèle de base (1.1) qui n'intègrent pas les effets aléatoires peuvent être exploités. En effet, les résidus ordinaires entre les données et le modèle de base estimé sont $Y_{ij} - \hat{p}$ où $\hat{p}$ est l'estimateur de la probabilité p. S'il y a un effet joueur, il est évident que les $Y_{ij} - \hat{p}$ vont dépendre des p_i. Par exemple, si le joueur i_0 fait souvent claquer le flipper (p_{i_0} grand) alors les résidus ordinaires $Y_{i_0 j} - \hat{p}$ vont souvent être grands, et vice-versa. Ceci montre que les résidus du modèle de base contiennent de l'information sur les effets aléatoires et que l'on peut envisager d'exploiter ces résidus afin de spécifier la distribution des effets aléatoires.

Analogie avec la caractérisation des planètes extrasolaires

Les planètes extrasolaires sont des planètes gravitant autour d'étoiles autres que notre soleil. Il est difficile d'observer une planète extrasolaire de manière directe car son soleil est trop lumineux. Toutefois, des techniques basées sur la vitesse radiale ou encore sur le rayonnement lumineux de l'étoile ont été développées afin de les détecter et d'en déterminer des caractéristiques. La figure 1.3 montre les séries temporelles du rayonnement lumineux de l'étoile HD 209458 les 9 et 16 septembre 1999 (Charbonneau et al., 2000). Sous le modèle 'étoile seule', ces séries devraient être centrées autour de 1.00 le 9 septembre et autour de 0.95 le 16 (le bruit étant principalement dû au scintillement atmosphérique). Cependant, ces séries présentent chacune un décrochage qui est attribué au passage devant l'étoile d'une planète : en passant entre l'étoile et l'observateur, la planète réduit le rayonnement lumineux de l'étoile perçu par l'observateur.

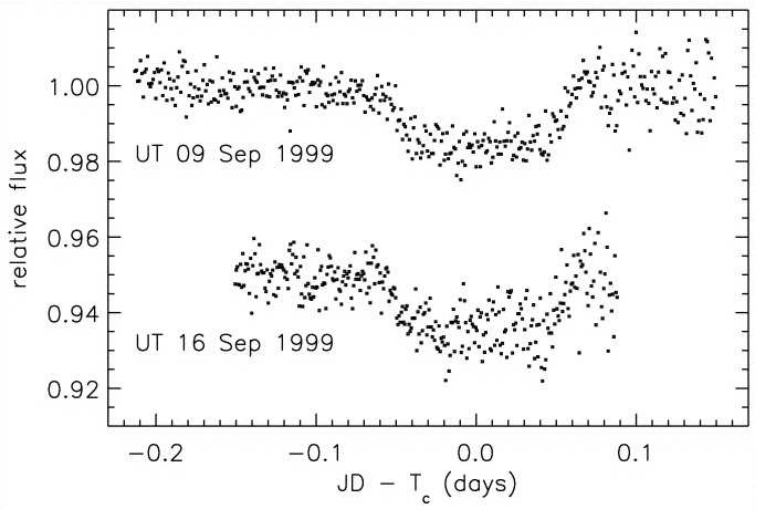


Fig. 1.3. Séries temporelles du rayonnement lumineux de l'étoile HD 209458 durant les 9 et 16 septembre 1999. Source : Charbonneau et al. (2000).

Ainsi l'élément 'planète' qui est non observé (ou caché) est détecté grâce à des écarts au modèle 'étoile seule' (les décrochages du rayonnement lumineux). Une fois que l'élément 'planète' est détecté, le modèle 'étoile + planète' prévaut. Mais comment déterminer les caractéristiques (masse, diamètre, distance à l'étoile) de cette planète qui reste cachée ? En fait, de l'information sur ces caractéristiques est contenue dans les écarts au modèle 'étoile seule', c'est-à-dire dans les décrochages du rayonnement lumineux de l'étoile. C'est en analysant ces écarts que Charbonneau et al. (2000) évaluent la masse, le diamètre et la distance à l'étoile de la planète HD 209458.

Faisons l'analogie entre ce contexte et le nôtre. Le modèle 'étoile + planète' est analogue à notre modèle hiérarchique. Le modèle 'étoile seule' est analogue à notre modèle de base. Tout comme Charbonneau et al. (2000) utilisent des écarts au modèle 'étoile seule' pour évaluer des caractéristiques de la planète extrasolaire qui est cachée, nous utilisons des résidus au modèle de base pour spécifier la distribution des effets aléatoires qui sont eux aussi cachés.

Une décomposition des résidus à la base de la méthode de spécification de la distribution des effets aléatoires

Savoir que les résidus du modèle de base contiennent de l'information sur les effets aléatoires ne suffit pas pour spécifier la distribution de ces effets. Pour ce faire, il faut donner le lien entre les résidus et les valeurs des effets aléatoires. Dans les chapitres 7 et 8, nous déterminons ce lien pour deux types de modèles hiérarchiques : l'un d'eux contient des effets aléatoires partagés, l'autre des effets aléatoires corrélés spatialement (dans ce cas les effets aléatoires forment un champ aléatoire caché). Dans les deux cas nous démontrons l'existence d'une décomposition asymptotique des résidus dont une des composantes fait intervenir de manière explicite les effets aléatoires. Cette décomposition nous permet d'estimer (ou restaurer) les valeurs des effets aléatoires puis, en se basant sur ces valeurs estimées, de

spécifier la distribution des effets. Dans les chapitres 7 et 8, des simulations sont effectuées pour évaluer l'efficacité de la méthode de spécification qui est ensuite appliquée à différents jeux de données (propagation à courte distance de la rouille brune, propagation à longue distance de la rouille jaune, radioactivité sur Rongelap Island, mortalité des vignes dans le Languedoc). La méthode de spécification est ensuite discutée au chapitre 9 d'un point de vue technique, et au chapitre 10 d'un point de vue pratique (exploitation de cette méthode pour l'analyse de la variabilité de la propagation).

Première partie

Modéliser et estimer la propagation spatiale des maladies aériennes des végétaux

2

Introduction à la propagation des maladies aériennes des végétaux : biologie et modélisation

La lutte contre les épidémies frappant les cultures constitue un enjeu central en agriculture. L'épidémiologie végétale apporte son concours à ce dessein par l'étude des mécanismes liant (i) la maladie étudiée, (ii) l'hôte, (iii) l'environnement et (iv) l'action humaine. En effet, c'est en comprenant les mécanismes d'une épidémie que l'on peut envisager de l'enrayer ou de la ralentir efficacement par l'emploi de moyens de lutte adaptés. Prenons l'exemple d'un champignon qui se transmet d'une année sur l'autre par les résidus de la plante qui sont laissés dans le champ après la récolte ; dans ce cas alterner les cultures permet d'enrayer l'épidémie car le champignon présent sur un hôte une année donnée ne trouve pas l'année suivante cet hôte lui permettant de se développer. Prenons maintenant le cas d'une maladie à dissémination aérienne ; dans ce cas, l'alternance des cultures ne permet pas d'enrayer l'épidémie puisque la maladie peut venir d'un champ distant. Ainsi, acquérir de la connaissance sur la propagation d'une maladie, sur sa capacité de dissémination à diverses distances, doit aider à déterminer des moyens de lutte adaptés.

Ce chapitre décrit les deux maladies à dissémination aérienne, les rouilles brune et jaune du blé, qui sont étudiées dans ce manuscrit. Il présente des modèles existants qui peuvent être utilisés pour décrire la propagation des maladies des végétaux. Les limites de ces modèles sont discutées. Puis, afin de développer des modèles contournant ces limites, des principes de modélisation sont énoncés.

2.1 Deux exemples de maladies aériennes des végétaux : les rouilles brune et jaune du blé

2.1.1 Impact économique des rouilles du blé

Le blé est l'une des principales céréales dans l'alimentation de l'humanité. Avec 16% de la production annuelle mondiale entre 1995 et 1999, l'Europe des 15 était durant cette période le premier producteur mondial de blé ; en 1998, la France a exporté 15 des 40 millions de tonnes de blé produites (Bonjean and Angus, 2001).

L'importance de la culture du blé aux niveaux alimentaire et économique justifie la lutte contre les maladies frappant cette culture. Parmi ces maladies, les rouilles brune et jaune

sont les plus importantes au niveau mondial. Cependant, les dégâts peuvent être inégaux : leur effet est évalué entre 5 et 15% de baisse de rendement au Canada et entre 3 et 5% en Europe de l'Est.

Face aux rouilles, les acteurs de la filière blé balancent entre (i) accepter des pertes de rendement considérables et (ii) payer le coût de la tranquillité en traitant les champs avec des fongicides (au prix de problèmes environnementaux) et/ou en recherchant des variétés de blé résistantes aux rouilles.

2.1.2 Description sommaire de la biologie des rouilles

La biologie et l'épidémiologie des rouilles des céréales en général et du blé en particulier sont l'objet d'ouvrages spécifiques (Bushnell and Roelfs, 1984; Roelfs and Bushnell, 1985; Roelfs et al., 1992). Des ouvrages de référence récents (Lucas, 1998; Lepoivre, 2003; Agrios, 2005) donnent un aperçu plus large sur les maladies des plantes. Il existe enfin des ouvrages ou chapitre d'ouvrage traitant de l'épidémiologie de ces maladies (Campbell and Madden, 1990; Rapilly, 1991; Jones, 1998; Kranz, 2003; Sache, 2003).

Les rouilles brune (*Puccina triticina* ou leaf rust ou brown rust) et jaune (*Puccina striiformis* ou stripe rust ou yellow rust) du blé sont des champignons biotrophes (qui se développent sur du tissu vivant) dont les symptômes sont des lésions formées sur les feuilles de la plante. A l'intérieur de ces lésions, sont produites des spores qui sont les agents de la dissémination de la maladie. En effet, une fois produite une spore peut être libérée par le vent ou la pluie, transportée dans l'air puis déposée. Si les conditions sont favorables, une spore déposée sur une feuille de blé germe et pénètre le tissu foliaire et donne une nouvelle lésion. Cet enchaînement constitue un cycle épidémique. La succession de plusieurs de ces cycles constitue une épidémie.

La durée d'un cycle épidémique et l'évolution d'une épidémie dépendent grandement de l'état sanitaire du couvert végétal et des conditions météorologiques. Un couvert végétal vigoureux est un terrain favorable pour l'épidémie de rouille. La température et l'humidité (ou l'eau libre) sont les facteurs météorologiques principaux modulant l'épidémie. Les conditions météorologiques optimales pour le développement de la maladie diffèrent entre rouilles jaune et brune. Globalement, la rouille jaune est favorisée par les climats de types océanique, doux et humide ; la rouille brune est adaptée à des environnements climatiques plus contrastés, elle se développe également sous les climats de type continental et méditerranéen qui sont en général moins propices au développement de la rouille jaune.

Les lésions de la rouille brune sont de petits disques dont le diamètre est de l'ordre du millimètre ; en revanche, les lésions de la rouille jaune croissent le long des nervures de la feuille et peuvent s'étendre sur une surface de quelques centimètres carrés (Berger et al., 1997). Le diamètre d'une spore de rouille brune ou jaune varie entre 16 et 30 μm (Eversmeyer and Kramer, 2000). Cependant, les spores de la rouille brune sont dispersées seules alors que les spores de la rouille jaune sont dispersées en amas de quelques spores à cause d'une

couche mucilagineuse les entourant (Rapilly and Fournet (1968) cité par Schermesser (1996) et Geagea et al. (1999)).

2.1.3 Une définition de la propagation spatiale

Les rouilles peuvent se propager entre feuilles voisines, entre plantes d'un même champ, entre champs, et même entre continents, et ce suite à un seul ou plusieurs cycles épidémiques. La propagation au sens large est comprise comme l'avancement dans l'espace de la maladie durant un ensemble de cycles épidémiques. **Nous définissons la propagation spatiale comme l'avancement dans l'espace de la maladie au cours d'un seul cycle épidémique**. Dans les chapitres 3, 4 et 5, nous nous intéressons à la propagation uniquement spatiale des rouilles qui est le moteur de l'épidémie (c'est ce qui fait qu'une maladie se propage).

La propagation spatiale est le résultat (i) de la production des spores dans les lésions mères, (ii) de la dispersion des spores, et (iii) de l'infection de plantes de blé par les spores, infection qui a pour conséquence l'apparition de lésions filles

$$\text{Propagation spatiale} = \text{Production} + \text{Dispersion} + \text{Infection}.$$

Dans cette décomposition de la propagation spatiale, la dispersion est le résultat de la libération, du transport et du dépôt des spores. Les principaux facteurs de la dispersion sont le vent et la pluie. Le vent peut libérer les spores produites par une lésion grâce à des rafales. Il peut transporter les spores sur de courtes distances (Aylor, 1987), principalement grâce aux turbulences qu'il crée au niveau du couvert végétal, et sur de longues distances (Sackett and Mundt, 2005), voire sur des distances continentales (Brown and Hovmøller, 2002; Hovmøller et al., 2002). Campbell and Madden (1990) traitent de la dispersion sur l'ensemble de ces distances. La pluie libère et transporte les spores sur de courtes distances par impaction et/ou par *splashing* (Geagea et al., 1999, 2000). Elle permet également de lessiver l'air des spores qu'il contient et de déposer ces spores sur le couvert végétal (Sache et al., 2000). Aylor (1978) et McCartney and Fitt (1998) décrivent plus précisément les mécanismes biophysiques intervenant dans la dispersion des spores par le vent et par la pluie.

2.1.4 Propagation spatiale et lutte contre les épidémies

La propagation spatiale rend compte de la capacité de dissémination de la maladie à différentes distances. Mieux comprendre ce qui se passe à chaque distance, c'est à dire mieux comprendre quels mécanismes prédominent aux différentes échelles spatiales, est important dans l'optique de la lutte contre l'épidémie. En effet, aujourd'hui on lutte contre les maladies aériennes des végétaux principalement en utilisant des variétés résistantes et des applications étendues et préventives de fongicides. Parce que les résistances peuvent être contournées et parce que les fongicides sont polluants et peuvent voir leur efficacité diminuée suite à la sélection de populations pathogènes résistantes, d'autres mesures de lutte sont étudiées ou

appliquées : les mélanges variétaux[1], la gestion multi-parcellaire des variétés[2], les traitements locaux[3]. L'efficacité de ces mesures de lutte dépend de la capacité de dissémination du pathogène à différentes distances. L'évaluation des effets de ces mesures nécessite donc une connaissance multi-échelle de la propagation spatiale.

2.2 Modèles existants permettant de décrire la propagation spatiale

Dans cette section nous présentons trois types de modèles de propagation spatiale. Ces modèles détaillent plus ou moins les étapes production, dispersion, infection. Certains peuvent être ajustés aux données de propagation, d'autres sont utilisés pour prédire et simuler. Dans la section 2.3, nous donnerons les limites de ces modèles de propagation spatiale au vu des objectifs que nous nous fixons.

2.2.1 Modèles pour le gradient de maladie

La propagation à partir d'une source ponctuelle de spores a été décrite par le gradient de maladie noté g. Donnons la définition de g. Soit $Q(x, y)$ une mesure quantitative de la maladie présente sur un hôte situé en (x, y) et due aux spores émises par la source, la valeur $g(x, y)$ du gradient de maladie est l'espérance de $Q(x, y)$. La mesure quantitative de la maladie sur l'hôte est, par exemple, le nombre de lésions sur une feuille ou encore le pourcentage de surface foliaire malade sur une feuille. Dans la littérature classée 'épidémiologie végétale', c'est généralement la version 1D du gradient de maladie qui est utilisée pour décrire la propagation (McCartney and Fitt, 1985; Aylor, 1990) : $g(r)$ est l'espérance de $Q(x, y)$ où $\sqrt{x^2 + y^2} = r$.

Des modèles paramétriques ont été proposés pour le gradient de maladie g. Les deux modèles les plus utilisés (McCartney and Fitt, 1985; Aylor, 1990) sont le gradient exponentiel sous sa forme 1D

$$g(r) = a \exp(-r/b)$$

et le gradient loi-puissance sous sa forme 1D

$$g(r) = ar^{-b},$$

où a et b sont deux paramètres positifs. Le gradient loi-puissance est également appelé gradient de Gregory (1968) ou de Pareto (Minogue, 1989). D'autres modèles paramétriques

[1] Le mélange variétal consiste à intégrer dans les semences diverses variétés de blé dont les résistances sont différentes dans le but de ralentir l'épidémie. En fait l'idée est que la contamination de proche en proche s'opère plus difficilement. De plus on peut espérer que l'une des variétés ne soit pas ou ne soit que partiellement attaquée et ainsi garantir une partie de la récolte.

[2] La gestion multi-parcellaire des variétés consiste à semer des variétés caractérisées par des résistances différentes dans des champs différents. Un agencement réfléchi des variétés dans les champs doit permettre de réduire le risque qu'une épidémie saute de champs en champs.

[3] Le traitement local consiste à appliquer un fongicide sur un foyer de maladie afin de ralentir l'épidémie. Si la dispersion des spores se fait principalement à courte distance, alors traiter le foyer peut grandement diminuer la pression d'inoculum.

plus flexibles ont été proposés dans le but de trouver des formes ayant la propriété de mieux s'ajuster aux données de propagation. On peut citer par exemple le gradient de Mundt and Leonard (1985, $g(r) = a(c + r)^{-b}$) qui généralise le gradient loi-puissance, et le gradient exponentiel-puissance qui généralise le gradient exponentiel ($g(r) = a\exp(-br^c)$, Sackett and Mundt, 2005).

Ces modèles sont simples et leurs paramètres peuvent être estimés en ajustant les modèles à des données expérimentales de propagation. Les expériences de propagation consistent généralement (Gregory, 1968; Aylor, 1987; Sackett and Mundt, 2005)

- à placer une source de spores (une plante ou un groupe de plantes volontairement infectées et donc sur lesquelles il y a des lésions mères) dans une parcelle expérimentale, et
- à mesurer, après que les lésions filles soient apparues et à différentes distances de la source, la quantité de maladie.

Les données sont donc des mesures quantitatives de la maladie à différentes distances de la source. La mesure utilisée est généralement une mesure agrégée (e.g. nombre total de lésions filles sur 21 feuilles, Aylor, 1987). Le nombre de données varie généralement entre 4 et 10 (Gregory, 1968; Aylor, 1987; Fitt et al., 1987; Mundt, 1989; Sackett and Mundt, 2005). L'ajustement d'un gradient de maladie aux données, i.e. l'estimation de ses paramètres, est généralement opéré en linéarisant le modèle[4] et en utilisant le critère des moindres carrés (régression linéaire ordinaire).

Les paramètres des gradients de maladie n'ont pas de sens physique. Cependant ils donnent des indications sur la puissance de la source de spores (capacité de production) combinée à la probabilité d'infection des spores au travers du paramètre a, et sur la décroissance de la concentration de maladie avec la distance au travers du paramètre b. Ainsi, à l'aide du gradient de maladie, on peut comparer les capacités de dissémination de deux maladies ou d'une même maladie dans deux situations différentes[5].

2.2.2 Modèles de simulation

La propagation spatiale à partir de sources réparties dans l'espace a été décrite par des modèles où sont explicités les trois étapes de la propagation spatiale (production de spores, dispersion des spores et infection des plantes par les spores). Ces modèles sont intégrés à des modèles spatio-temporels construits pour simuler des épidémies (e.g. Minogue and Fry, 1983; Lett and Østergård, 2000).

Exemple

Présentons la composante 'propagation spatiale' du modèle de Lett and Østergård (2000).

[4] Par exemple, pour l'exponentielle : $\log\{g(r)\} = \log a - r/b$, pour la loi-puissance : $\log\{g(r)\} = \log a - b\log r$.
[5] Capacité de dissémination dans des directions différentes ou sur des variétés différentes de l'hôte.

Production : la quantité potentielle de spores qui sont produites par une lésion en un jour et qui conduiraient à des lésions filles si le champ était infini et sain, suit une loi de Poisson de moyenne α. Le paramètre α est appelé facteur de multiplication journalier potentiel de la maladie.
Dispersion : chaque spore produite est déposée dans le plan

– à une distance r de la lésion mère distribuée selon la densité

$$f(r) = \frac{2}{\pi\mu}\frac{1}{1+(r/\mu)^2}\delta_{r\geq 0}$$

où $\delta_{r\geq 0}$ vaut 1 si $r \geq 0$ et 0 sinon,

– dans une direction uniformément distribuée dans $[0^o, 360^o[$,

et ce indépendamment des autres spores.
Infection : quand une spore est déposée sur la plante i, elle donne une lésion avec la probabilité $1 - y_i/y_{max}$ où y_i est la surface foliaire atteinte de la plante et y_{max} est la surface foliaire totale de la plante ($y_{max} - y_i$ est la surface foliaire libre). Si la spore donne une lésion, la surface foliaire atteinte y_i est incrémentée de la surface s occupée par une lésion.

Les paramètres de ce modèle sont, pour une maladie donnée, mesurés par ailleurs (utilisation de résultats donnés dans la littérature).

Choix de la fonction de dispersion (FDD)

La fonction de dispersion (FDD), notée $h : \mathbb{R}^2 \to \mathbb{R}$, est définie comme la densité de probabilité de la position aléatoire de dépôt dans $\mathbb{R}^2$ d'une spore émise par une source ponctuelle située en l'origine. Les modèles de simulation intègrent une FDD de manière explicite[6]. Par exemple, on peut montrer (Tufto et al., 1997) que la FDD intégrée au modèle de Lett and Østergård (2000) est

$$h(x,y) = \frac{1}{2\pi r}\frac{2}{\pi\mu}\frac{1}{1+(r/\mu)^2},$$

où (x, y) est dans le plan $\mathbb{R}^2$ et $r = \sqrt{x^2+y^2}$

Le choix d'une FDD contraint la répartition spatiale de la maladie. Diverses formes ont donc été proposées pour h afin de refléter des décroissances différentes de la concentration de spores avec la distance et afin de refléter l'anisotropie de la dispersion des spores. Nous présentons ici certaines de ces formes.

Modèles empiriques pour la FDD

Des densités de probabilité paramétriques ont été proposées pour la FDD h. Par exemple, Tufto et al. (1997) présentent la fonction exponentielle-puissance

$$h(x,y) = \frac{c}{2\pi b^{2/c}\Gamma(2/c)}\exp\left(-\frac{r^c}{b}\right)$$

[6] les modèles pour le gradient de dispersion intègrent également une FDD, mais de manière implicite : on considère généralement que le gradient de maladie est proportionnel à la FDD.

qui généralise la fonction exponentielle ($c = 1$)

$$h(x, y) = \frac{1}{2\pi b^2} \exp(-r/b),$$

la fonction de Weibull

$$h(x, y) = \frac{c\ r^{c-2}}{2\pi b} \exp\left(-\frac{r^c}{b}\right)$$

qui généralise la fonction gaussienne ($c = 2$ et $\sigma^2 = b/2$)

$$h(x, y) = \frac{1}{2\pi\sigma^2} \exp(-r^2/2\sigma^2),$$

où b et c sont des paramètres strictement positifs. Les différents modèles qui ont pu être proposés varient selon le type de la décroissance (décroissance à l'origine plus ou moins forte, queue de distribution plus ou moins lourde). L'idée est que les caractéristiques physiques de la spore (ou de l'amas de spores dans le cas de la rouille jaune) jouent sur la décroissance de la FDD.

Bien que Minogue (1989) prête aux FDD exponentielle et loi-puissance des interprétations physiques, bien que Bicout and Sache (2003) montrent que la FDD exponentielle peut être dérivée, sous certaines conditions, du comportement des spores dans l'atmosphère, ces formes paramétriques ont d'abord été choisies pour leur simplicité mathématique. On parle de modèles empiriques. Notons que toutes les FDD précédentes sont des fonctions isotropes. Les modèles suivants permettent d'introduire de l'anisotropie dans la fonction de dispersion.

Modèles quasi-mécanistes pour la FDD

Tufto et al. (1997), Stockmarr (2002), Klein et al. (2003) et Bicout and Sache (2003) proposent des modèles paramétriques pour la FDD qui sont dérivés analytiquement de modèles quasi-mécanistes[7] décrivant le comportement de particules (e.g. de spores) dans l'atmosphère. Ces modèles sont dits quasi-mécanistes car seuls les mécanismes majeurs auxquels sont soumises les particules sont pris en compte. Par exemple, les mouvements des spores sont modélisés par des mouvements browniens tridimensionnels avec tendances[8] (*drifts*), et les temps de déposition des spores sont modélisés par des temps d'arrêt[9]. Ne prendre en compte que les mécanismes majeurs permet d'obtenir une fonction de dispersion ne contenant que peu de paramètres (à peu près autant que les modèles empiriques). Ainsi, sous certaines hypothèses, Tufto et al. (1997) obtiennent la fonction de dispersion suivante

$$h(x, y) = \frac{1}{2\pi r\sqrt{\gamma}} \exp\left\{\frac{1}{(\tau_x x + \tau_y y)\gamma} - \left(\frac{1}{\gamma} + \frac{\tau_x^2}{\gamma^2} + \frac{\tau_y^2}{\gamma^2}\right)^{1/2} r\right\},$$

où $r = \sqrt{x^2 + y^2}$ et γ, τ_x et τ_y sont des paramètres fonctions des tendances des mouvements browniens et de la loi du temps de déposition des spores. Dans l'équation ci-dessus, le premier terme dans l'exponentielle permet de modéliser une anisotropie de la dispersion.

[7] La formulation est suggérée par Klein et al. (2003).

[8] Les tendances servent à décrire l'effet moyen du vent horizontal ou la gravité.

[9] Le temps d'arrêt peut être le moment où la trajectoire de la particule rencontre une hauteur donnée (la hauteur des maïs femelles dans Klein et al. (2003)).

Modèles physiques ou mécanistes pour la FDD

McCartney and Fitt (1985) et Aylor (1990) présentent trois modèles physiques ou mécanistes : le modèle de panache gaussien, le modèle gradient-diffusion et le modèle lagrangien de marche aléatoire, qui peuvent être utilisés pour décrire la dispersion (libération, transport et dépôt) des spores. Les deux premiers modèles correspondent à l'approche eulérienne : l'évolution dans le temps de la concentration de spores est décrite en tout point de l'espace. Le dernier modèle correspond à l'approche lagrangienne : les trajectoires des spores dans l'atmosphère sont décrites. Ces modèles détaillent nombre de mécanismes auxquels sont soumises les spores. Leurs paramètres ont un sens physique et sont mesurés expérimentalement. Les FDD dérivées de ces modèles physiques ont soit des expressions analytiques très lourdes avec de nombreux paramètres, soit des formes numériques obtenues par simulation (Klein et al., 2003).

2.2.3 Modèles stochastiques à potentiel

La propagation spatiale à partir de sources réparties dans l'espace a été décrite par des modèles stochastiques à potentiel dont les paramètres peuvent être estimés. Décrivons-en le principe et donnons-en deux exemples.

Principe des modèles stochastiques à potentiel

Considérons I individus répartis sur un réseau spatial. Dans le modèle, un individu peut être sain (état 0) ou atteint par la maladie m parmi M maladies (états $1, \ldots, M$) ; un individu atteint par la maladie $m \in \{1, \ldots, M\}$ ne peut plus évoluer (les états $1, \ldots, M$ sont absorbants). La figure 2.1 montrent les transitions possibles entre états pour un individu donné.

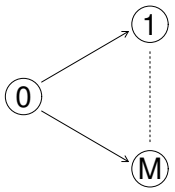


Fig. 2.1. Graphe des transitions entre états pour les modèles stochastiques à potentiel.

On attribue, à chaque individu i sain (état 0) au temps $t-1$, un potentiel $W_{i,u}^t$ d'être dans l'état $u \in \{0, \ldots, M\}$ au temps suivant t. Le potentiel $W_{i,u}^t$ dépend des états pris par les autres individus entre le temps initial 0 et le temps t :

$$W_{i,u}^t = \alpha(u) + \sum_{\substack{j=1 \\ j \neq i}}^{I} \beta\{u, U_j^{0:t}, d(i,j)\},$$

où $U_j^{0:t} = \{U_j^0, \ldots, U_j^t\}$ est le vecteur des états pris par l'individu j entre les temps 0 et t, et $d(i,j)$ est une distance entre les individus i et j. Les termes $\alpha(u)$, $u = 1, \ldots, M$, sont des potentiels de base liés aux probabilités conditionnelles qu'un individu soit attaqué par les différentes maladies sachant que tous les autres individus sont sains. Ces termes permettent de prendre en compte le risque d'être infecté par des sources de maladie autres que les individus $1, \ldots, I$ (attaque exogène). Les $\beta\{u, U_j^{0:t}, d(i,j)\}$ sont des termes d'interaction entre individus distants. Si $U_j^{0:t-1}$ est non nul (i.e. si l'individu j a été malade avant la date t), alors $\beta\{u, U_j^{0:t}, d(i,j)\}$ traduit la pression infectieuse générée par j sur l'individu i. Si $U_j^{0:t-1}$ est nul et U_j^t est non nul (i.e. si l'individu j devient malade à la date t), alors $\beta\{u, U_j^{0:t}, d(i,j)\}$ est une mesure de la concomitance[10] d'une attaque sur i et j entre $t-1$ et t. Ainsi construit, le potentiel reflète la production de l'inoculum exogène (les potentiels de base α) et endogène (les interactions β) et la dispersion de la maladie (variations de $\beta\{u, U_j^{0:t}, d(i,j)\}$ quand la distance $d(i,j)$ varie).

La probabilité conditionnelle de trouver l'individu i dans l'état u au temps t sachant qu'il était sain à $t-1$ et connaissant les états $U_j^{0:t}$ est définie par

$$\mathbb{P}[U_i^t = u \mid U_i^{t-1} = 0, \{U_j^{0:t} : j \neq i\}] = \frac{\exp(W_{i,u}^t)}{\sum_{u'=0}^{M} \exp(W_{i,u'}^t)}.$$

L'ensemble des probabilités conditionnelles reflètent l'infection.

Exemple 1

Chadœuf et al. (1992) proposent un modèle de Gibbs spatio-temporel pour décrire les épidémies de M maladies sur un réseau carré (application : propagation de deux champignons sur une plantation d'hévéas). Dans ce modèle, $\alpha(0)$ vaut 0 par convention, et $\alpha(1), \ldots, \alpha(M)$ sont des paramètres inconnus. Par convention également, le terme d'interaction $\beta\{u, U_j^{0:t}, d(i,j)\}$ vaut 0 si on s'intéresse au potentiel pour i de rester sain ($u = 0$) ou si le voisin j est resté sain ($U_j^0 = \cdots = U_j^t = 0$) ou si i et j ne sont pas voisins sur le réseau. Dans les autres cas, le terme d'interaction $\beta\{u, U_j^{0:t}, d(i,j)\}$ prend une valeur qui dépend de la maladie u considérée, de la maladie qui touche le voisin j de i, et du temps depuis lequel ce voisin est malade. Les termes d'interactions non nuls sont des paramètres inconnus. Dans ce modèle, la dispersion ne s'effectue qu'entre voisins directs. Chadœuf et al. (1992) proposent une méthode permettant d'estimer les potentiels de base et les paramètres d'interaction, et de tester l'égalité des différents paramètres. Les tests servent à comparer, pour deux maladies différentes, leur pouvoirs infectieux (liés aux capacités de production des individus malades) production de l'inoculum (pouvoir infectieux) ou la capacité de dispersion de deux maladies.

Exemple 2

Chadœuf et Calonnec (communication personnelle) proposent un modèle stochastique à potentiel dans lequel chaque individu malade génère en tout point de l'espace un potentiel

[10] Les individus i et j peuvent être attaqués en même temps ou bien l'un deux avoir été attaqué puis avoir transmis la maladie à l'autre entre les deux dates d'observation $t-1$ et t.

infectieux qui décroît avec la distance entre le point et l'individu malade. (application : propagation de l'oïdium dans une parcelle de vigne). Dans ce modèle il n'y a qu'une maladie ($M = 1$) et donc seulement deux états (sain : 0 et malade : 1). Par convention $\alpha(0) = 0$, et $\alpha(1)$ est un paramètre inconnu représentant le risque, constant dans la parcelle, d'être infecté par une source extérieure. D'autre part, les paramètres d'interaction valent

$$\beta\{u, U_j^{0:t}, d(i,j)\} = \delta_{U_j^t=1} \frac{b}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{d(i,j)^2}{2\sigma^2}\right\},$$

où b et σ sont des paramètres inconnus, et $d(i,j)$ est la distance euclidienne entre i et j, et $\delta_{U_j^t=1}$ vaut 1 si $U_j^t = 1$ et 0 sinon. b reflètent la production d'inoculum par un individu malade et $(\sqrt{2\pi}\sigma)^{-1} \exp\{-d(i,j)^2/2\sigma^2\}$ reflètent la dispersion. Dans ce modèle, la dispersion a lieu à toutes les distances (pas seulement entre voisins). Chadœuf et Calonnec proposent une méthode permettant d'estimer les paramètres.

2.3 Nos principes de modélisation

Expériences et objectifs

Des expériences de propagation des rouilles ont été réalisées par les scientifiques du laboratoire d'Epidémiologie Végétale et Ecologie des Populations (INRA Versailles-Grignon). L'objectif général de l'ensemble de ces expériences était de mieux comprendre comment se développent les épidémies des maladies aériennes des végétaux cultivés. Afin de mieux comprendre quels mécanismes sont prépondérants à quelles échelles, trois séries différentes d'expériences ont été réalisées :

- propagation spatiale à courte distance (moins de 50 centimètres) de la rouille brune,
- propagation spatio-temporelle à moyenne distance (de 1 à 25 mètres) des rouilles brune et jaune, et
- propagation spatiale à longue distance (de 10 à 200 mètres) de la rouille jaune.

Ces trois échelles d'observation offrent des visions différentes de la propagation. Les expériences de propagation spatiale à courte distance témoignent notamment de la dispersion des spores entre feuilles proches et de l'hétérogénéité des feuilles en terme de propension à être infectée. Les expériences de propagation spatio-temporelle à moyenne distance témoignent notamment de l'itération des cycles épidémiques et du développement de foyers secondaires de maladie. Les expériences de propagation spatiale à longue distance témoignent notamment de l'anisotropie de la dispersion des spores.

Pour mieux comprendre le développement épidémique des maladies aériennes des végétaux, nous voulons développer un cadre générique[11] de modélisation/estimation permettant de remplir les objectifs suivants

1. estimer la propagation des maladies aériennes des végétaux à partir des données issues des expériences mentionnées ci-dessus,

[11] Le terme "un cadre générique" doit être compris comme "un cadre adapté non pas seulement aux rouilles du blé mais plus largement aux maladies aériennes des végétaux".

2. analyser la variabilité des données, et
3. intégrer les différentes échelles spatiales mises en jeu.

Limites des modèles existants au vu des objectifs précédents

Le gradient de maladie décrit la tendance moyenne de la propagation. Il ne permet donc pas de décrire la variabilité autour de cette tendance. Les modèles de simulation, eux, peuvent rendre compte d'une part de la variabilité autour de la tendance moyenne. Cependant, leurs paramètres ne sont pas estimés à partir de données de propagation, mais déterminés par ailleurs. Les paramètres des modèles stochastiques à potentiel, eux, peuvent être estimés à partir des données de propagation. Mais, ces modèles ne décrivent que des variables absence/présence de maladie. De plus, aucun de ces modèles n'offrent réellement une approche multi-échelle de la propagation.

Principes de modélisation

Compte-tenu des objectifs énoncés ci-dessus et des limites des modèles existants pour remplir ces objectifs, nous proposons des modèles qui à la fois, (i) sont assez simples afin d'être ajustés aux données expérimentales, (ii) partagent des caractéristiques communes afin d'intégrer et comparer les échelles, (iii) présentent des caractéristiques différentes afin de refléter des sources de variabilité des données propres à chaque échelle considérée. Plus précisément, nous avons appliqués les principes suivants de modélisation.

Principe 1. Les modèles proposés doivent être assez simples (nombre limité de paramètres) et doivent être adaptés à la nature des observations pour qu'ils puissent être ajustés aux données expérimentales de propagation.

Principe 2. Les modèles proposés doivent décrire les trois étapes de la propagation spatiale (production des spores, dispersion des spores et infection des feuilles par les spores) pour qu'ils puissent rendre compte de la variabilité des données due à chacun de ces trois temps.

Principe 3. Les modèles proposés pour les différents types d'expériences doivent partager des composantes communes afin de permettre l'intégration des différentes échelles et de faciliter une vision multi-échelle de la propagation spatiale.

Principe 4. Les modèles proposés pour les différents types d'expériences doivent se distinguer en intégrant des composantes différentes. Ceci permettra de marquer ce que l'on voit et ce que l'on ne peut voir selon l'échelle considérée.

Dans les chapitres 3 et 4, nous analysons les données de propagation spatiale à courte et longue distances en construisant deux modèles basés sur les quatre principes précédemment énoncés. Les paramètres de ces modèles sont estimés par maximum de vraisemblance.

3

Propagation spatiale à courte distance de la rouille brune

Une série d'expériences a été réalisée afin de mieux comprendre les premiers stades du développement d'un foyer de rouille brune du blé. Dans cette série, des expériences de propagation spatiale à courte distance (de 0 à 50 centimètres) à partir d'une feuille infectée (source de spores) ont été menées afin, principalement, d'évaluer la décroissance de la concentration de maladie due à la source quand on s'éloigne de celle-ci. La concentration de maladie a été mesurée en comptant le nombre de lésions sur chaque feuille avoisinant la source de spores.

Les données montrent, comme attendu, une décroissance du nombre de lésions par feuille avec la distance à la source. Elles montrent également une variabilité locale extra-poissonienne possiblement causée par l'hétérogénéité des feuilles en terme de propensions à être infectées. Nous proposons un modèle de fragilité (*frailty model*) qui permet de décrire la décroissance du nombre de lésions par feuille avec la distance, tout en tenant compte de la sur-dispersion des données. Dans ce modèle, une fonction déterministe appelée 'potentiel infectieux' intègre la force[1] de la source et la dispersion des spores ; des effets aléatoires (fragilités) modélisent les propensions non observées des feuilles à être infectées ; le nombre de lésions sur une feuille suit une loi de Poisson dont la moyenne est le produit entre la valeur locale du potentiel infectieux et la fragilité (non observée) de la feuille.

Les paramètres du modèle ont été estimés par maximum de vraisemblance pour cinq jeux de données. Des tests ont rejeté l'égalité des potentiels infectieux pour les 5 jeux de données, mais n'ont pas rejeté l'égalité des densités des fragilités, indiquant que l'hétérogénéité des feuilles induit la même variabilité quelque soit le jeu de données considéré.

Le modèle estimé a ensuite été confronté aux données. Une grande part de la variabilité des données est capturé par le modèle de fragilité. Toutefois, l'étude des écarts entre modèle et données suggère notamment que les fragilités pourraient être structurées spatialement. L'étude de la structuration des fragilités est poursuivie dans la section 7.5 de ce manuscrit.

[1] Capacité de production de spores

A frailty model to assess plant disease spread from individual count data

By Samuel Soubeyrand, Ivan Sache, Christian Lannou and Joël Chadœuf.

Spread of airborne plant diseases from a propagule source is classically assessed by fitting a disease gradient to aggregated data coming from field experiments. But, aggregating data decreases information about processes involved in disease spread. To overcome this problem, individual count data can be collected ; it was done in the case of short-distance spread of wheat brown rust. However, for such data, the disease gradient is a limited model since heterogeneity of hosts is ignored. So, we propose a parametric frailty model in which the frailties represent propensities of hosts to be infected. The model is used to assess dispersal of propagules and heterogeneity of hosts. A model check leads us to address questions about assumptions made in the model.

3.1 Introduction

In botanical epidemiology, assessing spread of airborne diseases of plants is of major concern (Aylor, 1990; Campbell and Madden, 1990; Fitt et al., 1987; McCartney and Fitt, 1998). It contributes to understand dynamic of epidemics and, consequently, to assess disease impact on crop growth and crop yield. The spreading process of diseases of interest can be described as follows. Propagules are produced at a given location. Generally because of wind and/or rain, they are released, transported and deposited on other areas (propagule dispersal process). When conditions are conducive, some of the deposited propagules succeed in infecting hosts (host infection process). Disease spread is so the result of both propagule dispersal and host infection processes. Propagule dispersal is well studied, whereas host infection is often ignored because it depends on hardly-observable host features influencing propensities of hosts to be infected. For the brown rust of wheat for example, the infection of a leaf by a spore depends on the physiological state of the leaf (hydric status, nitrogen content) and on the microclimate at the leaf scale (temperature, wetness) which are difficult to measure in field experiments.

To assess disease spread from a field experiment, an exponential or power-law disease gradient is commonly fitted to aggregated data, i.e. disease measures done on sets of hosts (Aylor, 1987; Fitt et al., 1987). To better understand processes involved in disease spread, the epidemiologist can use individual data, i.e. disease measures done on individual hosts. However, in this case, the disease gradient is a limited model because it does not include

heterogeneity of hosts which can cause overdispersion of individual data and can lead to misleading inference for the parameters of the disease gradient (Hinde and Demétrio, 1998).

In this paper, we propose a frailty model to describe individual count data

in the disease spread context. A deterministic parametric function models the expected dispersal of propagules, and frailties are included to model heterogeneity of hosts. Frailty models are usually developed in survival analysis (Nielsen et al., 1992), but our frailty model is adapted to the disease spread context. In particular, the frailty is viewed as a weight in [0,1] characterizing a host and, consequently, cannot obeys a classical frailty distribution, that is the mathematically convenient gamma distribution or the log-normal distribution whose supports are $\mathbb{R}^+$. So, we use a parametric distribution in [0,1]. The frailties, which are assumed to depend on biological characteristics at the leaf scale, are assumed to be independent and identically distributed. Estimating the model allows us to quantify the propagule dispersal process and the host infection process. In addition, checking the model allows us to detect sub-processes possibly involved in disease spread and not included in the model. For instance, it is suggested that the host frailties are slightly spatially dependent. In the discussion, we suggest an approach to learn about this spatial dependence via model residuals.

The dataset we consider comes from a field experiment conduced to assess short-distance spread of wheat brown rust ; Section 3.2.1 details the experiment and Section 3.2.2 provides biological aspects of heterogeneity of hosts. Section 3.3 presents the frailty model. Section 3.4 derives maximum likelihood estimators for the parameters ; their uncertainties are assessed by using a normal approximation and parametric bootstrap. Model parameters are estimated in Section 3.5. Section 3.6 checks the model : model outputs and residuals are analyzed with plots and tests. Section 3.7 discusses the analysis from both the biological and the statistical points of view.

3.2 Biological content

In the experiment described in the next paragraph, short-distance spread of wheat brown rust was measured to better understand local epidemic spread and pathogen lesion distribution within a field crop (Robert, 2003). Short-distance spread of wheat brown rust was already measured by Aylor (1987) : he counted lesions on sets of plants (aggregated data). In contrast, we counted lesions on individual leaves (individual data). With such individual count data, we expected (i) to get more accurate estimators for the parameters of the dispersal function, (ii) to quantify the variability of data due to leaf-scale variations of leaf conditions and, consequently, (iii) to gain insight into disease spread.

3.2.1 Field experiment

An experimental field of wheat was sown in October 2001. Its length was 30 m and it contained 9 rows 18.4 cm apart (row -4 to row +4 in the top panel of figure 3.1). Within

this field, 14 flag leaves, lined up along row 0 every two meters, were inoculated with brown rust. The flag leaf of a wheat plant is the first leaf below the spike. The inoculated flag leaves are called thereafter spore sources. Exogenous infection (from non-artificial sources) was at most avoided by applying a fungicide three weeks before the artificial inoculation. About two weeks after the inoculation, daughter lesions appeared on leaves surrounding the sources. The daugther lesions were counted for all the flag leaves in the neighborhood of 5 of the 14 spore sources (one lesion count per leaf). The 5 retained spore sources were the ones around which the plant canopy was healthy before the experiment, and homogeneous in plant density and nutritional state. Daugther lesions were not counted around the 9 other spore sources. The neighborhood of a spore source, thereafter called sampling zone, is defined by a rectangle with dimensions 80 cm (-40 cm to +40 cm) and 18.4*3 cm (rows -1, 0 and +1). It is drawn in the bottom panel of figure 3.1. Leaf locations were not exactly measured : leaves were located in small rectangular sets, called quadrats. The quadrats partition the sampling zone in 36 parts which are drawn in the bottom panel of figure 3.1). The farthest quadrats from the spore source are twice larger than the closest quadrats because the lesion count was expected to be almost constant between 20 and 30 cm and between 30 and 40 cm from the source.

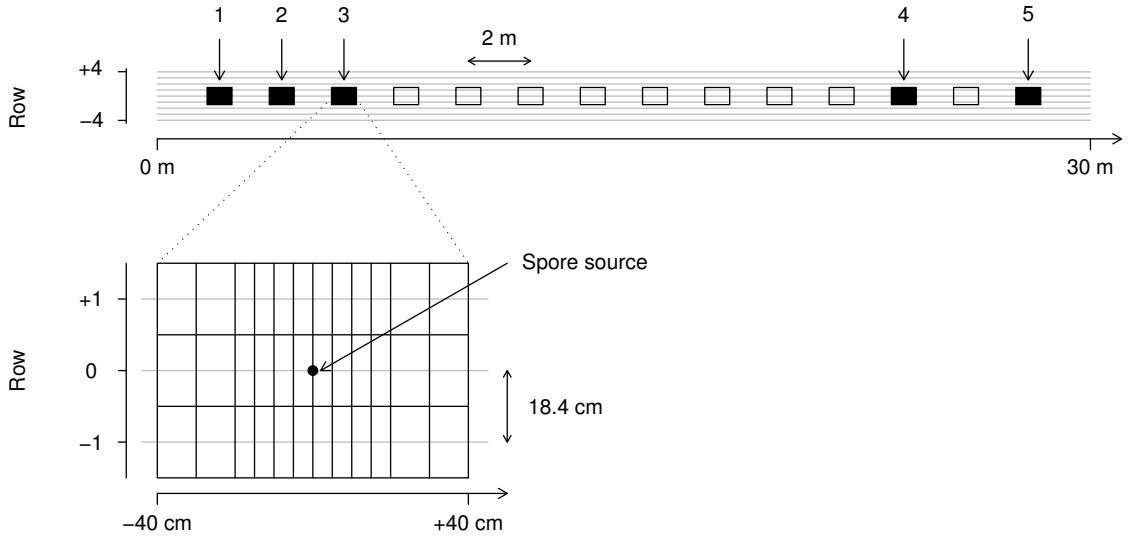


Fig. 3.1. Field experiment. Top : experimental field ; black-filled rectangles are the 5 subexperiments taken into consideration. Bottom : sampling map for each subexperiment ; lesions are counted for all the leaves located in drawn quadrats.

Thus, the field experiment consists in 5 subexperiments denoted by index i in $\{1, \ldots, I = 5\}$ (see top panel of figure 3.1). Observed variables are quadrats $\{A_{ij} \subset \mathbb{R}^2 : i = 1, \ldots, I, j = 1, \ldots, J_i\}$ and lesion counts $\{N_{ijk} : i = 1, \ldots, I, j = 1, \ldots, J_i, k = 1, \ldots, K_{ij}\}$. A_{ij} denotes the surface of quadrat j of subexperiment i. For all i, the number of quadrats is $J_i = 36$. N_{ijk} denotes the count of daughter lesions on leaf k of quadrat j of subexperiment i.

3.2.2 Heterogeneity of individuals

Leaf-scale variations of leaf conditions cause heterogeneity of leaves in their propensities to be infected. Indeed, as a biotrophic fungus, brown rust infects more easily vigorous leaves than non-vigorous leaves (Rapilly, 1991). Consequently, differences in nutritional and hydric status among leaves, which induce difference in vigor among leaves, result in differences in propensities of leaves to be infected. Note that differences in nutritional and hydric status happens even if the plant canopy seems visually homegeneous in plant state. Moreover, the success of propagules to infect leaves depends on microclimatic conditions such as temperature and wetness at the leaf scale (Campbell and Madden, 1990; Rapilly, 1991). Consequently, difference in 3D-geometry among leaves (size, shape, position), which induces differences in microclimatic conditions among leaves, results on differences in propensities of leaves to be infected.

Thus, the dynamic of epidemics of brown rust within a field crop of wheat can be better understood by assessing the role, in disease spread, of propensities of leaves to be infected. Note that neither the propensities nor their factors are observed. Moreover, contrary to propagule dispersal, propensity of a leaf to be infected, as used in this paper, is not a common notion in botanical epidemiology. It is different from susceptibility which is a genetically-determined propensity to be infected (Rapilly, 1991). Note that in the field experiment, wheat is genetically homogeneous. It can be more likely connected with receptivity which is an environmentally-determined propensity to be infected (Rapilly, 1991). However, in Rapilly (1991), receptivity varies with time whereas our propensity varies with individuals. To avoid any confusion, we use the term frailty. In statistics, the frailty is a random effect and, precisely, the unobserved propensity of a leaf to be infected is viewed as a random effect in this paper. Also, the propensity of a leaf to be infected is, differently speaking, the frailty of a leaf facing a disease attack.

3.3 The frailty model for disease spread

Let $N_1, \ldots, N_K$ be random counts of lesions on K leaves under the influence of a single spore source located at 0 in $\mathbb{R}^2$. Let $X_1, \ldots, X_K$ denote leaf locations in $\mathbb{R}^2$. We model the distribution of lesion counts as follows.

3.3.1 Infectious potential and dispersal function

Assuming that transports of spores are independent (McCartney, 1994) and identically distributed, the infectious potential S_{ab} is defined as the product between a quantity $a > 0$ of spores produced by the source, called source strength, and a dispersal function f_b with dispersal parameter b

$$S_{ab}(x) = a f_b(x), \quad \forall x \in \mathbb{R}^2.$$

The quantity $S_{ab}(x)$ is a measure of the risk of infection at x in $\mathbb{R}^2$, and the function S_{ab} represents an intensity of spores produced by the source.

Let D be the random location of deposition in $\mathbb{R}^2$ of a spore emitted at 0. f_b is its density function. We assume that

$$f_b(x) = \frac{1}{2\pi b^2} \exp\left(-\frac{||x||}{b}\right), \quad \forall x \in \mathbb{R}^2,$$

where $b > 0$ and $||.||$ is the $\mathbb{R}^2$-Euclidean distance. f_b is chosen isotropic because data do not provide evidence for anisotropic spread. Its exponential form is obtained under the assumptions that spores move in radial half lines and that the probability of deposition in the infinitesimal interval $[r, r + dr]$ is constant whatever the already traveled distance r (Tufto et al., 1997). Given D belongs to any radial half line, the conditional density function of D is exponential ; this 1-dimensional dispersal function is widely used in botanical epidemiology as a disease gradient (Aylor, 1990; McCartney and Fitt, 1998).

3.3.2 Leaf frailties

The propensity of leaf k ($k = 1, \ldots, K$) to be infected, which determines the proportion of spores succeeding to infect leaf k, is influenced by unobserved leaf features. Therefore, it is modeled as a random variable Z_k, called leaf frailty, which varies between 0 and 1. As leaf frailties are assumed to depend on biological characteristics at the leaf scale, they are modeled as independent and identically distributed random variables. As no biological assumption was available to choose the density of the leaf frailties, we uses the following polynomial form

$$f_{cd}(z) = \{cz^2 + dz + e(c,d)\}^2, \quad 0 \leq z \leq 1,$$

where $e(c,d) = -c/3 - d/2 + \sqrt{\Delta(c,d)}/2$ and $\Delta(c,d) = -16c^2/45 - d^2/3 - 2cd/3 + 4$ to ensure $\int_{[0,1]} f_{cd} = 1$. Frailty parameters (c,d) are constrained by $\Delta(c,d) \geq 0$ (elliptical area). The polynomial form for f_{cd} allows us to get a flexible density with only two parameters, and to speed up the maximization of the log-likelihood as an integration over [0,1] is replaced by a sum of 5 terms.

3.3.3 Conditional distribution of lesion counts

Lesion counts $N_1, \ldots, N_K$ conditional on frailties $Z_1, \ldots, Z_K$ and leaf locations $X_1, \ldots, X_K$ are assumed to be independent and to obey Poisson distributions with intensities

$$Z_k S_{ab}(X_k) = Z_k \frac{a}{2\pi b^2} \exp\left(-\frac{||X_k||}{b}\right), \quad k = 1, \ldots, K.$$

3.4 Estimation method

We are interested in estimating, for subexperiment i in $\{1,\dots,I\}$, the source strength, the dispersal parameter and the frailty parameters, under the constraint that leaf locations are restricted to quadrats. In the next sections, we derive maximum likelihood estimators for these parameters and assess their uncertainty using a normal approximation and parametric bootstrap.

3.4.1 Likelihood function

Consider subexperiment i (i fixed in $\{1,\dots,I\}$). Observed variables are $\{N_{ijk} : j = 1,\dots,J_i, k = 1,\dots,K_{ij}\}$ where N_{ijk} is the count of lesions on leaf k located in quadrat j with surface $A_{ij} \subset \mathbb{R}^2$. Assume unobserved leaf locations X_{ijk} are independent and uniformly distributed in quadrats A_{ij}. Then, as under $\theta = (a,b,c,d)^T$ variables $N_{ijk}|X_{ijk}, Z_{ijk}$ are Poisson distributed with intensities $Z_{ijk}S_{ab}(X_{ijk})$, the probability that N_{ijk} equals n in $\mathbb{N}$ given A_{ij} is

$$p_\theta^{ij}(n) = \int_0^1 \left[\frac{1}{|A_{ij}|} \int_{A_{ij}} \exp\{-zS_{ab}(x)\} \frac{\{zS_{ab}(x)\}^n}{n!} dx \right] f_{cd}(z)dz.$$

Expanding f_{cd} in monomials : $f_{cd}(z) = \sum_{m=0}^4 \gamma_{cd}(m) z^m$, $z \in [0,1]$, the log-likelihood $\sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \log\{p_\theta^{ij}(N_{ijk})\}$ of θ for subexperiment i can be written, up to a constant,

$$l_{K_i}^i(\theta) = \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \log \left\{ \sum_{m=0}^4 \gamma_{cd}(m) \frac{(N_{ijk}+m)!}{N_{ijk}!} \int_{A_{ij}} \frac{1 - F_{S_{ab}(x)}(N_{ijk}+m)}{S_{ab}(x)^{m+1}} dx \right\}, \qquad (3.1)$$

where $F_\lambda(u) = (1 - e^{-\lambda u})$ if $u > 0$, 0 otherwise, and $K_i = \sum_{j=1}^{J_i} K_{ij}$ is the total number of leaves for subexperiment i. Let $\hat{\theta}_i$ be the maximum likelihood estimator (MLE) of θ for subexperiment i obtained by maximizing $l_{K_i}^i(\cdot)$.

3.4.2 Estimator accuracy

To know the uncertainty of the estimator of θ, i.e. to get confidence intervals for the parameters, the behavior of $\hat{\theta}_i$ must be assessed. We assess the behavior of $\hat{\theta}_i$ by providing its asymptotic distribution when K_i tends to infinity. Since we are interested in estimating disease spread within a well-identified bounded domain, we use fixed-domain asymptotic (Stein, 1999), that is the number of leaves K_i is increased in a fixed spatial domain. The determination of the asymptotic distribution of $\hat{\theta}_i$ is not standard because counts of lesions are independent but non identically distributed (i.n.i.d.). However, by using theorems for i.n.i.d. variables (Hoadley, 1971; Philippou and Roussas, 1973), we show that $\hat{\theta}_i$ is consistent and, under θ, the limiting distribution of $\sqrt{K_i}(\hat{\theta}_i - \theta)$ is a centered normal distribution. A sketch of proof is provided in Appendix A.1.

In Table 3.1, we provide, for various leaf densities, coverage probabilities of the 95%-confidence ellipsoid for θ obtained from the normal approximation of $\hat{\theta}_i$, thereafter called

standard ellipsoid. For each leaf density, 100 subexperiments were drawn under the frailty model with parameters estimated for subexperiment 5 (see Table 3.2). For each subexperiment, the standard ellipsoid was computed. The coverage probability is the proportion of standard ellipsoids including the true parameters. In the application, the mean number of leaves per subexperiment is 275. For such a leaf number, the standard ellipsoid is actually a 70%-confidence ellipsoid.

Tab. 3.1. *Coverage probabilities of the standard ellipsoid. Leaf density : number of leaves sampled in each small quadrat (the number of leaves sampled in each large quadrat is two times the leaf density, see Figure 3.1). Leaf number : total number of leaves per subexperiment.*

Leaf density	3	5	10	20	30
Leaf number	144	240	480	960	1440
Coverage (%)	56	68	72	85	93

As the confidence level 95% is not achieved by using the normal approximation, we use parametric bootstrap (Efron and Tibshirani, 1993) to compute 95%-confidence intervals for the parameters. Confidence intervals are computed as follows. Once the parameters of the frailty model are estimated, we simulate 200 independent bootstrap datasets under the estimated parameters by preserving the numbers of leaves per quadrat. Then, for each bootstrap dataset, we compute the bootstrap MLE of the parameters. For any parameter, the endpoints of the 95%-confidence interval, called percentile interval, are the 5th and the 195th ordered values of the 200 bootstrap estimates.

3.5 Results

3.5.1 Dataset and overdispersion

Figure 3.1 represents the field experiment together with the locations of the five subexperiments. The number of leaves per subexperiment ranks from 256 to 294 (mean=275). For all the subexperiments, the percentage of infected leaves is high, varying between 93.7% and 98.0%. The count of lesions per leaf is very variable, ranking from 0 to 816. Left panel of Figure 3.2 summarizes the distributions of the lesion count (y-axis) for the five subexperiments (x-axis). The y-axis is logarithmic. All the distributions are very skewed and show similar shapes even if some statistics such as the median vary.

Right panel of Figure 3.2 shows overdispersion of data. It plots sample variance per quadrat versus sample mean per quadrat (stars) together with estimated variance per quadrat versus estimated mean per quadrat (dots), estimated values being obtained under a model without frailty. The model without frailty ($N_{ijk}|X_{ijk} \sim \text{Poisson}\{S_{ab}(X_{ijk})\}$) is fitted with a least squares criterion. A 95%-confidence zone under the estimated model without frailty is drawn (grey zone). It is computed by performing 499 Monte-Carlo simulations under the estimated model. It is the smallest zone which contains 95% of the points corresponding to

simulated variance per quadrat versus simulated mean per quadrat. It corresponds to the region where neither overdispersion nor underdispersion are detected. Unlike the line stating variance equals mean (which is also drawn), it takes into account variations of lesion counts due to variations of the infectious potential within each quadrat. Overdispersion appears clearly since the cloud of sample points (stars) is over the simulated confidence zone (grey zone).

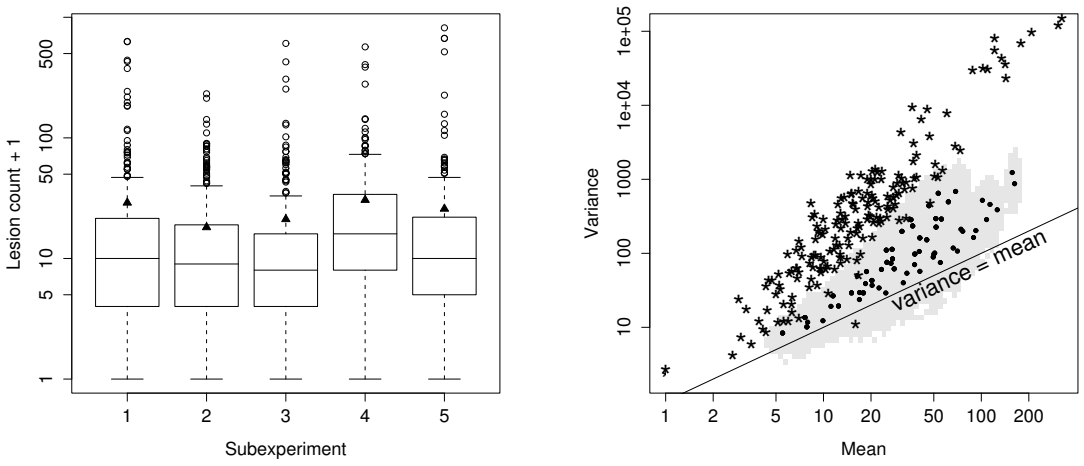


Fig. 3.2. Data dispersion. Left : distribution per subexperiment of lesion counts (plotted in log-scale) ; triangles are the sample means. Right : variance per quadrat versus mean per quadrat both plotted in log-scale ; stars for sample statistics, dots for estimated statistics under the model without frailty, grey zone for a 95%-confidence zone obtained under the model without frailty.

3.5.2 Parameter estimation

Assuming the subexperiments are isolated, i.e. each spore source only contributes to the infection of its surrounding leaves, the log-likelihood $l(\cdot)$ for the five subexperiments is the sum

$$l(\theta_i : i = 1, \ldots, 5) = \sum_{i=1}^{5} l^i_{K_i}(\theta_i), \tag{3.2}$$

where $\theta_i = (a_i, b_i, c_i, d_i)^T$ is the vector of parameters for subexperiment i, and $l^i_{K_i}(\cdot)$ is the log-likelihood for subexperiment i (Equation (3.1)). If the subexperiments do not share parameters then log-likelihoods $l^i_{K_i}(\cdot)$, $i = 1, \ldots, 5$, can be separately maximized to estimate the parameters.

At first, we test if the subexperiments share some parameters by using three maximum likelihood ratio tests whose null hypotheses are equality of the source strengths ($a_1 = \cdots = a_5$), equality of the dispersal parameters ($b_1 = \cdots = b_5$), and equality of the frailty parameters ($c_1 = \cdots = c_5$ and $d_1 = \cdots = d_5$). To achieve a global significance level less

than or equal to 5%, the significance level for each of the three tests is 1.667% (Bonferroni procedure, Miller Jr. (1981)). The global significance level is not necessarily 5% because of dependence between tests. The source strengths and the dispersal parameters cannot be accepted as equal for the five subexperiments ($p = 0.0004$ and $p = 0.0042$, respectively), whereas equality of the frailty parameters is not rejected ($p = 0.0307$, see also Figure 3.3). The source strengths are variable because disease inoculations were carried out by applying a mixture talc/spores on concerned leaves, and this method does not allow to control the resulting count of lesions. The significant difference between the dispersal parameters may be due to varying local conditions (local turbulence, spore source orientation, unexpected spore sources). Remark that no clear relationship appears between the source strengths and the dispersal parameters. On the other side, the subexperiments having been carried out simultaneously and in homogeneous zones (see Section 3.2.1), same frailty distributions were expected as long as they were related to crop features. In the following, we consider that the subexperiments share the frailty parameters.

Parameter estimates are provided in Table 3.2 together with their standard and percentile intervals (see Section 3.4.2). Likelihood ratio tests for a and b are corroborated by the study of the overlap of the percentile intervals and by tests based on the bootstrap estimates (not shown). The dispersal parameters b for subexperiments 1 and 4 are high compared with the others. In fact, unexpected spore sources are suspected in quadrat [row=0, distance=35cm] for subexperiment 1 and in quadrat [row=0, distance=12.5cm] for subexperiment 4. Unexpected spore sources are unexpected lesions, appeared despite of the preventive treatment (Section 3.2.1), which induce daughter lesions simultaneously with artificial spore sources. Daughter lesions due to artificial and unexpected sources are indistinguishable and, consequently, are counted together. If there exists an unexpected source in the study domain, but not at the artificial source location, then a higher estimate for the dispersal parameter is expected. Rejection of equality of the dispersal parameters may be partly due to such events.

Figure 3.3 presents estimated density functions of the leaf frailty when frailty parameters are shared by the 5 subexperiments (solid line) or when they are not (dashed lines). The slight differences between the dashed lines corroborate that frailty parameters are not statistically different. Frailty is not uniformly distributed between 0 and 1, but shows a high peak around 0 corresponding to leaves with low propensities to be infected. The solid line shows rebounds at $z = 0.6$ and $z = 1$ surely because the density function f_{cd}, as a constrained polynomial of degree 4, is not enough flexible to be, for example, constant on [0.4,1]. However, the mass of segment [0.4,1] being less than 0.05, the eventual mis-estimation of f_{cd} on [0.4,1] is of minor importance.

Tab. 3.2. Parameter estimates (1st rows) together with their standard intervals (2nd rows) and percentile intervals (3rd rows). Definitions of standard and percentile intervals are provided in Section 3.4.2. Estimates for c and d are shared by the 5 subexperiments.

	Subexperiment				
	1	2	3	4	5
$a.10^{-6}$	1.77	0.89	0.89	2.01	1.31
	(1.63,1.91)	(0.74,1.03)	(0.75,1.04)	(1.87,2.15)	(1.16,1.45)
	(1.37,2.25)	(0.63,1.15)	(0.65,1.14)	(1.52,2.52)	(0.96,1.69)
b	19.3	17.0	13.9	19.0	14.5
	(17.4,21.2)	(15.1,18.9)	(12.0,15.8)	(17.1,20.9)	(12.6,16.3)
	(16.3,23.3)	(14.7,22.0)	(12.0,16.7)	(16.0,22.9)	(12.8,17.3)
c			8.00		
			(7.54,8.46)		
			(7.16,8.35)		
d			-10.29		
			(-10.71,-9.87)		
			(-10.59,-9.71)		

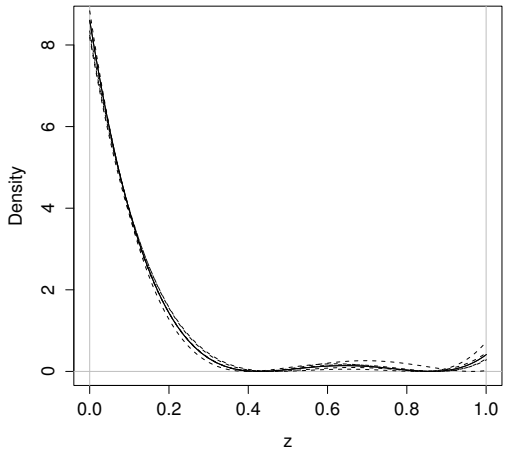


Fig. 3.3. Estimated density function of the leaf frailty when frailty parameters are shared by the five subexperiments (solid line) and when they are not shared (dashed lines, one for each subexperiment).

3.6 Model check

3.6.1 Is overdispersion handled ?

Model check is first performed by comparing the variability of sample data to the variability achieved under the estimated frailty model. Figure 3.4 plots sample variance per quadrat versus sample mean per quadrat (stars) together with estimated variance per quadrat versus estimated mean per quadrat under the frailty model (dots). A 95%-confidence zone is computed by performing 499 Monte-Carlo simulations under the estimated model, and drawn (grey zone). The straight line stating variance equals mean gives a benchmark.

Whereas the cloud of sample points and the simulated confidence zone do not overlap in Figure 3.2, they strongly do in Figure 3.4. Thus, overdispersion of individual data is handled by the frailty model.

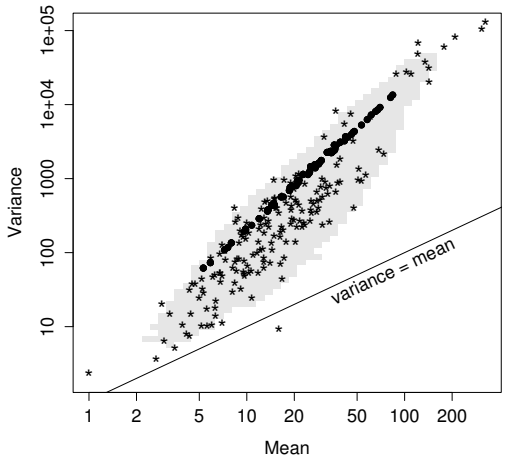


Fig. 3.4. Variance per quadrat against mean per quadrat both plotted in log-scale : stars for sample statistics, dots for estimated statistics under the frailty model, grey zone for a 95%-confidence zone obtained under the frailty model.

3.6.2 Is the dispersal function exponential ?

Second, model check is performed by comparing the sample distribution to the estimated distribution of the lesion count per leaf for various distances from the spore source (see Appendix A.2). For each subexperiment, both distributions coincide almost at every distance. They do not coincide only near the source, i.e. in the two quadrats in contact with the source, for subexperiments 1, 3 and 5. Indeed, in these quadrats, the sample mean and the sample standard deviation are underestimated by the estimated model.

To solve this problem, we fitted the frailty model by using a broader class of dispersal functions : the class of exponential power functions (Tufto et al., 1997). The fit was significantly better for subexperiments 1, 3 and 5. Nevertheless, the variability of the estimated source strengths was biologically unrealistic and leads us to not use exponential power functions as dispersal functions. To conclude, the exponential dispersal function seems appropriate for subexperiments 2 and 4, and further investigations are needed for the others to understand the underestimation of the disease level near the source. Section 3.7.1 discusses the choice of the dispersal function : biological elements suggest the use of a mixture of dispersal functions.

3.6.3 Are the frailties i.i.d. ?

Third, model check is performed by looking at the spatial structure of the standardized residuals per quadrat (thereafter called residuals) defined for all i in $\{1, ..., I\}$ and for all j in $\{1, ..., J_i\}$ by

$$\epsilon_{ij} = \left\{ \frac{K_{ij}}{\mathbb{V}_\theta(N_{ijk})} \right\}^{1/2} \left\{ \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} N_{ijk} - \mathbb{E}_\theta(N_{ijk}) \right\}.$$

Expressions of the expected value of the lesion count $\mathbb{E}_\theta(N_{ijk})$ and of its variance $\mathbb{V}_\theta(N_{ijk})$ are provided in Appendix A.3. Under the model, the residuals have expected values 0, variances 1 and covariances 0.

To look at the spatial structure of the estimated residuals, we use permutation tests which do not require any distributional assumption. The tests are only applied to subexperiments 2 and 4 as the disease level near the source is underestimated for the other subexperiments (see Section 3.6.2).

We performed Mantel's tests (Manly (1997)). It consists in seeing whether the correlation between absolute differences in estimated residuals and reciprocal distances between quadrats is negative. A significantly negative correlation is an evidence that close quadrats tend to have similar estimated residuals. The observed correlations and the p-values are corr $= -0.100$ and $p = 0.020$ for subexperiment 2, and corr $= -0.068$ and $p = 0.062$ for subexperiment 4 (p-values were computed by permuting 9999 times the estimated residuals). The low p-values and the observed correlations close to 0 indicate that there is a fairly small effect of spatial correlation.

We also performed an independence test described in Peyrard et al. (2005). It consists in computing the sample semivariogram of the estimated residuals and seeing whether it goes out of its 95%-confidence envelopes obtained by permuting the estimated residuals. Figure 3.5 shows the sample semivariogram of the estimated residuals of subexperiments 2 and 4, together with its 95%-confidence envelopes based on 999 permutations. The first dot of the semivariogram slightly goes out of the envelopes. It suggests that the estimated residuals are slightly spatially dependent.

The spatial dependence between the estimated residuals may be caused by spatially dependent frailties. Such a hypothesis can be biologically justified : factors as fertilization and water levels influencing leaf frailties can be spatially structured, and so would be the frailties. In Section 3.7.2, we discuss more realistic models than our model with i.i.d. frailties : these models include a dependence structure. In Section 3.7.3, we suggest an approach to learn about the dependence structure by estimating our frailty model.

3.7 Discussion

To assess short-distance spread from a single source of the brown rust of wheat, we have measured counts of lesions on individual hosts. The assessment was expected to be more

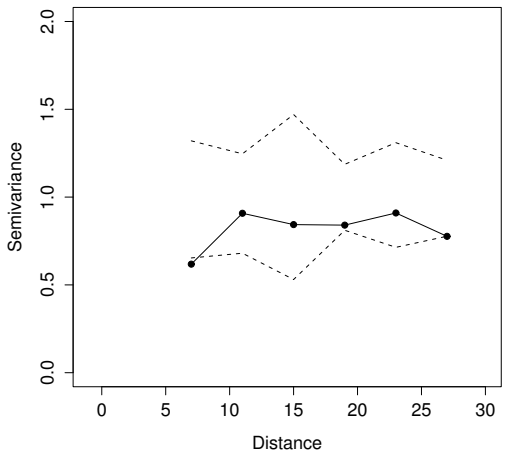


Fig. 3.5. Sample semivariogram for standardized residuals (dots linked by the solid line) together with 95%-confidence envelopes (dashed lines) computed by performing 999 permutations.

accurate by counting lesions on individual hosts (individual data) rather than counting lesions on sets of hosts (aggregated data) as it is commonly done in botanical epidemiology. To analyze the individual count data, we have proposed, estimated and checked a frailty model. In the model, a dispersal function characterizes propagule dispersal, and frailties characterize propensities of hosts to be infected by propagules.

Analyzing individual count data with the frailty model, rather than analyzing aggregated data with a disease gradient, effectively allows to get a more accurate assessment of the spread (see appendix A.4). In addition, estimating and checking the model leads to investigate biological and statistical questions about the dispersal function of propagules and the dependence structure of host frailties. These questions are discussed in the following sections.

3.7.1 The dispersal function of propagules

The dispersal function is assumed to be exponential. In the model check (Section 3.6.2), we have seen that this class of functions underestimates the level of disease near the propagule source (in the first centimeters) for three of the five subexperiments. Klein et al. (2003) and Tufto et al. (1997) propose other classes of dispersal functions which give insight into how propagules are transported by wind. However, these functions would not help us if, as we suspect, the level of disease can be underestimated near the source because of other possible dispersal mechanisms occurring at very short range (few centimeters). The dispersal mechanisms we have in mind are rain splashing and leaf rubbing. Rain splashing causes dispersal of propagules in splash droplets, and is a known dispersal mechanism (Geagea et al., 2000; McCartney and Fitt, 1998). Leaf rubbing may cause transport of spores from a diseased host to another because of physical contacts between the leaves ; it has so far never been proved for airborne diseases of plants and would justify an experimental approach. To

take into consideration different dispersal mechanisms in the modeling, a mixture of dispersal functions would be used. But more information about the potential of each dispersal mechanism is needed to assess which mixtures can be realistic.

3.7.2 The dependence structure of host frailties

With our model, not only do we assess disease spread, but, by estimating the distribution of the host frailty, we also describe the variability of propensities of hosts to be infected. In this study, the host frailties are assumed to be i.i.d. However, the analysis of model residuals in Section 3.6.3 has suggested that host frailties are slightly spatially structured. A biological reason is that factors influencing host frailties, such as fertilization and water levels, can be spatially structured. Therefore, proposing and estimating a model including a dependence structure for the frailties would be an advance in describing the variability, in space, of propensities of hosts to be infected.

Such a model could be still developed in the framework of frailty models. For instance, Henderson et al. (2002) propose Bayesian models, with gamma-distributed and spatially-structured frailties, to describe spatial variation in leukemia data. Or it could be developed in the framework of Cox processes (Diggle, 1983) which are particularly adapted when environmental factors are spatially and stochastically heterogeneous, as can be factors influencing frailties. In this framework, frailties (unobserved random variables) are replaced by an unobserved random field. The product between this random field and the infectious potential is a stochastic intensity function whose support is host surfaces. Then, lesions follow a Poisson process conditional on a realization of the stochastic intensity function.

3.7.3 An approach to learn about the dependence structure

Whatever be the framework for a model including a dependence structure, a spatial distribution needs to be specified. However, no quantitative knowledge in biology is available to specify this spatial distribution. Such quantitative knowledge can be acquired through the estimation of our i.i.d.-frailty model. Indeed, first, information about the marginal distribution of the frailty is provided, second, information about the spatial dependence of the frailties could be learned through residuals estimated under the i.i.d.-frailty model.

Let us explain how we expect to learn about the spatial dependence of the frailties from residuals. Let H_0 denote the null hypothesis "frailties are i.i.d.", and H_1 denote the alternative hypothesis "frailties Z_{ijk} depend on quadrat parameters which are spatially structured", e.g. the frailty parameter c is replaced by quadrat parameters c_{ij}, $i = 1, \ldots, I$ and $j = 1, \ldots, J_i$. Under H_1, residuals obtained by estimating the model assuming H_0 true are functions of the quadrat parameters. Therefore, by exhibiting a relationship linking the estimated residuals and the quadrat parameters, we can expect to get information about quadrat parameters through residuals. The aim of such a residual analysis is between recognizing the form of departure from a model (McCullagh and Nelder, 1989) and using residuals to modify a model (Box and Jenkins, 1976, pp 298-299).

In this approach, the i.i.d.-frailty model, by acting as a filter which catches the tendencies of propagule dispersal and host heterogeneity, could be a tool to investigate poorly-known underlying processes involved in disease spread.

Nous remercions Corinne Robert et Lise Frezal pour leur investissement dans les expériences biologiques, et Rachid Senoussi pour ses commentaires sur cette étude.

4

Propagation spatiale à longue distance de la rouille jaune

Deux expériences de propagation spatiale à longue distance (de 10 à 200 mètres) à partir d'une source (assimilée à un point au vu des distances mises en jeux) de spores de rouille jaune ont été menées pour évaluer les risques d'infection entre parcelles agricoles distantes et pour mieux comprendre la formation de foyers secondaires de maladie éloignés du foyer primaire (la source). Les observations sont des nombres de feuilles malades dans des placettes (*trap plots*) d'un mètre carré réparties autour de la source.

Les données montrent une décroissance du nombre de feuilles malades par placette avec la distance à la source, décroissance qui dépend de la direction considérée. Cette dépendance de la décroissance avec la direction est attribuée à une anisotropie de la dispersion des spores. Pour caractériser l'anisotropie, nous proposons deux fonctions angulaires : l'une reliée à la quantité de spores déposées dans chaque direction ; l'autre donnant, pour chaque direction, la distance moyenne entre la source et les positions de dépôt des spores. Une approche non-paramétrique permet d'estimer ces fonctions et de leur attribuer des formes paramétriques. Les formes paramétriques sont ensuite intégrées à un modèle de propagation spatiale décrivant, en tout point de l'espace, le nombre de feuilles malades par mètre carré. Ce modèle est basé sur un modèle sous-jacent décrivant le nombre de lésions par feuille. Dans ce modèle sous-jacent, un potentiel infectieux intègre la force de la source et la dispersion anisotrope des spores ; le nombre de lésions sur une feuille suit une loi de Poisson dont la moyenne est la valeur locale du potentiel infectieux.

Les paramètres du modèles ont été estimés par maximum de vraisemblance pour les deux jeux de données issues des deux expériences. Pour chaque expérience, des tests ont rejeté la constance des deux fonctions angulaires, montrant que l'anisotropie est significative ; d'autres tests ont rejeté la proportionnalité des deux fonctions angulaires, montrant que la direction dans laquelle les spores vont le plus loin en moyenne n'est pas forcément la direction dans laquelle le plus grand nombre de spores sont déposées.

Un test montre que le modèle proposé s'ajuste relativement bien aux données. Le modèle parvient donc à capturer une grande part de la variabilité des données. Toutefois, une étude des écarts entre modèle et données suggère que le potentiel infectieux est perturbé par des effets structurés spatialement. Cette étude sera menée dans les sections 5.3.1 et 8.7.

Anisotropy, in direction and in distance, of the dispersal of yellow rust of wheat : experiments in large field plots and estimation

By Samuel Soubeyrand, Ivan Sache, Jérôme Enjalbert, Alicia Sanchez and Joël Chadœuf

Long distance dispersal of spores generally presents anisotropy. This anisotropy can appear in the mean number of spores deposited along a given direction, and in the mean distance that a spore travels in a given direction. Specific experiments, based on trap plots in large field plots and the use of a rare inoculum, together with a statistical methodology are proposed to study this effect. Anisotropies are characterized by two functions : a directional density function and a mean distance function. First, a nonparametric approach is developed to estimate these functions and to help in choosing a parametric model. Second, the parametric model is estimated. Whatever the experiment, the two anisotropies present different shapes, i.e. the number of spores dispersed in a given direction is not proportional to the mean distance traveled by these spores.

4.1 Introduction

Yellow rust of wheat, caused by *Puccinia striiformis*, is a major cause of yield loss in wheat (Roelfs et al., 1992; Oerke et al., 1994). The pathogen is spread by airborne spores dispersed locally and over long distances. Yellow rust is a typical focal disease, probably because of the steepness of the spore dispersal gradient (Rapilly, 1979). Most spores are liberated from sporulating lesions in clusters of 2-20 spores and are deposited closer to the source than would be spores dispersed as singletons (Rapilly et al., 1970; Rapilly, 1977, 1979). However, a few liberated spores are carried away by wind gusts on larger distances and form daughter (secondary) foci far away from the primary (mother) focus (Zadoks and Van den Bosch, 1994). Spore dispersal gradients from a point source have been characterized experimentally and theoretically for several airborne pathogens (Fitt et al., 1987) ; in yellow rust, gradients have been assessed locally (up to 15-30 m; Emge and Shrum, 1976; Rapilly et al., 1970) and, recently, on a larger scale (up to 80 m; Sackett and Mundt, 2005). Epidemiology and population genetics studies have revealed the continental dispersal of yellow rust, as reviewed Brown and Hovmøller (2002).

On small scales (a few meters), spore dispersal is isotropic, because local turbulence is more important than prevailing winds (Zadoks et al., 1969), and primary disease foci are roughly circular (Sache and Zadoks, 1996; Van den Bosch et al., 1988, 1990). On larger scales, however, spore dispersal is anisotropic and the shape of the disease foci is no

longer circular (Emge and Shrum, 1976). This shape is the result of the distributions in distance and direction of the dispersal units (individual spores or spore clusters[1] Rapilly, 1977). There is no hint that these two distributions should be identical. Visualization of the spread of individual spores or spore clusters from a disease focus in field conditions is technically impossible. The classical approach of spore dispersal is based on sampling of disease along lines radiating from the source (Emge and Shrum, 1976; Sackett and Mundt, 2005). Anisotropy in dispersal direction cannot be correctly assessed with such a design since the number of sampling directions is finite and the non-sampled area increases with increasing distance from the source (Gregory, 1945; Cammack, 1958). A systematic, short-scale, sampling of a big plot is no longer possible when the scale of the experiment increases. In contrast, anisotropy of spore dispersal both in distance and direction can be experimentally assessed with a large array of small trap plots arranged around a source plot artificially inoculated (Kingsolver et al., 1984; Sache and Zadoks, 1996; Underwood et al., 1959). The choice of the inoculum, the respective size of the source and trap plots, as well as the time schedule of the experiment, must be decided according to the three following epidemiological features of a polycyclic, airborne disease like yellow rust : (i) dispersal on larger scales involves a very limited spore load and disease caused by these spores is often below practical detection thresholds ; (ii) it is impossible to avoid contamination by external sources of inoculum (Danial et al., 1993) especially when the experimental scale increases ; and (iii) the succession and juxtaposition of spore production cycles makes it difficult to trace back specific contamination events caused by long-distance dispersal. Therefore, the experiment should be done with a significant trap plot area, an easily identifiable inoculum and during a single generation of the pathogen.

The experiment proposed above allows to visually assess anisotropy of spore dispersal. For quantitatively assessing it, different kinds of mathematical models can be used. First, classical gradient functions (e.g. exponential and power law) can be fitted to data collected in several directions (Gregory, 1968; Emge and Shrum, 1976) ; as said above, anistropy is not correctly addressed by these models. Second, mechanistic models based on the physical description of spore dispersal (McCartney and Fitt, 1985; Aylor, 1990) can be used to describe the transport of spores, but they don't differentiate anisotropy in direction and distance : in average, the spores travel further in their preferred dispersal direction. These models also require a very detailed description of the physical environment (Aylor, 1998) not easily obtainable in field conditions. Third, the so-called quasi-mechanistic models (Tufto et al., 1997; Klein et al., 2003; Stockmarr, 2002) require a much simpler characterization of the environment, but they do not address either the differential anisotropy in direction and distance.

In this paper, we propose a combined experimental and statistical methodology of analysis of spore dispersal anisotropy in direction and distance, as it occurs in the dispersal of wheat yellow rust. The experimental design is an array of trap plots in which spore dispersal

[1] Amas de spores.

from a source plot is assessed. In the modeling approach used to describe spore dispersal, anisotropy is defined by two functions characterizing dispersal in distance and direction, respectively. The first function describes the direction taken by the spores during their travel whereas the second one provides the expected distance between the spore source and its deposition location, given the direction taken by the spore. The two anisotropy functions are not constrained by any prerequisite model ; however, for assessing them, we propose a modeling and estimating framework. In a first step, a non-parametric analysis allowed us to figure out the shape of the two anisotropy functions ; in a second step, a parametric analysis was performed using the information gained from the first, non-parametric analysis.

4.2 Material and method

4.2.1 Field experiments

Field layout

The field experiments were performed during the 2001-2002 cropping season in two locations of north-western France, Le Neubourg (Upper-Normandy ; 49°15'N, 0°87'W ; elevation, 130 m a.s.l.), which has no recent history of yellow rust epidemics, and Thiverval-Grignon (Ile-de-France ; 48°50'N, 1°57'E ; elevation, 130 m a.s.l.), which has a long story of natural and experimental yellow rust epidemics. The distance between the two locations is c. 100 km.

Wheat cv. Soissons, resistant to yellow rust, was sown in a 400 x 75 m plot in Le Neubourg on 16 October 2001 and in a 250 x 300 m plot in Thiverval-Grignon on 23 October 2001. Crop management followed the local standard practices. Two weeks after sowing, 1 x 1 m plots, hereafter referred to as the "trap plots", were delimited in the main plots on a 24 x 20 m grid in Le Neubourg and on a 25 x 24 m grid in Thiverval-Grignon (see figures 1 and 2 respectively). There were 72 trap plots in Le Neubourg and 187 trap plots in Thiverval-Grignon. In Le Neubourg, the main plot was completely flat and the trap plots were arranged on a square grid. In Thiverval-Grignon, the main plot was located on a hillside and the two axes of the grid made an angle of 40 degrees. In each trap plot, the emerging seedlings of wheat cv. Soissons were removed with a hoe and three rows of wheat cv. Victo, highly susceptible to yellow rust, were sown manually. Emergence of wheat seedlings in the trap plots was homogeneous in Le Neubourg but heterogeneous in Thiverval-Grignon. In Le Neubourg, the number of leaves per trap plot was assessed at 600. In Thiverval-Grignon, seedling density in the trap plots was visually assessed on 30 January 2002 on a 0 (no seedlings) to 7 (fully established crop) scale ; the different levels correspond to different numbers of leaves ranking from 0 to 900.

Inoculation procedure

A clonal isolate of P. striiformis, pathotype 106E139 (nomenclature according to Johnson et al., 1972), a pathotype no longer present in the natural airborne yellow rust populations

and not virulent against the commercial wheat crops in use in the region (De Vallavieille-Pope et al., 2000; Villaréal et al., 2002), was increased in the laboratory according to standard procedures described in Villaréal et al. (2002). In brief, seedlings of wheat cv. Michigan Amber, highly susceptible to yellow rust, were grown in 5 x 5 cm square pots until the unfolding of the second leaf and sprayed with P. striiformis spores suspended in mineral oil (Soltrol® 130, Phillips Chemical Company, Phillips, TX). After 16 h in a dew chamber at 8°C in the dark, the pots with the inoculated seedlings were transferred in plastic batches (40 pots per batch, 20 batches) kept in spore-proof cabinets inside a greenhouse. Two weeks after inoculation, rust sporulated profusely on the seedlings. The batches were sealed into polyethylene bags and brought into the plots for inoculation on 4 April 2002 in Le Neubourg and on 5 April 2002 in Thiverval-Grignon. On each site, 10 batches of sporulating seedlings constituting the hereafter called "source plot" were set up in an area of the main plot from which all wheat seedlings had been removed with a hoe. In Le Neubourg, no prevailing wind directional effect was expected and the source plot was set up near the center of the main plot (see figure 4.1). In Thiverval-Grignon, prevailing western winds were expected to contribute mostly to spore dispersal and the source plot was skewed westwards from the center of the main plot (see figure 4.2). The batches were taken out of the polyethylene bags, which were immediately removed from the field and discarded. Two weeks later, the batches of sporulating seedlings were placed into sealed polyethylene bags, removed from the plots and discarded. In all cases, experimentators cared to avoid any contact with the trap plots. In the trap plots and the surrounding field sown with the resistant cultivar, wheat was at the end of tillering when the batches with sporulating seedlings were placed in the plots.

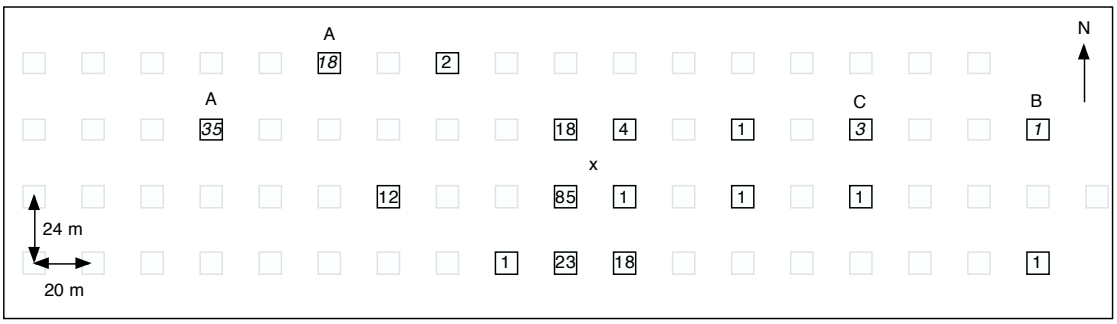


Fig. 4.1. Scale map of the field experiment performed in Le Neubourg. The location of the source plot is marked with X. Each square represents a 1 x 1 m trap plot. The size of the trap plots has been exaggerated for the sake of clarity. Numbers inside the squares are the numbers of diseased leaves collected, if any, at the end of the experiment. Numbers in italics indicate plots contaminated with exogenous inoculum. The letters stand for the identified pathotypes : A, 233E169 - V17 ; B, 233E137 - V17 ; C, 237E141 - V17. All other contaminated plots were found to be infected with pathotype 106E139 from the source plot.

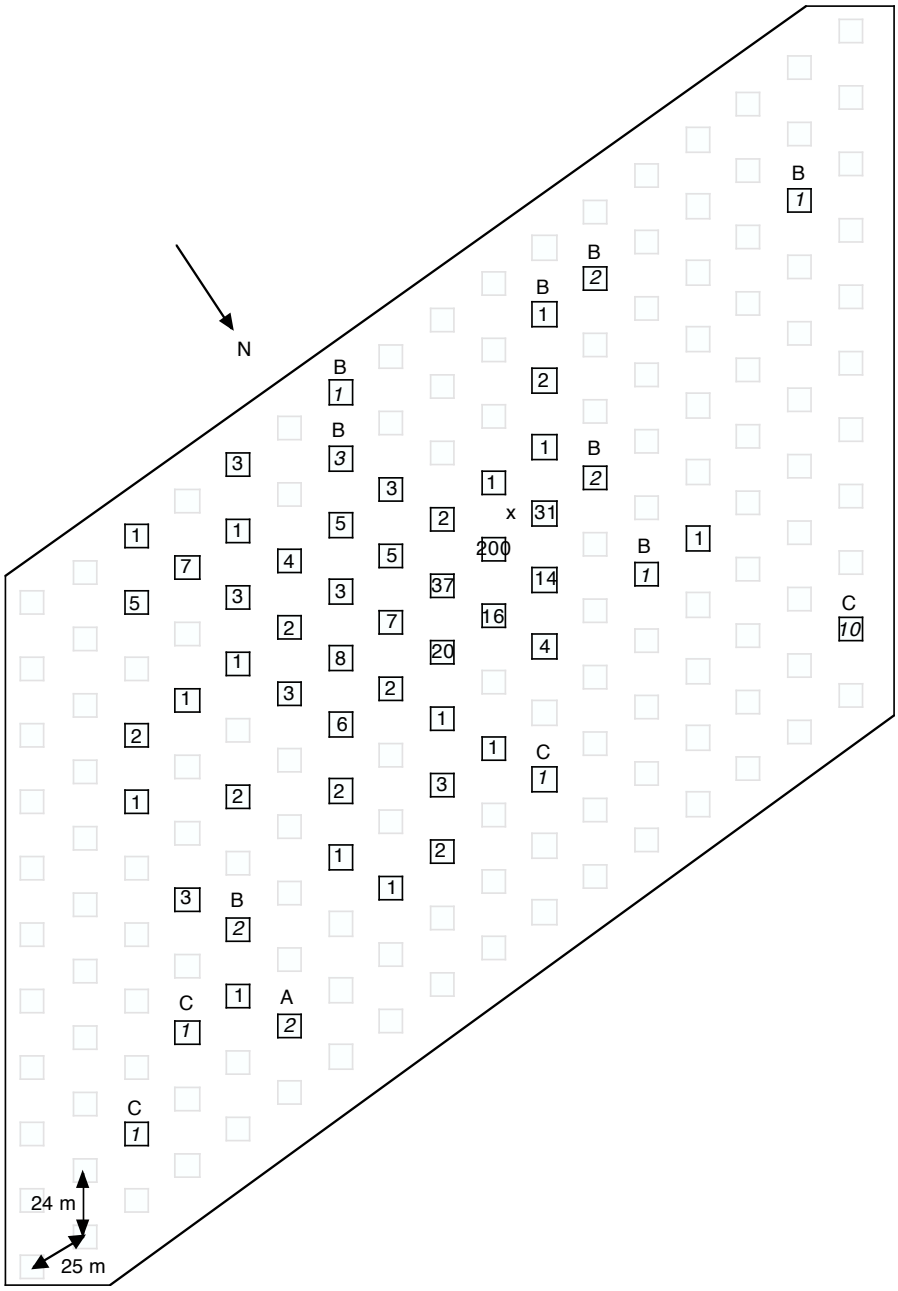


Fig. 4.2. Scale map of the field experiment performed in Thiverval-Grignon. The location of the source plot is marked with X. Each square represents a 1 x 1 m trap plot. The size of the trap plots has been exaggerated for the sake of clarity. Numbers inside the squares are the number of diseased leaves collected, if any, at the end of the experiment. Numbers in italics indicate plots contaminated with exogenous inoculum. The letters stand for the identified pathotypes : A, 45E140 ; B, 233E137 ; C, 237E141 - V17. All other contaminated plots were found to be infected with pathotype 106E139 from the source plot.

Disease assessment

All trap plots were checked for diseased leaves on 8 April in Le Neubourg and 10 April in Thiverval-Grignon. All diseased leaves were counted, excised and placed individually in a paper envelope. There was no attempt to count individual lesions on leaves since yellow rust does not form stripes on seedlings but groups of minute pustules covering the whole width of the leaf. The two fields were sprayed with a fungicide after the collection of the leaves.

Pathotype identification

Pathotypes were identified by the classic approach of determining the virulence or avirulence of the isolates on a set of differential hosts with known resistance genes or factors. The set of differential used for routine testing of the French isolates (De Vallavieille-Pope et al., 2000) is similar to the core set of differential cultivars recommended by the European yellow rust experts (COST 817 Bayles et al., 2000). Due to the low genetic diversity of the populations of yellow rust in France (De Vallavieille-Pope et al., 2000), it was assumed that the isolates harvested in the trap plots belonged either to the pathotype 106E139 used to inoculate the source plot or to one of the few pathotypes usually found in the most recent surveys of the French populations of yellow rust (M. Leconte, personal communication). These pathotypes can be distinguished with a subset of ten cultivars from the standard set of differential hosts table 4.1.

Differential host[(1)]	Gene[(2)]	Pathotype[(3)]				
		45E140	106E139	233E137-V17	233E169-V17	237E141-V17
Clement	*Yr9, Yr2*	a	a	V	V	V
Suwon 92/Omar	*Yr Suwon92/Omar*	a	V	V	V	V
Vilmorin 23	*Yr3*	V	V	V	V	V
Heines Kolben	*Yr6, Yr2*	V	a	a	V	V
Lee	*Yr7*	a	V	a	a	V
Chinese 166	*Yr1*	V	a	V	V	V
Carstens V	*Yr Carstens V*	a	a	a	a	a
Compair	*Yr8*	a	a	a	a	a
Hybrid 46	*Yr4*	a	V	V	a	V
VPM 1	*Yr17*	a	a	V	V	V

Tab. 4.1. Identification of the pathotypes collected in the experimental plots after inoculation of a subset of the standard set of differential hosts used to identify pathotypes of yellow rust in France. V : virulent ; a : avirulent on le 0-4 scale (McIntosh et al., 1995). (1) See the complete standard set in Bayles et al. (2000) and De Vallavieille-Pope et al. (2000). (2) Nomenclature of the genes for resistance to yellow rust according to McIntosh et al. (1995). (3) Pathotype designation according to Johnson et al. (1972) ; -V17 : virulent for *Yr17*.

Spores were increased by rubbing the leaf collected in the field against the primary leaf of a healthy seedling of wheat cv. Victo in order to produce the quantity of inoculum required for tests. The inoculated seedlings were handled as described above. Two weeks

after inoculation, the newly formed spores were vacuumed and kept in a dessiccator filled with glycerol until use. Inoculation and incubation of the subset of cultivars with the spores to be tested followed the procedure described above. The inoculated seedlings were then placed into a growth chamber at 17°C for a 16-h light period with a light intensity of 300 mol quanta m^{-2} s^{-1} and at 14°C for an 8-h dark period (Villaréal et al., 2002). Infection types on the differential hosts were assessed 15 days after inoculation using a 0-4 assessment scale (McIntosh et al., 1995). Compatible interactions (host susceptibility and pathogen virulence) were defined as infection types 3 (abundant sporulation, with chlorosis) and 4 (abundant sporulation without chlorosis).

4.2.2 Notations and definitions of the anisotropy functions

Notations used in the statistical analysis

Arbitrarily, the centre of the source plot is assigned at the origin of the planar space, i.e. the experimental field. Trap plots are indexed by i in $\{1, \ldots, I\}$, where I is the number of trap plots in an experimental field ($I = 72$ for Le Neubourg and $I = 187$ for Thiverval-Grignon). Locations of the centres of trap plots are denoted by $x_1, \ldots, x_I$ in Cartesian coordinates. Total numbers of leaves in trap plots are denoted by $n_1, \ldots, n_I$. Numbers of leaves diseased by the inoculated pathotype 106E139 in trap plots $1, \ldots, I$ are denoted by $y_1, \ldots, y_I$. Only leaves diseased by pathotype 106E139 are taken into account in the statistical analysis since we are interested in assessing the dispersal only from the source plot.

For a spore dispersed from the origin of the planar space, let (Φ, R) denote, in polar coordinates, the random location of deposition of the spore. Φ is called the random dispersal direction. R is called the random dispersal distance.

Upper characters Y_i, Φ and R are used to denote random variables whereas lower characters y_i, ϕ and ρ are used to denote deterministic or observed values.

Directional dispersal function

Let f denote the directional density function. For direction ϕ in $[0^o, 360^o[$, $f(\phi)d\phi$ is the probability that the random dispersal direction Φ is in $[\phi - d\phi/2, \phi + d\phi/2]$, where $d\phi$ is an infinitesimal arc length. Such an event occurs if the spore is deposited in an increasing width plot with angle $d\phi$ radiating from the source in direction ϕ (see figure 4.3). f is the probability density function of the random direction Φ and reflects how directions are taken by spores.

Mean distance function

Let the mean distance function describe, for any direction ϕ in $[0^o, 360^o[$, how far a spore travels in average. Two versions, say $g^<$ and $g^=$, of this function are considered. First, $g^<(\phi)$ is the mean dispersal distance of spores deposited over an increasing width plot centred

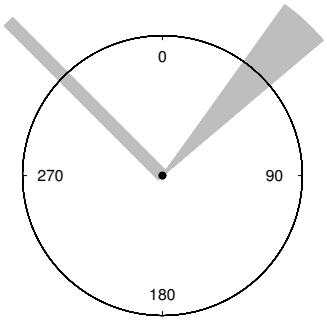


Fig. 4.3. Graphical representations of two sorts of plots radiating from the source location (black dot). Left gray shape : constant width plot ; right gray shape : increasing width plot.

at ϕ (see figure 4.3). Second, $g^{=}(\phi)$ is the mean dispersal distance of spores deposited over a constant width plot radiating from the source in direction ϕ (see figure 4.3). On an experimental point of view, the version $g^{=}$ is the one which can be generally estimated. Indeed, when a disease gradient in a given direction ϕ is assessed in Sackett and Mundt (2005) for instance, the mean dispersal distance which could be estimated corresponds to $g^{=}(\phi)$ because sampling units which are lined up have common shapes and common sizes.

4.2.3 Nonparametric framework

Kernel estimator for the directional dispersal function f

A kernel density estimator, say $\hat{f}$, is built to estimate the directional density function f. Let $\Phi_1, \ldots, \Phi_N$ be random dispersal directions of N spores. Assume $\Phi_1, \ldots, \Phi_N$ are independent and identically drawn from the probability distribution f. The kernel density estimator of $f(\phi)$, $\phi \in [0^o, 360^o[$, is defined by

$$\hat{f}(\phi) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{b} K\left(\frac{\phi - \Phi_n}{b}\right), \tag{4.1}$$

where K is called kernel function and b is called bandwidth (Bosq and Lecoutre, 1987; Fisher, 1995; Silverman, 1986). Each dispersal direction Φ_n contributes to the value of $\hat{f}(\phi)$ at a level depending on the closeness between ϕ and Φ_n. Closer Φ_n from ϕ is, larger its contribution is. Therefore, if there are a lot of Φ_n close to ϕ then the value of the estimated density $\hat{f}$ at ϕ is high. On the contrary, if there are few Φ_n close to ϕ then the value of the estimated density $\hat{f}$ at ϕ is low. The closeness is defined through the kernel function K and the bandwidth b.

Kernel estimator for the mean distance function $g^{<}$

A kernel smoother, say $\hat{g}^{<}$, is used to estimate the mean distance function $g^{<}$. Let $\Phi_1, \ldots, \Phi_N$ and $R_1, \ldots, R_N$ be, respectively, random dispersal directions and distances of

N spores. Assume couples $(\Phi_1, R_1), \ldots, (\Phi_N, R_N)$ independent and identically distributed. The kernel smoother of $g^<(\phi)$, $\phi \in [0^o, 360^o[$, is defined by

$$\hat{g}^<(\phi) = \sum_{n=1}^{N} \frac{K\left(\frac{\phi - \Phi_n}{b}\right)}{\sum_{m=1}^{N} K\left(\frac{\phi - \Phi_m}{b}\right)} R_n, \tag{4.2}$$

where K is the kernel function and b is the bandwidth (Bosq and Lecoutre, 1987; Hastie and Tibshirani, 1990). The estimator $\hat{g}^<(\phi)$ of the expected dispersal distance is a weighted mean of dispersal distances $R_1, \ldots, R_N$. Closer Φ_n from ϕ is, larger the weight of R_n in the mean is.

Choices of the kernel function and the bandwidth

For both the kernel density estimator $\hat{f}$ and the kernel smoother $\hat{g}^<$, the kernel function

$$K(\phi) = \begin{cases} 0.9375(1-\phi^2)^2 & \text{if } -1 \leq \phi \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

suggested by Fisher (1995, p. 26) was used, and the bandwidth b was selected by cross-validation (see Hastie and Tibshirani, 1990; Loader, 1999).

Computation of the kernel estimators

Ideally, the kernel density estimator and the kernel smoother should be computed by randomly sampling N spores and by observing their deposition locations $(\Phi_1, R_1), \ldots, (\Phi_N, R_N)$. However, data at our disposal consists in numbers of diseased leaves in trap plots. Therefore, the estimators of the anisotropy functions were computed as follows. The planar space was partitioned into rectangles centred at the trap plots (see figure 4.4) and the observed number of diseased leaves in a trap plot was assumed to be proportional to the number of spores deposited in its associated rectangle. For each index i, y_i locations were generated independently and uniformly in the rectangle associated with trap plot i. $N = \sum_{i=1}^{I} y_i$ locations, say $(\Phi_1, R_1), \ldots, (\Phi_N, R_N)$, were obtained and estimators $\hat{f}$ and $\hat{g}$ (equations (4.1) and (4.2)) were computed using these locations. As the estimates of f and g depends on the generated locations, the procedure was repeated 100 times. 100 estimates for f and 100 estimates for g were obtained. For each anisotropy function, the final estimate was the mean of the 100 estimates obtained.

4.2.4 Parametric framework

A parametric model is proposed to estimate both the directional density function f and the mean distance function $g^=$. Firstly, an underlying model integrating f and $g^=$ is defined to model numbers of lesions on leaves. Then, the underlying model is derived to get a model adapted to our data which consists in numbers of diseased leaves in trap plots.

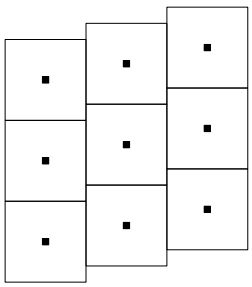


Fig. 4.4. Trap plots (small filled rectangles) and their associated rectangles (large rectangles) which partition the planar space.

Underlying model

An underlying model describes the number of lesions per leaf. For any leaf located at x, the number of lesions is assumed to follow a Poisson distribution with mean $S(x)$, where $S(x)$ is called infectious potential at x. $S(x)$ measures the expected quantity of spores which are deposited at location x in the planar space and which succeed to infect.

The infectious potential at location x, generated by a point source located at the origin of the planar space, is defined in function of the anisotropy functions f and $g^{=}$ by

$$S(x) = s_0 \, \frac{f(\phi)}{g^{=}(\phi)^2} \exp\left\{-\frac{\rho}{g^{=}(\phi)}\right\}, \tag{4.3}$$

where (ϕ, ρ) are the polar coordinates of x. In equation (4.3), parameter s_0 represents the strength of the source of spores. The quantity $S(x)dx/s_0$ is the probability that a spore dispersed from the source at the origin lands in the infinitesimal rectangular area dx centred at x. The infectious potential function is the 2-dimensional version of the exponential dispersal gradient (McCartney and Fitt, 1985; Aylor, 1990). Note that in this setting, it can be shown that $g^{<}(\phi) = 2g^{=}(\phi)$ for any ϕ in $[0^o, 360^o[$.

Parametric forms for the anisotropy functions

Results of the nonparametric estimation of the anisotropy functions (see section 4.3 below) suggested that unimodal parametric forms could be used. The directional dispersal function f is assumed to be a von Mises function parameterized by μ and σ (Fisher, 1995, p. 48)

$$f(\phi) = \{2\pi I_0(\sigma)\}^{-1} \exp\{\sigma \cos(\phi - \mu)\} \tag{4.4}$$

$$I_0(\sigma) = (2\pi)^{-1} \int_0^{2\pi} \exp\{\sigma \cos(\theta - \mu)\} d\theta. \tag{4.5}$$

This is an unimodal density distribution over $[0^o, 360^o[$. μ is the expected dispersal direction and σ controls the variability of dispersal directions around μ. $\sigma = 0$ means that dispersal

directions are uniformly distributed in $[0^o, 360^o[$, and $\sigma = \infty$ means that dispersal directions are all equal to μ. Figure 4.5 shows different von Mises functions for different values of σ. Von Mises functions are unimodal functions, i.e. there is only one main dispersal direction.

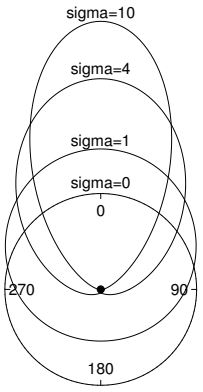


Fig. 4.5. Von Mises functions for different values of the parameter σ ("sigma") which controls the variability around the mean 0^o.

The distance dispersal function $g^=$ is assumed to be proportional to a von Mises function parameterized by ν and κ

$$g^=(\phi) = g_0\{2\pi I_0(\kappa)\}^{-1} \exp\{\kappa \cos(\phi - \nu)\}. \tag{4.6}$$

ν is the direction at which the expected dispersal distance is the larger and equals $g_0 \exp(\kappa)/\{2\pi I_0(\kappa)\}$. κ controls the variability of the expected dispersal distance as the dispersal direction varies. $\kappa = 0$ means that dispersal distances are the same whatever the dispersal directions. g_0 is a multiplicative constant and measures how far spores travel.

Derivation of a model adapted to the data

The previous paragraphs define an underlying model which models numbers of lesions on leaves. A model, adapted to our data, was derived from it. Let us describe it. Suppose the source plot is small enough to be viewed as a point source. Then, it generates an infectious potential with the form (4.3). Consider trap plot i centred at x_i. Suppose the trap plot is small enough to consider that the infectious potential S is constant over the trap plot and equals $S(x_i)$. Since the number of lesions on a leaf at x_i is Poisson distributed with mean $S(x_i)$, the number of diseased leaves Y_i in trap plot i follows a binomial distribution with size n_i and probability $p(x_i)$

$$Y_i \sim \text{Binomial}\{n_i, p(x_i)\}, \tag{4.7}$$

where

$$p(x_i) = 1 - \exp\{-S(x_i)\}. \tag{4.8}$$

Remember that n_i is the total number of leaves located in trap plot i. $p(x_i)$ equals $1 - \exp(-S(x_i))$ because it is the probability for a leaf at x_i to be diseased or, in other words, the probability that at least one lesion is on the leaf.

Estimation, tests and confidence intervals for the parameters

In the model, there are 6 parameters : the source strength s_0, the direction parameters μ and σ, and the distance parameters ν, κ and g_0. Let the vector of parameters be $\lambda = (s_0, \mu, \sigma, \nu, \kappa, g_0)$. λ is estimated by maximum of likelihood (Rohatgi, 2003). Intuitively, the maximum likelihood estimator, say $\hat{\lambda} = (\hat{s_0}, \hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\kappa}, \hat{g_0})$, of λ is the set of parameters which has the most support of data in the likelihood sense. Formally, $\hat{\lambda}$ maximizes the loglikelihood function

$$l(\lambda) = \sum_{i=1}^{I} \log \mathbb{P}_\lambda(Y_i = y_i),$$

where

$$\mathbb{P}_\lambda(Y_i = y_i) = C_{n_i}^{y_i} p(x_i)^{y_i} \{1 - p(x_i)\}^{n_i - y_i}$$

is the probability that a binomial variable with size n_i and probability $p(x_i)$ (equation (4.8)) equals y_i. Maximization of the loglikelihood function is made using a Newton-Raphson algorithm. The maximum likelihood estimate of f is obtained by replacing μ and σ by $\hat{\mu}$ and $\hat{\sigma}$ in equation (4.4). The maximum likelihood estimate of g is obtained by replacing ν, κ and g_0 by $\hat{\nu}$, $\hat{\kappa}$ and $\hat{g}_0$ in equation (4.6). The probability function p is estimated by replacing λ by $\hat{\lambda}$ in expressions (4.3) and (4.8).

Likelihood ratio tests (LRT, Dacunha-Castelle and Duflo, 1982; Rohatgi, 2003) were performed to see (i) whether the directional density function is uniform ($\sigma = 0$) or not ($\sigma > 0$) and (ii) whether the mean distance function is constant over $[0^o, 360^o[$ ($\kappa = 0$) or not ($\kappa > 0$). For each test, the LRT statistic was parametrically bootstrapped 200 times to get its distribution under the null hypothesis (McLachlan, 1987). The p-value of the test is given by the proportion of bootstrap LRT statistics over the observed LRT statistic.

95%-confidence intervals for parameters s_0, μ, σ, ν, κ and g_0 were computed using parametric bootstrap (Efron and Tibshirani, 1993). For each site, 200 independent bootstrap datasets were simulated under the fitted parametric model. For each bootstrap dataset, the maximum likelihood estimators of the parameters were computed. For any parameter, 200 bootstrap estimates were obtained ; the endpoints of the 95%-confidence interval are the 5th and the 195th ordered values of the 200 bootstrap estimates.

Goodness-of-fit statistics were computed to see whether the observed number of leaves for Le Neubourg and Thiverval-Grignon are suitable values under the fitted parametric models. For each site, 1000 samples were simulated from the corresponding fitted model. For each trap plot, 1000 simulated numbers of diseased leaves were obtained. Two goodness-of-fit statistics were defined. First, the percentage of trap plots for which the observed numbers of diseased leaves were between the 25th and the 975th ordered values of the simulated

numbers of diseased leaves. Second, the percentage of trap plots for which the observed numbers of diseased leaves were between the minimum and the maximum simulated number of diseased leaves. High percentages reflect good fits.

Computations and graphics were made using the R Statistical Software (http ://www.r-project.org).

4.3 Results

Field experiment

In Le Neubourg, diseased leaves were found in 17 of the 72 trap plots (23.6%). The number of diseased leaves per plot ranged from 1 to 85 ; seven plots had only one diseased leaf, three plots had 2-5 diseased leaves, six plots had 12-35 diseased leaves and one plot had 85 diseased leaves. Spores collected in a given trap plot belonged to a single pathotype, regardless of the number of diseased leaves per plot. The inoculated pathotype 106E139 was identified in 13 of the 17 infected trap plots (76.5%). Two of the four naturally contaminated plots were infected by pathotype 233E169-V17, which formed two large foci with 18 and 35 diseased leaves, respectively. The two other naturally contaminated trap plots were infected by pathotype 233E141-V17 (one diseased leaf) and 233E137-V17 (three diseased leaves). The maximal dispersal distance of pathotype 106E139 was 175 m from the source plot. Figure 4.2 shows the spatial pattern of the numbers of diseased leaves for Le Neubourg.

In Thiverval-Grignon, diseased leaves were found in 56 of the 187 trap plots (29.9%). The number of diseased leaves per trap plot ranged from 1 to > 200 ; 20 plots had only one diseased leaf, 30 plots had 2-10 diseased leaves, five plots had 14-37 diseased leaves and one plot had > 200 diseased leaves. Spores collected in a given trap plot belonged to a single physiologic race, regardless of the number of diseased leaves per plot. The inoculated pathotype 106E139 was identified in 43 out of the 56 infected trap plots (76.8%). Eight of the 13 naturally contaminated plots were infected by pathotype 233E137, four by pathotype 237E141-V17 and one by pathotype 45E140. Pathotype 237E141-V17 formed one large disease focus (10 infected leaves) whereas the other naturally contaminated trap plots had only 1-3 diseased leaves. The maximal dispersal distance of pathotype 106E139 was 225 m from the source plot. Figure 4.1 shows the spatial pattern of the numbers of diseased leaves for Thiverval-Grignon.

Nonparametric Estimation

Figures 4.6(a) and (c) show the nonparametric estimates of the directional density functions f for Le Neubourg and Thiverval-Grignon. Figures 4.7(a) and (c) show the nonparametric estimates of the mean distance functions $g^{<}$ for Le Neubourg and Thiverval-Grignon. The shapes of these functions suggest that the directional density functions and the mean

distance functions are more or less unimodal. Thus, the von Mises form proposed for the anisotropy functions in the parametric model seems to be appropriate.

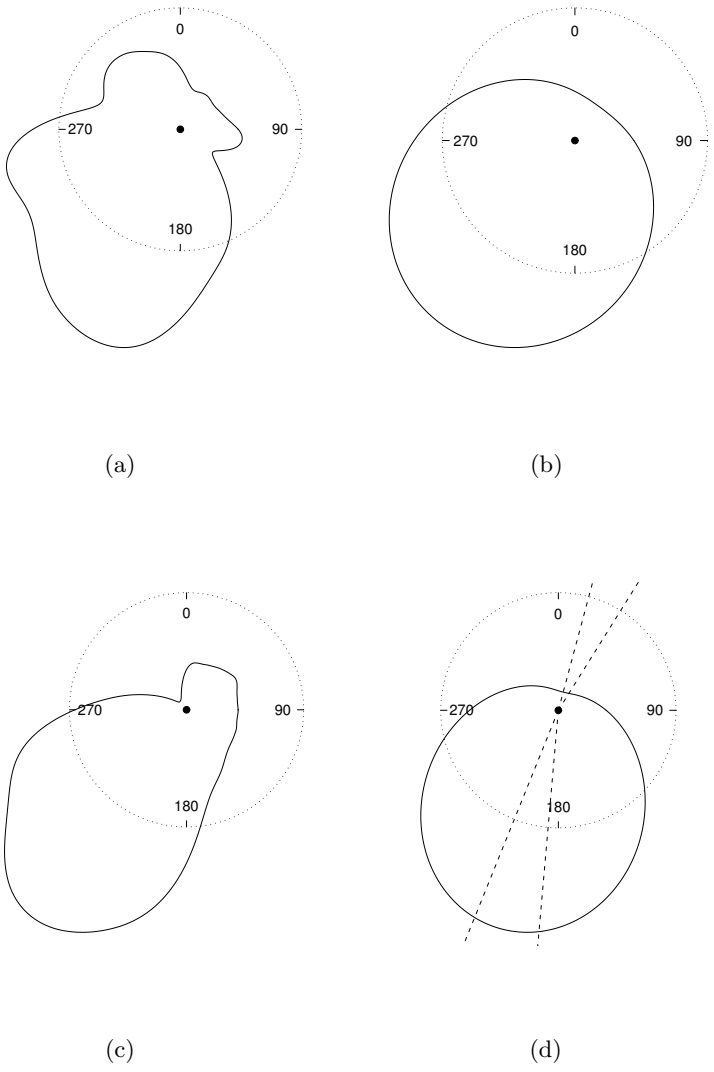


Fig. 4.6. Directional density functions (solid lines) for Le Neubourg (top) and Thiverval-Grignon (bottom). Left : nonparametric estimates ; right : parametric estimates. Dotted lines : uniform directional density function. Point : Source location.

Parametric Estimation

Figures 4.6(b) and (d) show the parametric estimates of the directional density functions f for Le Neubourg and Thiverval-Grignon. The parametric estimates are smoother than the nonparametric estimates (figures 4.6(a) and (c)) but show more or less the same main dispersal direction, i.e. spores are mainly dispersed to the bottom - bottom left direction for both experiments. Note that, for Le Neubourg and Thiverval-Grignon, the likelihood

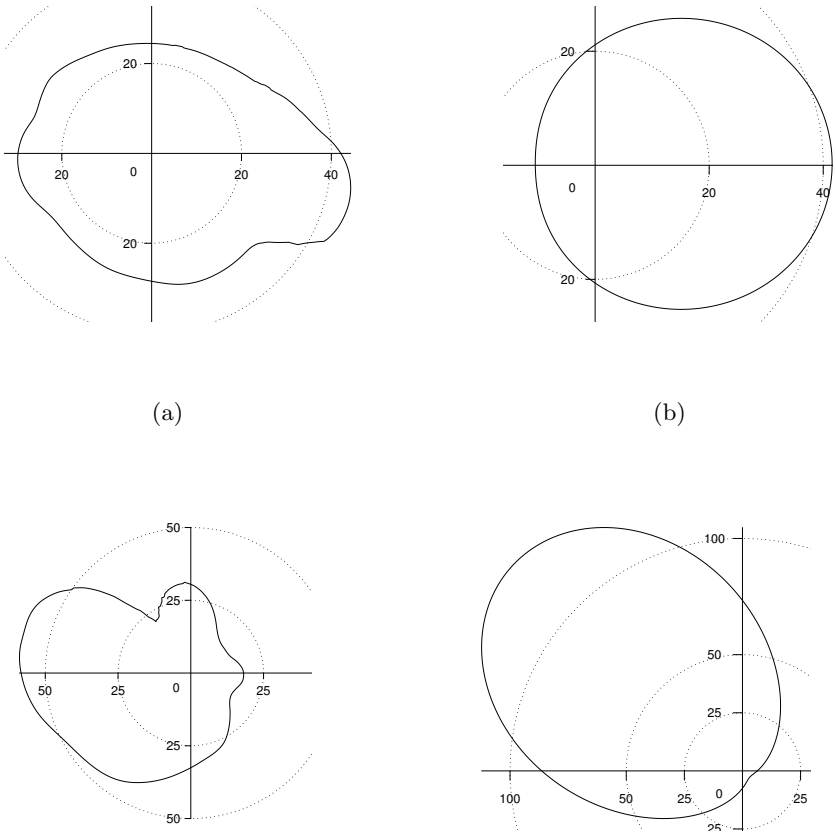


Fig. 4.7. Mean distance functions (solid lines) for Le Neubourg (top) and Thiverval-Grignon (bottom). Left : nonparametric estimates ; right : parametric estimates. Origin : source location ; unit : meter.

ratio tests reject at the risk level 5% the uniformity of the directional density functions (see table 4.2). To illustrate the departure of the directional density function from the uniform density, the estimated percentages of spores deposited over the two sectors drawn in figure 4.6(d) were computed. The estimated percentage for the bottom orientated sector is 18% instead of 5%. The estimated percentage for the top orientated sector is 0.1% instead of 5%.

	Le Neubourg	Thiverval-Grignon
M00 vs M10	0	0
M00 vs M01	0	0
M10 vs M11	0.04	0
M01 vs M11	0	0

Tab. 4.2. Bootstrap likelihood ratio tests. M00 : model with uniform directional density function and constant mean distance function ($\sigma = 0$ and $\kappa = 0$) ; M10 : model with constant mean distance function ($\sigma > 0$ and $\kappa = 0$) ; M01 : model with uniform directional density function ($\sigma = 0$ and $\kappa > 0$) ; M11 : complete model ($\sigma > 0$ and $\kappa > 0$).

Figures 4.7(b) and (d) show the parametric estimates of the directional density functions $g^{=}$ for Le Neubourg and Thiverval-Grignon. At a first sight, the parametric estimates differ from the nonparametric estimates (figures 4.7(a) and (c)) especially because the nonparametric estimates are estimates of $g^{<}$. However, for both experiments, the directions at which the mean dispersal distance is the largest are more or less the same. Note that, for Le Neubourg and Thiverval-Grignon, the likelihood ratio tests reject at the risk level 5% the constancy of the mean distance functions (see table 4.2).

Let us go further in the description of the mean distance function for Thiverval-Grignon. The mean distance function estimated from the parametric model was drawn together with the map of disease measures in figure 4.8. The bottom part of the mean distance function seems correctly estimated. Now consider what happened for the top part of the function. Almost no diseased leaves have been sampled at the top and top-left of the source. It is confirmed by figure 4.6(d). So, the top part of the estimated mean distance function must be considered with care.

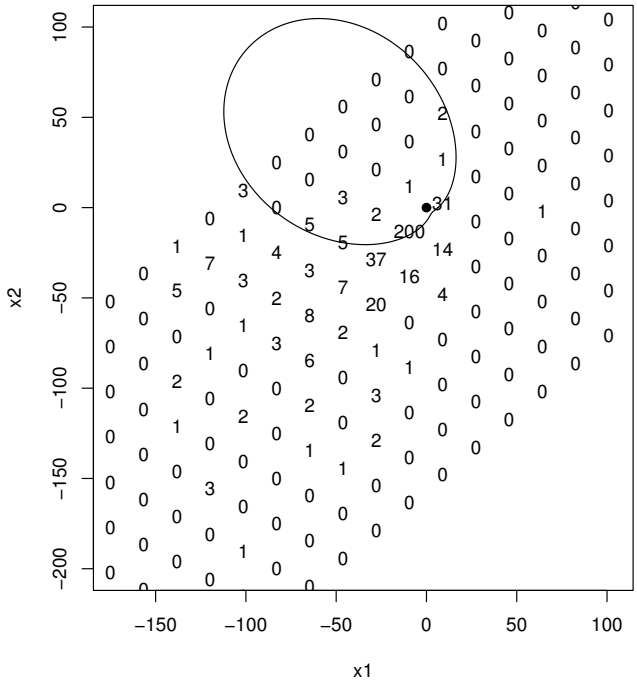


Fig. 4.8. Mean distance function (solid lines) estimated under the parametric model together with the map of disease measures for Thiverval-Grignon. Point at the origin : source location ; unit : meter.

Finally, for Le Neubourg and Thiverval-Grignon, the directional density functions on the one hand and the mean distance functions on the other hand do not coincide, that is directions preferred by spores (figure 4.6) do not correspond to directions at which the mean dispersal distance is the largest (figure 4.7). The difference is significant since the

95%-confidence intervals for parameters μ and ν do not overlap (see table 4.3 which shows estimates and confidence intervals of the parameters).

	Le Neubourg		Thiverval-Grignon	
	Estimate	CI	Estimate	CI
s_0	168	(143 ;208)	1371	(840 ;1377)
μ	216	(201 ;227)	194	(184 ;200)
σ	1.69	(1.28 ;2.09)	2.49	(2.41 ;3.37)
ν	89	(58 ;131)	311	(296 ;321)
κ	0.69	(0.41 ;1.14)	1.73	(1.51 ;1.98)
g_0	147	(99 ;270)	280	(164 ;375)

Tab. 4.3. Estimates and 95%-confidence intervals (CI) of parameters s_0, μ, σ, ν, κ and g_0 for Le Neubourg and Thiverval-Grignon.

Estimation of the probabilities for leaves to be infected

Figures 4.9 and 4.10 show the maps of probabilities for leaves to be infected (equation (4.8)) obtained by estimating the parametric model. This sort of maps, by combining the directional density function and the mean distance function estimated under the parametric model, allows to assess the resulting anisotropy.

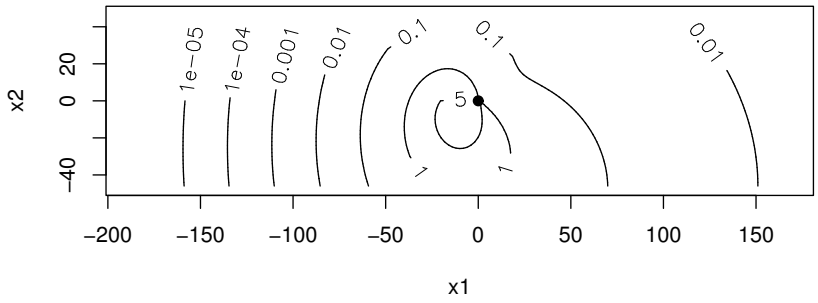


Fig. 4.9. Probabilities (%), estimated under the parametric model, for leaves to be infected by the source plot in Le Neubourg. Point : source location ; unit : meter.

Goodness-of-fit statistics

Table 4.4 shows that about 90% of the observed numbers of diseased leaves are suitable values under the fitted parametric models. These statistics do not take into account the uncertainty about the parameters due to the estimation. Taking into account the uncertainty should increase the percentages. So the percentages in table 4.4 underestimate the goodness-of-fit.

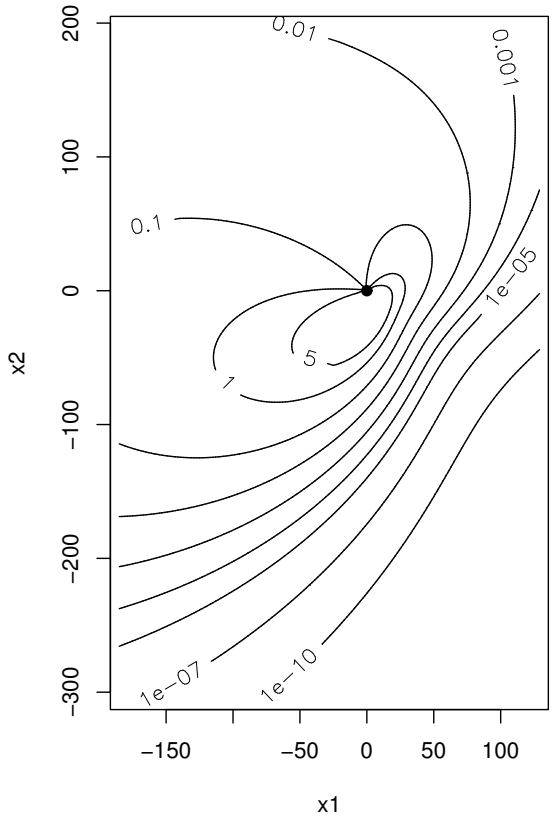


Fig. 4.10. Probabilities (%), estimated under the parametric model, for leaves to be infected by the source plot at Thiverval-Grignon. Point : source location ; unit : meter.

Le Neubourg	Thiverval-Grignon
89%	90%
97%	94%

Tab. 4.4. Goodness-of-fit statistics for Le Neubourg and Thiverval-Grignon. Percentages of trap plots for which the observed numbers of diseased leaves were between the 25th and the 975th ordered values of the 1000 simulated numbers of diseased leaves (1st row). Percentages of trap plots for which the observed numbers of diseased leaves were between the minimum and the maximum simulated number of diseased leaves (2nd row).

4.4 Discussion

Dispersal experiments in large field plots

The experimental design made of an array of small susceptible wheat plots scattered over a large resistant wheat plot allowed us to follow the dispersal of yellow rust from a source plot during a single generation of the pathogen. The discontinuous pattern of the trapping area decreased the number of data (infected leaves) compared to the data which would have been obtained with a continuous sampling scheme. However, discontinuous sampling on a

regular grid allowed us to observe disease all over the experimental area without strong presupposition on the main distance of dispersal.

Mapping disease dispersal up to a distance of 125 m and 225 m from the source in Le Neubourg and Thiverval-Grignon respectively was possible in spite of a fairly small trapping area. This is far beyond distances on which yellow rust dispersal has already been documented in field experiments (Emge and Shrum, 1976; Sackett and Mundt, 2005), and also far beyond the distance on which faba bean rust dispersal was studied using a similar, discontinuous design (Sache and Zadoks, 1996). Based on the knowledge of the genetic structure of the populations of the pathogen, we were able to identify contamination from external sources of inoculum ; on both experimental sites, less than 25% of the infected trap plots were contaminated by such sources of inoculum. Since the experiment was not designed to quantify external contamination, no attempt was made to interpret the level of this contamination and to trace back its origin. Due to the well-established long-distance dispersal ability of rust spores (Brown and Hovmøller, 2002), the external spores might come from a neighbouring field as well as from a very distant region. Migration of yellow rust spores over long distances, e.g. from England, has been indeed documented in the regions where we did the experiment (Bayles et al., 2000).

Anisotropy of spore dispersal

The anisotropy of disease dispersal was evidenced on maps showing the results of the two experiments (figures 4.1 and 4.2). The differences in the pattern of disease dispersal between the two sites were caused by local differences in weather conditions during the experiment. As disease dispersal is the result of spore dispersal and leaf infection, the anisotropy of disease dispersal could be caused by (i) anisotropy of airborne spore dispersal and/or (ii) variations of leaf receptivity from trap plot to trap plot. Individual variations in the leaf receptivity were included in a model describing the dispersal of wheat leaf rust on short distance around a diseased leaf (chapter 3). In the present, large-scale experiment, we considered that these individual variations were negligible and that the average leaf receptivity in a trap plot was the same all over the experiment area. Accordingly, only cause (i) was integrated in the modelling framework. The purpose of the modelling approach was not to perform a mechanistic interpretation of the local variations in spore dispersal ; rather, its purpose was to propose a generic modelling framework suitable for the analysis of anisotropy of airborne spore dispersal in a various range of experimental situations, various in terms of weather conditions and sampling schemes. Statistical tests (table 4.2) confirmed the significance of the anisotropy that was visually evidenced on maps showing the results of the experiments (figures 4.1 and 4.2).

Anisotropy functions

The anisotropy of spore dispersal in direction and in distance is not clearly visible on maps showing the results of the experiments, however, it was shown and quantified by the statistical analysis. To characterize the anisotropy of spore dispersal in direction and in

distance, two anisotropy functions were introduced and estimated : the directional density function and the mean distance function. The statistical analysis performed for the two experiments shown that the anisotropy functions can differ in their orientation. Factors such as the temporal sequence of the wind and the topography might explain the difference.

The anisotropy functions are complementary tools with the wide used disease gradient to quantify dispersal in a field experiment. On one hand, the disease gradient allows to assess the decrease of disease concentration with distance from the spore source, on the other the anisotropy functions allow to assess how directions are preferred by spores and how far in average spore travels in the various directions.

The anisotropy functions are defined without setting any model. However, for estimating them, two models were proposed : a nonparametric one and a parametric one. Data from the field experiments presented in this paper did not allow to get accurate nonparametric estimators of the anisotropy functions because, ideally, data should be deposition locations for a set of randomly sampled spores. That is the reason why the nonparametric approach was used to figure out what are the main directions and the shapes of the anisotropy functions. Such information were used to specify the form of the anisotropy functions in the parametric model. Unimodal von Mises functions were used but more flexible functions such as mixtures of von Mises functions and elliptical functions could be used. However such functions require more parameters and problems of identifiability may occur. Contrary to the nonparametric approach, the parametric one was adapted to the sampling scheme used in the field experiments. Note that increasing the quantity of data, that is increasing the quantity of trap plots, should allow to estimate more flexible parametric anisotropy functions and allow to get more accurate estimators. Note also that the estimations of the anisotropy functions are conditionnal on the ranges of distances which are observed. For instance, if the sampling units had been distributed at shorter distances from the source, we would have obtained different estimates of the anisotropy functions.

5

Bilan sur nos modèles de propagation : éléments intégrés et non intégrés

Les chapitres 3 et 4 présentent deux expériences de propagation spatiale des rouilles du blé. Pour analyser les données issues des expériences, deux modèles paramétriques ont été développés. Ces modèles permettent d'inférer sur la dispersion des spores, l'hétérogénéité des feuilles face à l'infection et l'anisotropie de la dispersion. Dans cette section, nous proposons un cadre commun de modélisation au sein duquel se situent les deux modèles. Ce cadre intègre des éléments connus de la propagation spatiale des maladies aériennes des végétaux et permet de construire des modèles adaptés aux données. Mais la capacité de ces modèles à résumer les données a des limites à cause de la non intégration de certains éléments, connus ou inconnus, qui jouent un rôle dans le phénomène étudié. Dans ce chapitre, nous discutons de la détection de tels éléments et abordons le sujet de leur caractérisation, sujet qui est traité plus profondément dans la deuxième partie de ce document.

5.1 Un cadre commun de modélisation

5.1.1 Modèle sous-jacent pour le nombre de lésions sur une feuille

Variables

Soit $N_1, \ldots, N_I$ les nombres de lésions sur I feuilles dont les positions sont $X_1, \ldots, X_I$ dans le plan $\mathbb{R}^2$. Soit $Z_1, \ldots, Z_I$ les fragilités caractérisant les feuilles.

Potentiel infectieux et fonction de dispersion

Les feuilles sont soumises à l'effet d'une source ponctuelle unique située en 0 dans le plan. Notons $S(x)$ la valeur en $x \in \mathbb{R}^2$ du potentiel infectieux créé par la source,

$$S(x) = s_0 h(x), \tag{5.1}$$

où s_0 représente la force de la source et $h(\cdot)$ la fonction de dispersion, c'est-à-dire la densité de probabilité pour la position aléatoire de dépôt d'une spore émise par la source. La forme de $h(\cdot)$ est donnée par

$$h(x) = \frac{f(\phi_x)}{g(\phi_x)^2} \exp\left\{-\frac{||x||}{g(\phi_x)}\right\}, \tag{5.2}$$

où $(\phi_x, ||x||)$ sont les coordonnées polaires de x, et f et g sont les fonctions d'anisotropie. Notons $(\Phi, R) \in [0, 2\pi[\times\mathbb{R}^+$ les coordonnées polaires de la position aléatoire d'une spore émise par la source. La fonction $f(\cdot)$ est la densité de probabilité de Φ. La quantité $g(\Phi)$ est l'espérance conditionnelle de la distance R sachant que la position aléatoire de dépôt (Φ, R) est dans une demi-bande dont l'origine est en 0, dont l'orientation est Φ et dont la largeur est infinitésimale (cf. figure 4.3).

Distribution des nombres de lésions $N_1, \ldots, N_I$

Les nombres de lésions $N_1, \ldots, N_I$ sachant les positions $X_1, \ldots, X_I$ et les fragilités $Z_1, \ldots, Z_I$ sont supposées être des variables aléatoires indépendantes distribuées selon des lois de Poisson de moyennes $Z_i S(X_i)$, $i = 1, \ldots, I$

$$N_i | X_i, Z_i \ \sim \ \text{indép. Poisson}\{Z_i S(X_i)\}, \quad i = 1, \ldots, I.$$

On a donc

$$\mathbb{P}(N_i = n | X_i, Z_i) = \exp\{Z_i S(X_i)\} \frac{\{Z_i S(X_i)\}^n}{n!}.$$

5.1.2 Dérivation du modèle sous-jacent pour obtenir des modèles adaptés aux données

Selon la nature des données et leur quantité, le modèle sous-jacent décrit dans la section précédente est inadapté ou pas assez parcimonieux. Toutefois, il peut être dérivé afin de parvenir à un modèle adapté et parcimonieux. Les paragraphes suivant montrent, dans différents cas, comment la dérivation peut être effectuée.

Mesures alternatives de la maladie

Le chapitre 3 traite de données représentant des nombres de lésions sur feuilles. Mais d'autres mesures de la maladie peuvent être utilisées (nombre de lésions sur un ensemble de feuilles, surface foliaire atteinte sur une feuille, sur un ensemble de feuilles, dans un mètre carré, feuille malade ou pas, nombre de feuilles malades dans un mètre carré (mesure utilisée au chapitre 4), champ contenant des plantes malades ou pas). La mesure choisie dépend par exemple de l'échelle à laquelle l'expérimentateur s'intéresse[1] ou encore de la possibilité d'obtenir telle où telle mesure[2].

Traitons l'exemple de la mesure indiquant si la feuille est malade ou pas, i.e. indiquant la présence ou l'absence de lésions sur la feuille. Notons Y_i cette variable pour la feuille i et posons que Y_i vaut 1 si la feuille i est malade, 0 sinon. Afin d'analyser les données, on a besoin de connaître le comportement de Y_i, c'est-à-dire la loi conditionnelle de Y_i sachant

[1] Le nombre de feuilles malades par mètre carré n'est pas une mesure adaptée s'il s'intéresse à la propagation à moins de 50 cm

[2] Pour la rouille jaune par exemple, parce qu'on peut difficilement distinguer les différentes lésions d'une même feuille, on utilise plutôt les mesures présence/absence de lésions ou surface des lésions que nombre de lésions

X_i et Z_i. Cette loi est déduite de la loi de N_i. En effet, $Y_i|X_i, Z_i$ suit une loi de Bernouilli de probabilité

$$\begin{aligned}\mathbb{P}(Y_i = 1|X_i, Z_i) &= \mathbb{P}[N_i \geq 1|X_i, Z_i] \\ &= 1 - \exp\{-Z_i S(X_i)\},\end{aligned}$$

i.e. la probabilité qu'il y ait au moins une lésion sur la feuille i.

Quelque soit la mesure de maladie utilisée, si elle s'exprime comme une fonction (déterministe ou aléatoire) du nombre de lésions sur une feuille, alors on peut en donner la loi. Cette loi peut être simple (Bernouilli dans l'exemple traité) mais peut aussi être compliquée et ne pas correspondre à une loi de probabilité usuelle.

Positions des feuilles censurées

Si la position de la feuille i n'est pas précisément connue mais qu'elle est censurée dans un sous-espace A_i du plan $\mathbb{R}^2$, alors on est intéressé non par la loi conditionnelle de N_i sachant X_i et Z_i mais par la loi conditionnelle de N_i sachant A_i et Z_i. Cette loi est un mélange de lois de Poisson. Pour déterminer ce mélange il faut donner la mesure de probabilité, disons λ, de X_i sachant que X_i est dans A_i. Il vient

$$\mathbb{P}(N_i = n|A_i, Z_i) = \int_{A_i} \mathbb{P}(N_i = n|X_i = x, Z_i)\lambda(dx).$$

Par exemple dans le chapitre 3 où les positions des feuilles sont censurées dans des rectangles, on suppose que la loi de $X_i|X_i \in A_i$ est uniforme sur le rectangle A_i.

Fragilités non observées

Dans les expérimentations que nous considérons, les fragilités ne sont pas des variables observées, que ce soit directement ou indirectement à l'aide de covariables explicatives. On est alors interessé par la loi conditionnelle de N_i sachant la position X_i seulement, les fragilités étant vues comme des variables aléatoires[3].

Traitons un cas particulier : supposons que les fragilités soient des variables aléatoires indépendantes. Alors la loi conditionnelle de N_i sachant X_i est un mélange de lois de Poisson. Pour déterminer ce mélange, il faut donner la loi des fragilités Z_i. Si l'on note μ la mesure de probabilité de Z_i sur son support $\mathcal{Z}$, il vient

$$\mathbb{P}(N_i = n|X_i) = \int_{\mathcal{Z}} \mathbb{P}(N_i = n|X_i, Z_i = z)\mu(dz).$$

Par exemple dans le chapitre 3 les fragilités sont supposées indépendantes et identiquement distribuées à densité dans $[0, 1]$.

On peut aussi faire face à des fragilités non observées mais structurées spatialement comme suggéré dans la section 3.7.2. Dans ce cas, on peut vouloir écrire la loi jointe de $N_1, \ldots, N_I$ sachant $X_1, \ldots, X_I$. Formellement elle s'écrit facilement, mais selon la structure

[3] C'est le cas dans les modèles de fragilités développés en analyse de survie (Nielsen et al., 1992).

spatiale, son écriture est plus ou moins exploitable. Pour manipuler des modèles incluant des variables aléatoires cachées et structurées spatialement, des techniques ont été développées. Les variables aléatoires cachées sont modélisées par un champ aléatoire caché pouvant être un champ de Markov (Green and Richardson, 2002; Hrafnkelsson and Cressie, 2003) ou encore un champ aléatoire à support continu (Diggle et al., 1998; Henderson et al., 2002; Hrafnkelsson and Cressie, 2003).

Anisotropies pour la direction et la distance de dispersion

Dans la formule (5.1) du potentiel infectieux, les fonctions f et g gouvernent les anisotropies pour, respectivement, la direction et la distance. Dans le chapitre 3, f et g sont constantes sur $[0, 2\pi[$ et donc la dispersion est isotrope. Dans le chapitre 4, f et g sont des fonctions paramétrées et unimodales dont les modes ne correspondent pas. D'autres cas peuvent être envisagés selon les conditions expérimentales : fonctions multimodales (s'il y a plusieurs directions privilégiées ou dans lesquelles les spores vont loin en moyenne), f ou g constante (anisotropie simple), un mode commun pour f et g. Toutefois, le choix pour f et g est contraint par le fait que f est une densité de probabilité, et que g est une fonction positive.

Combinaisons des modifications

Les paragraphes précédents présentent des modèles dérivés du modèle sous-jacent en ne modifiant qu'une caractéristique du modèle. Cependant, plusieurs modifications peuvent être effectuées afin d'adapter le modèle au contexte expérimental. Par exemple, dans le chapitre 3 les positions sont censurées, les fragilités ne sont pas observées, la dispersion est isotropique. Dans le chapitre 4 la mesures de la maladie est changée, les fragilités sont supposées égales, les fonctions d'anisotropie sont paramétrées et unimodales.

5.2 Résumer la connaissance en biologie

5.2.1 Eléments modélisés

Rappelons que les modèles proposés dans les deux précédents chapitres reflètent
- la décroissance de la quantité de spores ou de symptômes avec la distance à la source (chapitre 3 et 4),
- l'hétérogénéité des hôtes face à l'infection (chapitre 3), et
- la dispersion anisotrope des spores (chapitre 4).

Ces trois éléments ont été intégrés dans les modèles parce qu'ils jouent un rôle dans la propagation des rouilles (connaissance biologique) et parce qu'ils apparaissent de manière évidente dans les données expérimentales analysées. Par exemple, la figure 3.2 montre la sur-dispersion des données que l'on attribue à l'hétérogénéité des feuilles face à l'infection. En cela nos modèles sont des résumés, plutôt simples, de la connaissance en biologie et de l'observation.

La décroissance de la quantité moyenne de spores ou de symptômes avec la distance à la source est un élément largement documenté et intégré dans les modèles de propagation décrits au chapitre 3. Dans nos modèles, elle est introduite à travers le terme exponentiel de la fonction de dispersion (cf. section 5.1.1). Utiliser une décroissance exponentielle de la quantité de spores n'est pas original puisqu'un des modèles empiriques les plus utilisés, le gradient exponentiel, en fait l'hypothèse. L'originalité de nos modèles tient plutôt à l'intégration de composantes décrivant les éléments 'hétérogénéité des feuilles face à l'infection' et 'anisotropie de la dispersion des spores'. Intégrer ces composantes permet de mieux décrire la variabilité des données.

L'hétérogénéité des hôtes face à l'infection est prise en compte dans le modèle développé pour la propagation à courte distance en introduisant des fragilités individuelles caractérisant les feuilles. L'hétérogénéité des hôtes face à l'infection, bien que documentée (section 3.2.2), n'est pas intégrée dans les modèles mentionnés dans la section 2.2. Notre modèle, lui, permet la quantification de l'hétérogénéité des feuilles à travers la densité de probabilité des fragilités.

La dispersion anisotrope des spores est prise en compte dans le modèle développé pour la propagation à longue distance. Cette prise en compte est réalisée en faisant dépendre de la direction deux caractéristiques : la quantité de spores déposées d'une part, et la distance moyenne entre la source et les positions de dépôt des spores d'autre part. Ces caractéristiques sont supposées être des fonctions éventuellement différentes de la direction. La plupart des modèles mécanistes et quasi-mécanistes présentés dans la section 2.2.2 intègrent le caractère anisotrope de la dispersion des spores. Cependant, dans ces modèles, les anisotropies pour la direction et la distance sont contraintes l'une par l'autre. Dans notre modèle ce n'est pas le cas, et l'analyse statistique menée au chapitre 4 montre que les deux anisotropies peuvent effectivement être différentes.

5.2.2 Eléments modélisés et échelle

Rappelons que l'anisotropie de la dispersion n'est pas intégrée au modèle pour la propagation à courte distance[4] (chapitre 3), et que l'hétérogénéité des feuilles ne l'est pas au modèle pour la propagation à longue distance[5] (chapitre 4). Ceci est lié à un problème d'échelle. En effet, toutes les échelles n'offrent pas la même information sur le phénomène "propagation spatiale". On voit des éléments à certaines échelles qu'on ne voit pas à d'autres, par exemple l'hétérogénéité des feuilles et l'anisotropie de la dispersion. Dans l'autre sens, on voit des éléments qui sont communs à plusieurs échelles, c'est le cas de la décroissance de la quantité de spores avec la distance. Le cadre de modélisation décrit dans la section 5.1 permet d'utiliser des modèles différents mais ayant une base commune pour décrire des expériences

[4] Il n'y a pas d'évidence dans les données en faveur d'une dispersion anisotrope des spores.

[5] Une version du modèle intégrant des fragilités individuelles a été utilisée. Cependant, la matrice d'information de Fisher n'était pas inversible, autrement dit, le modèle n'était pas identifiable. En fait, les données étant agrégées (nombres de feuilles malades dans des quadrats d'un mètre carré), elles ne reflètent pas l'hétérogénéité des individus feuilles.

de propagation à différentes échelles. Ce cadre permet donc de voir comment les différents éléments de la propagation évoluent quand l'échelle change. Expliciter les différences entre échelles d'une part et les points communs entre échelles d'autre part permet une vision multi-échelle de la propagation spatiale.

Le concept d'échelle et l'évolution des éléments de la propagation lorsque l'échelle change est discuté plus avant au chapitre 10. Décrivons brièvement le contenu de ce chapitre. L'échelle est décomposée en trois dimensions (Dungan et al., 2002) : (i) l'échelle du phénomène étudié qui dépend des structures physiques et des processus agissant sur ces structures, (ii) l'échelle de l'échantillonnage qui dépend principalement du type de mesure utilisé pour quantifier la maladie, et (iii) l'échelle de l'analyse statistique qui dépend principalement de la variable modélisée. En général, les questions que l'on se pose sont liées à des structures physiques et aux processus y agissant dessus. Par exemple, les feuilles (la structure) ont-elles des propensions différentes à être infectées (le processus) ? Pour répondre à ce type de questions, on peut expérimenter et analyser les données issues de l'expérience. Pour ce faire, il faut choisir une échelle d'échantillonnage et une échelle d'analyse. Le chapitre 10 montre comment l'information apportée sur le phénomène étudié évolue lorsque les échelles d'échantillonnage et d'analyse changent.

5.3 Appréhender la variabilité des données qui reste inexpliquée par les modèles de propagation

En intégrant à nos modèles de propagation les fragilités (chapitre 3) et les fonctions d'anisotropie (chapitre 4), nous avons capturé une partie de la variabilité des données qui ne l'est pas lorsqu'un gradient de maladie isotrope est utilisé. Mais une part de la variabilité reste inexpliquée. Si l'on s'interroge sur les causes de cette variabilité inexpliquée, nos modèles de propagation peuvent être utilisés. En effet, pour schématiser,

- les modèles contiennent les éléments majeurs de la propagation et
- les écarts entre modèles et données contiennent les éléments non modélisés à l'origine de la part de variabilité qui reste inexpliquée.

Ainsi, analyser les écarts entre modèles et données peut permettre de détecter quels éléments non intégrés aux modèles jouent un rôle dans la variabilité inexpliquée des données, et de caractériser ces éléments.

5.3.1 Détection des éléments non modélisés

Les modèles que l'on peut dériver du cadre de modélisation sont bien-entendu incorrects. D'un point de vue plus positiviste, ce sont des approximations plus ou moins correctes de la "réalité", le niveau de correction étant jugé en fonction des questions posées. Par exemple, les modèles que nous avons proposés peuvent être jugés comme de bonnes approximations lorsqu'ils sont utilisés pour répondre aux questions suivantes : les feuilles de blé (homogènes génétiquement) sont-elles hétérogènes face à l'infection ? Quelle variabilité l'hétérogénéité

induit-elle dans les données ? Dans quelle direction les spores vont-elles préférentiellement ? Dans quelle direction les spores vont-elles le plus loin ? Ces deux directions sont-elles les mêmes ?

Après avoir répondu à ces questions grâce aux modèles proposés, de nouvelles questions peuvent être posées sur la part de la variabilité des données qui reste inexpliquée par les modèles. Cette part inexpliquée existe parce que des éléments jouant un rôle dans la propagation de la maladie n'ont pas été pris en compte[6]. Une deuxième étape peut alors être menée : détecter des éléments non modélisés.

Les modèles que nous avons proposés peuvent servir d'outils pour détecter les éléments non modélisés. En effet, ces modèles peuvent être vus comme des filtres appliqués aux données. Avec ces filtres, on isole dans les données d'une part ce qui relève des éléments connus et modélisés, et d'autre part ce qui relève des éléments non modélisés (connus ou non). Certains de ces éléments non modélisés peuvent être détectés par l'analyse des écarts entre modèles et données. Par exemple, une telle analyse a été conduite à la section 3.6. Cette analyse indique que les données de propagation à courte distance pourraient refléter (ou contenir de l'information sur) les éléments suivants

- la structuration spatiale de l'hétérogénéité des feuilles face à l'infection,
- le frottement entre feuilles comme mode additionnel de dispersion des spores,
- la propagation de la maladie depuis des sources de spores non introduites par l'expérimentateur.

Les données sur la propagation à longue distance pourraient, elles, refléter l'élément suivant qui n'a pas été modélisé

- la dispersion de nuages de spores.

Explicitons cet élément. Dans les modèles des chapitres 3 et 4, il est explicitement (section 3.3.1) ou implicitement supposé que les positions de dépôt des spores sont indépendantes et identiquement distribuées selon la fonction de dispersion. Cette hypothèse permet de définir le potentiel infectieux comme le produit entre la fonction de dispersion et la force de la source qui est un indicateur de la quantité de spores libérées. Cependant, on peut faire l'hypothèse que des spores libérées en même temps et par la même source sont transportées dans un même volume d'air et se déposent plus ou moins au même endroit. C'est ce que nous appelons la dispersion de nuages de spores. Sous cette hypothèse, les positions de dépôt de spores ne peuvent plus être indépendantes. A partir de là, deux cas se présentent.

- Dans le premier cas, le nombre de libérations de nuages de spores est supposé grand. Alors l'effet nuage n'apparaît que faiblement dans la répartition des positions de dépôts, et le nombre de spores déposées en tout point du plan ne s'écarte pas trop du potentiel infectieux défini comme le produit entre la fonction de dispersion et la force de la source.

[6] Il faut avoir à l'esprit que la présence de tels éléments a pu fausser l'inférence faite sur les éléments intégrés aux modèles : les estimateurs des paramètres peuvent être biaisés et les variances d'estimation sous- ou sur-estimées.

– Dans le deuxième cas, le nombre de nuages libérés est faible. Alors l'effet nuage est mis en évidence par des zones où l'intensité de la maladie est anormalement élevée, et le nombre de spores déposées en un point donné du plan peut s'écarter fortement du potentiel infectieux en ce même point.

Illustrons la dispersion de nuages de spores en analysant succinctement les données de propagation à longue distance. Notons $N(x)$ la quantité totale de feuilles et $Y(x)$ la quantité de feuilles malades dans un carré d'un mètre carré centré en x. Sous le modèle développé dans la section 4.2.2

$$\mathbb{E}\{Y(x)|x, N\} = N(x)p(x)$$

car $Y(x)$ suit une loi binomiale de taille $N(x)$ et de probabilité $p(x)$ qui est la probabilité qu'une feuille située en x soit malade. Ne conditionnons plus que par la position x :

$$\mathbb{E}\{Y(x)|x\} = \mathbb{E}\{N(x)|x\}p(x).$$

On dispose des données $y(x_1), \ldots, y(x_I)$ et $n(x_1), \ldots, n(x_I)$. Notons $\hat{y}(\cdot)$ et $\hat{n}(\cdot)$ les estimateurs nonparamétriques à noyau de $x \mapsto \mathbb{E}\{Y(x)|x\}$ et $x \mapsto \mathbb{E}\{N(x)|x\}$ caculés à partir de ces données. La fonction $\hat{p}(\cdot)$ définie par

$$\hat{p}(x) = \frac{\hat{y}(x)}{\hat{n}(x)}, \quad x \in \mathbb{R}^2,$$

est vue comme un estimateur de la fonction $p(\cdot)$. La figure 5.1 montre la carte de l'estimateur $\hat{p}(\cdot)$ calculé à partir des données de Thiverval-Grignon. Dans cette carte, nous nous intéressons à ce qui se passe à gauche de la source à la position (-150,-50). Cette zone correspond à une intensité de maladie anormalement élevée, le terme 'anormalement' traduisant donc un écart entre ce que l'on observe et ce que l'on attendait. Une des hypothèses envisageables pour la présence de cette zone à intensité élevée est la dispersion de nuages de spores.

5.3.2 Vers la caractérisation les éléments détectés

D'après la section précédente, des éléments non intégrés aux modèles sont suspectés jouer un rôle dans les données. En plus des éléments évoqués précédemment, d'autres éléments peuvent être envisagés dans d'autres contextes (l'hétérogénéité génétique des plantes face à l'infection, le rôle des facteurs climatiques (vent, pluie et humidité) dans la dispersion des spores). Afin de mieux comprendre la propagation spatiale des maladies aériennes, il est important de prouver l'existence de ces éléments et, le cas échéant, de les caractériser. Prouver l'existence de ces éléments ne peut se faire qu'au travers de nouvelles expériences biologiques. La conception et la réalisation de telles expériences est hors du propos de ce document. Cependant, sous l'hypothèse que les éléments suspectés existent, nous pouvons essayer de les caractériser à partir des données à notre disposition. Cette caractérisation, même si elle ne prouvera pas l'existence des éléments suspectés, pourra être utilisée pour concevoir les nouvelles expériences biologiques. Par exemple, savoir que la structuration

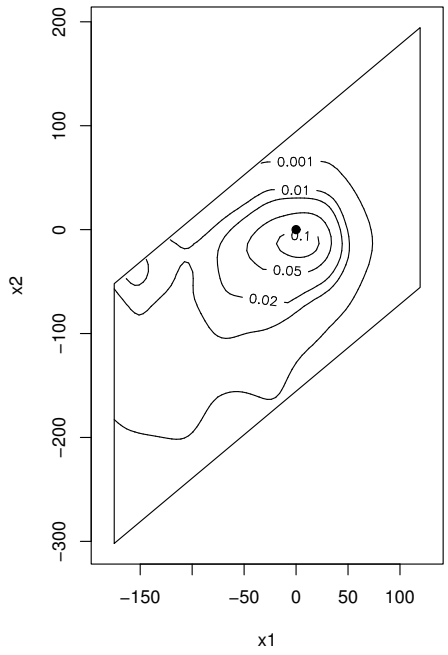


Fig. 5.1. Estimation non-paramétrique de la fonction de probabilité $p(\cdot)$ pour Thiverval-Grignon ($p(x)$ est la probabilité qu'une feuille située en x soit malade). Point : position de la source de spores ; unité des axes : mètre.

de l'hétérogénéité des feuilles peut être effective à l'échelle du décimètre doit guider la conception d'une expérience sur cet élément.

Les développements méthodologiques proposés dans la partie suivante de ce document ont justement pour objet la caractérisation d'éléments non modélisés mais dont on pressent le rôle dans le phénomène étudié.

Deuxième partie

Spécifier le second niveau d'un modèle hiérarchique en analysant des résidus

6

Introduction aux modèles hiérarchiques et à l'analyse de résidus

L'existence d'éléments biologiques non observés susceptibles de transparaître dans les données de propagation des rouilles a été évoquée dans la conclusion de la partie précédente (section 5.3). Les éléments en question sont notamment la structuration spatiale des fragilités et la dispersion de nuages de spores. Pour inférer sur ces éléments, des modèles plus complexes que ceux de la première partie peuvent être développés et ajustés aux données. Les modèles que nous envisageons sont des modèles hiérarchiques à deux niveaux

- 1^{er} niveau : les mesures de la maladie sont distribuées selon une loi de probabilité conditionnée par la réalisation d'un processus sous-jacent, processus qui correspondrait par exemple à la structuration spatiales des fragilités ou à la dispersion de nuages de spores,
- 2^{nd} niveau : le processus sous-jacent est distribué selon une loi de probabilité destinée à modéliser sa structure propre.

Les modèles hiérarchiques permettent de refléter, selon Green et al. (2003, p. 3), "*the presence of additional structure over the standard 'flat' statistical models*". Les *standard 'flat' statistical models* désignent les modèles hiérarchiques pour lesquels le second niveau est remplacé par un simple paramètre. Ils sont appelés, dans notre terminologie, 'modèles de base'.

Une des difficultés liée aux modèles hiérarchiques est la spécification, ou détermination, de la loi du processus sous-jacent car ce dernier n'est pas directement observé (on parle de processus caché). Cette difficulté est d'autant plus sérieuse que le processus sous-jacent est peu connu, comme c'est le cas pour la structuration des fragilités et la dispersion de nuages de spores. L'objet de cette deuxième partie est la spécification du second niveau d'un modèle hiérarchique. Pour ce faire, nous exploitons un processus résiduel obtenu en estimant le modèle de base associé au modèle hiérarchique considéré, et en mesurant l'écart entre modèle estimé et données. Cette approche est motivée par le fait que les résidus sont des fonctions de l'ignorance (i.e. de ce qui n'a pas été intégré au modèle) et, sous le modèle de base, le second niveau fait justement partie de l'ignorance. Dans le présent chapitre nous présentons les modèles hiérarchiques, discutons de l'information contenue dans les processus résiduels puis nous expliquons comment nous allons utiliser des "résidus" pour spécifier le second niveau d'un modèle hiérarchique.

6.1 Modèles hiérarchiques

6.1.1 Une définition

"*The word 'hierarchical' in 'Bayesian hierarchical models' refers to a generic model building strategy in which latent (that is, unobserved) quantities are organized into (a small number of) discrete levels with logically distinct and scientifically interpretable functions, with probabilistic relationships between them that capture inherent features of the data.*"

Cette définition du terme 'hiérarchique' proposée par Green et al. (2003, p. 2) correspond au modèle hiérarchique que nous avons défini dans l'introduction de ce chapitre et aux modèles hiérarchiques qui vont être décrits dans cette partie. Seule la référence au contexte bayésien ne nous convient pas. En effet, la statistique bayésienne vise à utiliser de manière explicite l'information disponible, i.e. connue a priori, sur le phénomène étudié. Ainsi, si l'on dispose d'information pouvant être mise sous la forme d'une distribution, dite a priori, pour une composante du modèle, alors, en combinant la distribution a priori et les données, une loi a posteriori pour cette composante peut être déterminée (Robert, 1992). Le contexte qui nous intéresse est tout à fait différent : dans notre cas une composante du modèle (le processus sous-jacent) est mal connue, et nous voulons la spécifier en se basant seulement sur les données. C'est pourquoi les modèles hiérarchiques auxquels nous nous intéressons sont hiérarchiques, mais non bayésiens.

6.1.2 Construction et exemples

Les modèles hiérarchiques sont variés (modèles de mélanges, GLMMs[1], GLMMs spatiaux, modèles de fragilités individuelles, partagées ou corrélées, modèles avec champ de markov caché), et sont appliqués à des domaines divers (géopolitique, épidémiologie humaine ou végétale, archéologie, environnement). Dans la suite nous décrivons certains types de modèles hiérarchiques et en donnons quelques exemples. Une attention particulière est portée aux modèles à effets partagés et à effets dépendants non partagés puisqu'ils sont étudiés aux chapitres 7 et 8.

Premier niveau d'un modèle hiérarchique

Soit $Y_1, \ldots, Y_N$ les variables réponses de N individus. Les variables $Y_1, \ldots, Y_N$ sont modélisées au premier niveau d'un modèle hiérarchique : conditionnellement à des effets aléatoires individuels $\alpha_1, \ldots, \alpha_N$, à un paramètre β et à des covariables $X_1, \ldots, X_N$, $Y_1, \ldots, Y_N$ sont indépendantes et identiquement distribuées selon une loi $\mathcal{L}$

$$Y_n|\alpha_n, \beta, X_n \ \sim \ \text{indép. } \mathcal{L}(\alpha_n, \beta, X_n).$$

[1] Generalized Linear Mixed Model.

Second niveau d'un modèle hiérarchique

Les effets aléatoires individuels $\alpha_1, \ldots, \alpha_N$, qui forment le processus sous-jacent, sont modélisés au second niveau du modèle hiérarchique. Différentes relations de dépendance entre les effets aléatoires peuvent être envisagées. Les paragraphes suivants décrivent des relations possibles.

Effets égaux

Si les effets individuels sont tous égaux, $\alpha_1 = \cdots = \alpha_N = \alpha$, alors α est vu comme un paramètre (aléatoire ou pas). Nous appelons 'modèle de base' le modèle pour $Y_1, \ldots, Y_N$ qui résulte de cette hypothèse.

Effets indépendants et identiquement distribués

Intégrer des effets aléatoires indépendants et identiquement distribués permet de modéliser une hétérogénéité non observée entre individus, hétérogénéité qu'aucune covariable observée ne structure. L'hétérogénéité non observée entre individus advient lorsque des covariables déterminantes pour le phénomène étudié ne sont pas introduites dans la modélisation. L'hétérogénéité non observée a pour conséquence la sur-dispersion des données. Un modèle hiérarchique pour lequel, au second niveau, les effets aléatoires individuels sont indépendants et identiquement distribués sert prioritairement à définir une loi de probabilité sur-dispersée par rapport à la loi obtenue avec le modèle de base. Donnons quelques exemples.

En accidentologie, la distribution binomiale négative est utilisée pour modéliser le nombre d'accidents pour un sujet donné (McCullagh and Nelder, 1989). Cette distribution peut être construite à partir d'un modèle hiérarchique. Au premier niveau de ce modèle, le nombre d'accidents du sujet n suit une loi de Poisson de moyenne α_n. Au second niveau, les effets individuels $\alpha_1, \ldots, \alpha_N$ sont indépendants et identiquement distribués selon une loi gamma commune, et modélisent une hétérogénéité non observée entre sujets.

En épidémiologie végétale, le modèle de fragilité développé au chapitre 3 est un modèle hiérarchique où les fragilités sont indépendantes et identiquement distribuées à densité polynomiale sur $[0, 1]$. Les fragilités reflètent l'hétérogénéité des feuilles face à l'infection.

En géopolitique, un modèle de fragilité est utilisé afin de modéliser les durées entre conflits consécutifs pour des paires de pays (Box-Steffensmeier and Jones, 2004). Au premier niveau de ce modèle hiérarchique, la distribution de la durée entre conflits consécutifs dépend d'une fonction de hasard. Pour chaque paire de pays, la fonction de hasard est le produit entre (i) une fonction de covariables liées à la stabilité des deux pays et à l'intensité de leurs échanges, et (ii) une fragilité qui reflète la propension non observée de la paire de pays à être en conflit. Au second niveau les fragilités (ou effets aléatoires) sont supposées

indépendantes[2] et identiquement distribuées selon une loi gamma. Les fragilités modélisent l'hétérogénéité non observée entre paires de pays.

En analyse d'influence, un modèle perturbé peut être utilisé pour détecter les données influentes (Critchley and Marriott, 2004). Le modèle perturbé est un modèle hiérarchique. Au premier niveau, la variable d'intérêt est distribuée selon une loi qui ne varie d'un individu à l'autre que par une perturbation ; au second niveau, les perturbations (ou effets aléatoires) sont supposées indépendantes et identiquement distribuées.

Effets partagés

Le second niveau d'un modèle hiérarchique peut être structuré par une covariable, dite structurante, qui permet de former des groupes (ou clusters) d'individus : les individus pour lesquels la covariable structurante est identique partagent le même effet aléatoire. La construction de tels modèles hiérarchiques est plus claire en ré-indiçant les variables de la manière suivante. Les Y_{ij}, $i = 1, \ldots, I$ et $j = 1, \ldots, J_i$, sont les variables modélisées au premier niveau du modèle hiérarchique. Conditionnellement à des effets aléatoires $\alpha_1, \ldots, \alpha_I$, à un paramètre β et à des covariables X_{ij}, les variables Y_{ij} sont indépendantes et identiquement distribuées selon une loi $\mathcal{L}$

$$Y_{ij}|\alpha_i, \beta, X_{ij} \ \sim \ \text{indép. } \mathcal{L}(\alpha_i, \beta, X_{ij}).$$

On parle de données groupées ou de données en clusters.

Prenons l'exemple d'une variable mesurée chez des patients soignés dans un certains nombre de centres médicaux. La variable d'intérêt peut être la réussite d'un traitement, la durée avant l'apparition d'une maladie, le temps d'attente avant la prise en charge par le personnel médical, ou encore le coût du traitement. Pour toutes ces variables, un effet 'centre médical' peut contribuer à l'hétérogénéité des données. Dans un tel contexte, des modèles linéaires mixtes généralisés (GLMM) (McCulloch and Searle, 2001) ou des modèles de fragilités partagées (Nielsen et al., 1992) peuvent être utilisés. Dans ces modèles, un effet aléatoire par centre médical est introduit, et tous les patients soignés dans un même centre partagent le même effet.

Les effets aléatoires $\alpha_1, \ldots, \alpha_I$ peuvent être supposés indépendants et identiquement distribués. Ce sera ainsi dans l'exemple précédent si aucune covariable observée ne permet de dire que certains centres médicaux ont tendance à avoir le même comportement. Mais les effets aléatoires peuvent aussi être supposés dépendants si la covariable structurante est une variable ordonnée. Donnons un exemple où l'espace induit une relation de dépendance entre des effets aléatoires partagés. Dans une étude des durées de vie pour des patients atteints par une leucémie, Henderson et al. (2002) construisent un modèle qui intégre des effets districts (ou cantons) : tous les patients d'un même district partagent le même effet district. Les

[2] Notons que l'hypothèse faite par Box-Steffensmeier and Jones (2004) sur l'indépendance des fragilités caractérisant les paires est forte : en effet, on aurait tendance à penser que deux paires qui ont en commun un pays ont des fragilités proches (un pays belliqueux à tendance à l'être avec tous les pays, un pays pacifiste à tendance à l'être avec tous les pays).

districts étant répartis dans l'espace, les auteurs font l'hypothèse que les effets districts sont dépendants et que la dépendance décroît avec la distance entre districts. L'analyse statistique montre la significativité des effets districts et de leur dépendance spatiale. Le temps peut aussi être une covariable structurante induisant de la dépendance entre effets aléatoires. Par exemple, tous les jours du mois précédant une élection, 1000 individus sont tirés au hasard et sondés sur leur intention de vote. Des effets aléatoires jours peuvent être introduits pour refléter les variations des intentions de votes dues aux interventions (non observées) des candidats. Si les effets des interventions des candidats sur les électeurs durent plusieurs jours, alors les effets jours forment un processus temporel corrélé.

Effets dépendants mais non partagés

Dans le précédent paragraphe, une covariable structurante permet de définir des groupes d'individus, les individus d'un même groupe étant soumis à un même effet 'groupe'. Dans ce paragraphe, la relation de dépendance est changée : la covariable structurante définit une distance entre individus telle que l'évènement 'distance nulle entre deux individus', qui se traduirait par un effet aléatoire partagé par les deux individus, est un évènement improbable. Ainsi, la covariable structurante ne permet pas de définir des groupes homogènes d'individus. Pour que la covariable structurante définisse une distance, il est nécessaire qu'elle soit ordonnée.

Construisons un modèle hiérarchique où l'espace joue le rôle de covariable structurante ordonnée telle qu'envisagée précédemment. Soit $Z_1, \ldots, Z_N$ les positions dans l'espace, noté $\mathcal{Z}$, de N individus. Les effets aléatoires individuels sont fonctions de $Z_1, \ldots, Z_N$; ils sont notés $\alpha(Z_1), \ldots, \alpha(Z_N)$. Au premier niveau du modèle hiérarchique, conditionnellement aux effets aléatoires $\alpha(Z_1), \ldots, \alpha(Z_N)$, à un paramètre β et à des covariables $X_1, \ldots, X_N$, les variables $Y_1, \ldots, Y_N$ sont indépendantes et identiquement distribuées selon une loi $\mathcal{L}$

$$Y_n|\alpha(Z_n), \beta, X_n \ \sim \ \text{indép. } \mathcal{L}\{\alpha(Z_n), \beta, X_n\}.$$

Au second niveau, les effets aléatoires individuels $\alpha(Z_1), \ldots, \alpha(Z_N)$ sont générés par un champ aléatoire. Ce champ est dit 'champ aléatoire caché' puisqu'il est introduit au second niveau du modèle hiérarchique et donc est non observé.

Présentons trois exemples de modèles hiérarchiques intégrant des effets dépendants spatialement. Premier exemple : dans une étude archéologique, pour déterminer les endroits d'une zone géographique où il y a eu une activité ancienne, la zone a été quadrillée par une grille carrée, et la concentration de phosphate en chaque nœud (ou site) de la grille a été mesurée (car une forte concentration de phosphate est un signe tangible d'une activité ancienne). Un modèle hiérarchique a été proposé par Besag et al. (1991). Au premier niveau, la concentration en phosphate Y_n pour le site Z_n suit une loi normale dont la moyenne dépend de la présence/absence d'activité ancienne en Z_n. Au second niveau, les présences/absences d'activité, notées $\alpha(Z_1), \ldots, \alpha(Z_N)$, sont générées par un champ aléatoire de Markov binaire. Cette hypothèse signifie qu'un site dont les sites voisins sont majoritairement anciennement

actifs, a tendance à être lui-même anciennement actif, et vice-versa. Dans cet exemple, l'espace est une grille et la relation de dépendance est à distance finie.
Deuxième exemple : pour déterminer l'intensité de la radioactivité sur Rongelap Island, Diggle et al. (1998) proposent un modèle hiérarchique modélisant les mesures de concentration de césium. Au premier niveau du modèle hiérarchique, la mesure de la concentration de césium Y_n faite en la position Z_n suit une loi de Poisson dont la moyenne $\alpha(Z_n)$ dépend de la position de la mesure. $\alpha(Z_n)$ représente l'intensité de la radioactivité en Z_n. Au second niveau, les intensités de radioactivité $\alpha(Z_1), \ldots, \alpha(Z_N)$ sont générées par un champ gaussien stationnaire avec une fonction de covariance exponentielle-puissance. Dans cet exemple, l'espace est continu et la relation de dépendance décroît de manière continue avec la distance. Ce type de modèles a été appellé GLMM spatial par Zhang (2002).
Troisième exemple : pour analyser les données de durées de vie de patients atteints par une leucémie, Henderson et al. (2002) proposent, outre le modèle évoqué plus haut, un modèle de fragilités spatialement dépendantes. Au second niveau de ce modèle hiérarchique, des fragilités individuelles caractérisant les patients sont générées par un champ aléatoire gamma dépendant d'une fonction de covariance de Matérn. Dans cet exemple aussi, l'espace est continu et la relation de dépendance décroît de manière continue avec la distance.

Dans ces trois exemples, l'hétérogénéité non observée entre individus, modélisée par des effets aléatoires binaires, gaussiens ou gammas, est d'autant moins grande que les individus sont proches spatialement. Cependant, dans aucun de ces exemples l'espace ne structure les individus en groupes. Les effets aléatoires sont donc dépendants mais non partagés.

6.1.3 Quelle spécification pour le second niveau ?

Une difficulté liée aux modèles hiérarchiques concerne le choix du second niveau, i.e. le choix de la distribution des effets aléatoires. En effet, dans les modèles hiérarchiques présentés ci-dessus, pourquoi les effets aléatoires sont-ils générés par des lois gamma ou normales indépendantes, des champs de Markov binaires, des champs gamma ou log-normaux ? Pourquoi choisir telle ou telle structure de dépendance ? Box-Steffensmeier and Jones (2004, p. 164), en discutant des modèles de fragilités, écrivent "*neither theory nor data typically provides much guidance for imposing a specific distribution on the frailties*". La remarque est plus généralement vraie pour les modèles hiérarchiques. En effet, la connaissance théorique du phénomène étudié permet rarement de spécifier la distribution des effets aléatoires puisque ceux-ci sont souvent utilisés pour refléter des mécanismes complexes et mal connus (Henderson et al., 2002). Les données, en l'état permettent difficilement de spécifier la distribution des effets aléatoires parce qu'elles sont des versions bruitées des valeurs des effets aléatoires.

Dans la pratique, le choix d'une spécification pour le second niveau d'un modèle hiérarchique est le plus souvent fait à cause de considérations liées à la commodité mathématique ou à l'usage. Par exemple dans les modèles de fragilités, il est commode, mathématiquement parlant, d'utiliser des fragilités de loi gamma car cela permet de calcu-

ler l'expression analytique de la vraisemblance des données observées (Clayton, 1991). Mais avec le développement de l'inférence via la technique MCMC[3], le choix de la loi gamma ne s'impose plus (Sargent, 1998). Pourtant, Henderson et al. (2002) préfèrent la loi gamma à la loi log-normale en particulier parce que "*log-normal frailty is rarely used in survival analysis*". Ainsi, dans le passé la loi gamma pour les fragilités était choisie par commodité mathématique, plus tard c'est l'usage qui a poussé Henderson et al. (2002) à faire ce même choix.

La spécification du second niveau est critique parce qu'en cas de mauvaise spécification les conclusions de l'analyse peuvent être fausses. En effet, les estimateurs des paramètres, les variances des estimateurs, les estimateurs des effets aléatoires peuvent être biaisés. Ceci est détaillé dans les introductions des chapitres 7 et 8 qui mentionnent les travaux de Neuhaus et al. (1992) et Monestiez et al. (2005).

L'inférence pouvant être faussée en cas de mauvaise spécification du second niveau d'un modèle hiérarchique, des méthodes permettant de spécifier ce second niveau sont nécessaires. Conscients de cet enjeu, des auteurs tels que Sastry (1997), Henderson et al. (2002) ou encore Christensen (2004) (i) proposent différentes spécifications pour le second niveau, (ii) construisent les modèles hiérarchiques correspondants, (iii) ajustent aux données chacun des modèles, et (iv) sélectionnent le modèle hiérarchique, i.e. la spécification du second niveau, qui réalise le "meilleur" ajustement aux données observées. Le problème de cette approche est qu'elle est "aveugle" puisque les différentes spécifications ne sont pas proposées au vu des valeurs des effets aléatoires. Donnons un exemple simple qui illustre cette idée. Considérons des données dont on veut spécifier la loi. Traçons l'histogramme de ces données. Si l'histogramme ne présente qu'un mode, qu'il est à peu près symétrique, que ses queues ne sont pas trop lourdes, on va essayer d'ajuster une loi normale aux données. Si l'histogramme présente deux modes bien distincts, alors rien ne sert d'ajuster une loi normale et on va plutôt s'orienter, par exemple, vers un mélange de deux normales. Le problème avec la sélection d'une distribution pour des effets aléatoires est que ceux-ci ne sont pas observés, que leur histogramme par exemple ne peut donc pas être tracé, et que finallement le choix des distributions est fait en aveugle. Ainsi, l'efficacité de la stratégie précédente de spécification du second niveau va reposer sur la taille de la classe de modèles que l'on est prêt à explorer.

La question suivante est donc posée :

Question 6.1. Peut-on développer une méthode 'non aveugle' permettant de spécifier le second niveau d'un modèle hiérarchique, c'est-à-dire la distribution de ses effets aléatoires ?

[3] Markov Chain Monte Carlo.

6.2 Processus résiduel

6.2.1 Que contient un processus résiduel ?

Un processus résiduel mesure l'écart entre des données et un modèle ajusté aux mêmes ou à d'autres données. Mais pourquoi y-a-t-il un écart entre les données et le modèle, et que représente cet écart ?

Les données sont vues comme les réalisations de variables, variables dont le comportement est sensé être mimé par le modèle. En général, le modèle ne mime qu'approximativement le comportement des variables car son concepteur a, volontairement ou non, ignoré et/ou simplifié des mécanismes jouant un rôle dans le phénomène étudié (dans la suite, la simplification des mécanismes est vue comme une forme d'ignorance). Ainsi, l'ignorance du concepteur induit un écart entre les données et le modèle, et donc le processus résiduel est une fonction de l'ignorance, c'est-à-dire une fonction de ce qui n'a pas été integré au modèle.

L'écart entre données et modèle peut également être fonction d'un biais d'estimation fait parce qu'un modèle (incorrect) est ajusté aux données. En effet, la méthode d'estimation peut essayer de compenser l'écart, dû à l'ignorance, entre données et modèle. Cela se traduit par des estimateurs biaisés (Henderson and Oman, 1999). Le processus résiduel, en plus d'être fonction de l'ignorance, est donc aussi fonction d'un biais d'estimation dû à l'ignorance.

L'écart entre données et modèle est également dû à l'échantillonage car les estimations des paramètres varient d'un échantillon de données à l'autre. Par conséquent, le processus résiduel est fonction de l'échantillonage.

La relation fonctionnelle entre le processus résiduel d'une part, et l'ignorance, le biais d'estimation dû à l'ignorance et l'échantillonage d'autre part, suscite l'idée qu'une analyse du processus résiduel peut permettre de réduire la part d'ignorance, i.e. la part de ce qui n'est pas pris en compte dans le modèle. Les exemples suivants montrent comment une telle analyse peut être conduite.

6.2.2 Exemples d'analyse d'un processus résiduel

Régression linéaire : résidus, hétéroscédasticité et non-linéarité (un cadre standard)

En régression linéaire, le résidu ordinaire est défini comme la différence entre la valeur observée et la valeur prédite de la variable modélisée. "Lorsque les hypothèses associées au modèle (de régression linéaire) sont correctes, les résidus (ordinaires) et les valeurs prédites sont non corrélées" (Antoniadis et al., 1992, p. 34). Par conséquent, le graphe où sont reportés en abscisse les valeurs prédites et en ordonnées les résidus ordinaires devrait exhiber un nuage de points formant une bande horizontale de largeur constante (Cook and Weisberg, 1982; Antoniadis et al., 1992) ; c'est le cas du graphe (a) de la figure 6.1. La figure 6.1 montre également deux cas pathologiques : le graphe (b) indique qu'une hétéroscédasticité (variance non constante du bruit) n'a pas été prise en compte, les graphes (c) et (d) indiquent qu'une

non linéarité n'a pas été prise en compte. En conclusion, ce qui a été ignoré par le modèle (hétéroscédasticité ou non linéarité) est contenu dans le processus résiduel.

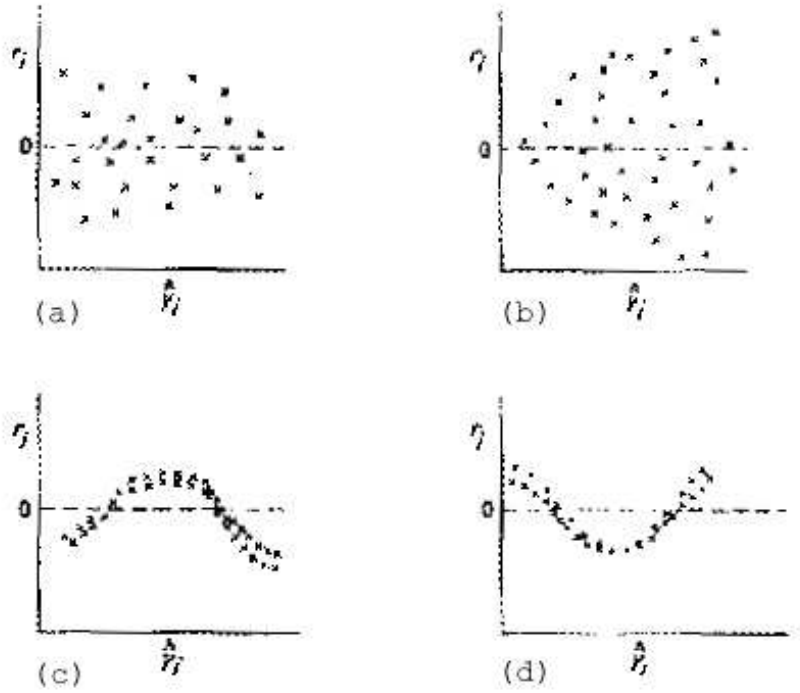


Fig. 6.1. Diverses formes du tracé des résidus ordinaires dans le cas de la régression linéaire. (a) ajustement satisfaisant, (b) présence d'hétéroscédasticité, (c) et (d) présence de non linéarité. Source : Cook and Weisberg (1982, p. 37).

Modèle hiérarchique de Hodges : résidus et covariable ignorée

Considérons le modèle hiérarchique à effets partagés suivant :

$$\begin{aligned} Y_{ij}|\alpha_i, \sigma^2, X_{ij}^{(1)} &\sim \text{indép. } \mathcal{N}(X_{ij}^{(1)}\alpha_i, \sigma^2) \\ \alpha_i|\mu, \Sigma, X_i^{(2)} &\sim \mathcal{N}(X_i^{(2)}\mu, \Sigma), \end{aligned}$$

où les Y_{ij} sont les variables modélisées, les $X_{ij}^{(1)}$ et les $X_i^{(2)}$ sont des covariables, les α_i sont des effets aléatoires partagés et σ^2, μ et Σ sont des paramètres inconnus. Aux deux niveaux du modèle, les covariables interviennent de manière linéaire et le bruit est gaussien. Dans ce cadre, Hodges (1998) développe des outils de diagnostic basés sur des processus résiduels. L'auteur construit par exemple des résidus qui sont fonctions d'une covariable ignorée dans la modélisation et candidate à être intégrée. Ce processus résiduel contient donc de l'information sur ce qui a été ignoré, en l'occurrence la covariable candidate.

Ecart à un processus ponctuel spatial modélisant l'expansion du pin noir

L'expansion, sur une génération, du pin noir sur le Causse Méjean a été étudiée par Debain (2003). Dans cette étude, les positions des pins parents et des pins enfants ont été relevées. Elles sont reportées à la figure 6.2(a). La figure 6.2(c) montre l'intensité des enfants estimée par lissage à partir des positions observées. Cette intensité est appelée 'intensité observée'. Afin d'évaluer la dispersion des graines depuis les parents, un modèle a été élaboré. Dans ce modèle, chaque parent émet, indépendemment des autres parents, un nombre poissonien de graines (dépendant de l'âge de l'arbre) dispersées indépendemment selon une fonction de dispersion isotrope. Les enfants sont donc réparties selon un processus ponctuel de Poisson non homogène. Ce modèle a été ajusté aux données. La figure 6.2(b) montre une simulation du modèle ajusté sachant les positions (observées) des parents. La figure 6.2(d) montre l'intensité des enfants estimée par lissage à partir des positions simulées de la figure 6.2(b). Cette intensité est appelées 'intensité simulée'. 100 simulations ont été réalisées, 100 cartes d'intensité simulée ont été obtenues. La figure 6.2(e) montre en chaque point de l'espace le pourcentage de fois où l'intensité simulée est plus grande que l'intensité observée, c'est-à-dire la p-valeur locale de l'intensité observée. Ce pourcentage est une mesure de l'écart entre données et modèle ajusté, il défini donc un processus résiduel.

Etudions visuellement la structure spatiale du processus résiduel (figure 6.2(e)). L'intensité des pins enfants est sous-estimée par le modèle dans les bandes verticales $[0; 200]$ et $[650; 900]$ (zones blanches). Dans le modèle, aucune information topographique ou climatique n'est introduite : la fonction de dispersion est isotrope et est la même pour tous les parents. Or les bandes verticales $[0; 200]$ et $[650; 900]$ correspondent à des couloirs dans lesquels le vent s'engouffre. Ainsi, plus de graines qu'attendues pourraient être déposées dans ces couloirs. Ceci expliquerait la sous-estimation de l'intensité des enfants dans ces zones et, par compensation, la sur-estimation ailleurs. Il semble donc que le processus résiduel contienne de l'information sur ce qui a été ignoré, i.e. sur le fait que le vent s'engouffre dans les couloirs transportant avec lui un surplus de graines.

Ecart au modèle de fragilité décrivant la propagation de la rouille brune

Dans la section 3.6.3 nous avons défini un processus résiduel qui est légèrement structuré spatialement alors qu'il ne devrait pas l'être sous le modèle utilisé. Nous avons conclu à la section 3.7.2 que le processus résiduel pourrait contenir de l'information sur une possible structuration spatiale des fragilités (cette structuration reflétant une structuration spatiale de l'hétérogénéité des feuilles en terme de propension à être infectées).

Erreurs du modèle déterministe CHIMERE

Le modèle CHIMERE[4] sert à la prédiction et à la simulation de la qualité de l'air à plusieurs échelles. Il modélise la concentration de divers polluants parmi lesquels l'ozone.

[4] Chemistry-transport model, Copyright (C) 2004 Institut Pierre-Simon Laplace, INERIS, LISA, C.N.R.S. (http ://euler.lmd.polytechnique.fr/chimere/).

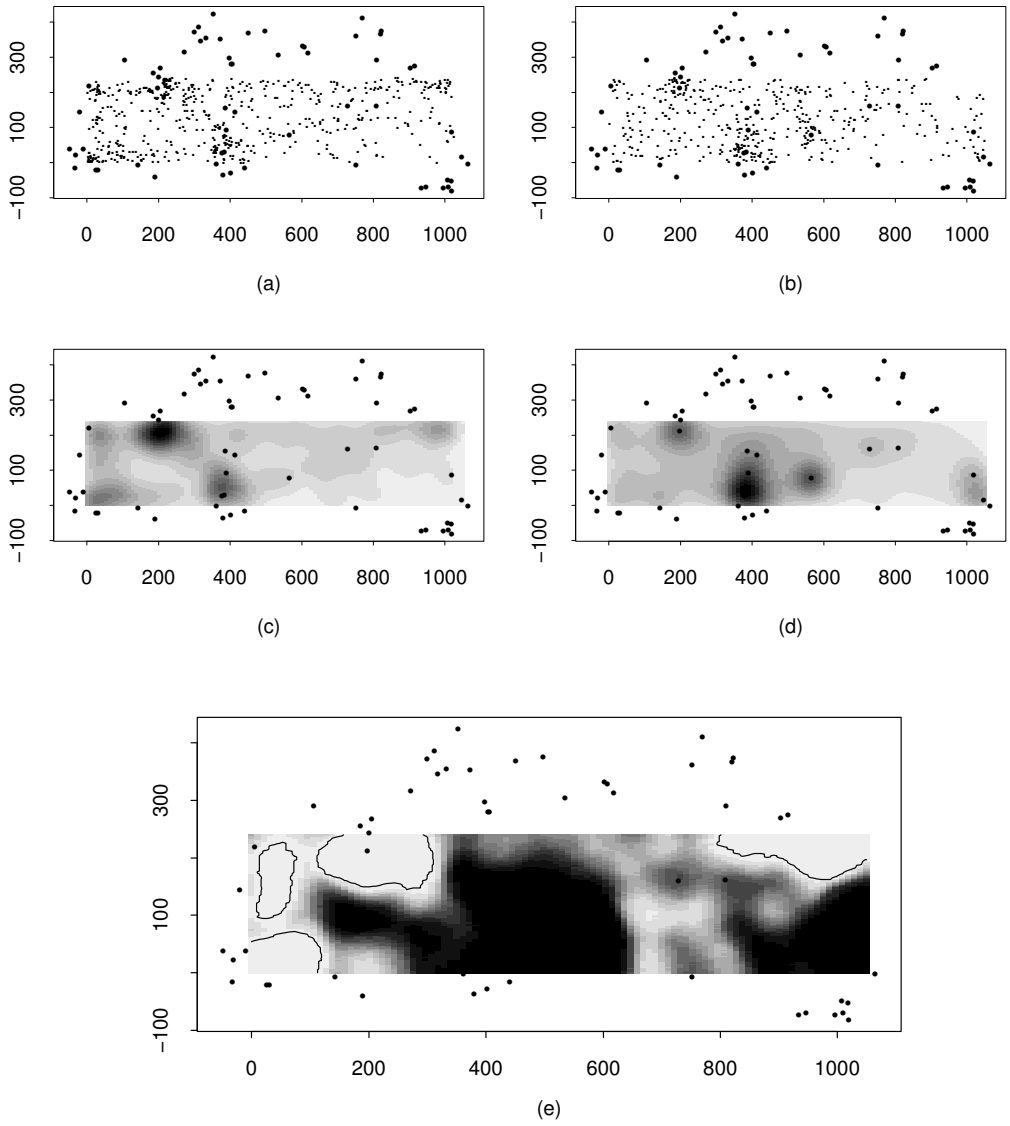


Fig. 6.2. Dispersion du pin noir sur le Causse Méjean. Sur tous les graphes, gros points : positions observées des parents. (a) positions observées des enfants (petits points), (b) positions simulées des enfants (petits points), (c) intensité observée des enfants obtenue par lissage de (a), (d) intensité simulée des enfants obtenue par lissage de (b), (e) pourcentage de fois où l'intensité simulée est plus grande que l'intensité observée. Palette de gris : noir pour les valeurs élevées, blanc pour les valeurs basses. Unité : mètre. Source : Joël Chadœuf.

C'est un modèle déterministe basé sur des équations chimiques et des équations de transport prenant en entrée des concentrations de polluants aux bornes du domaine d'étude et des données météorologiques.

CHIMERE et les modèles déterministes comparables, bien qu'intégrant de nombreux mécanismes, sont incorrects (comme tout modèle) et font des erreurs. Blond et al. (2003) et Grancher et al. (2005) présentent des méthodes visant à corriger les erreurs de CHIMERE en combinant les concentrations d'ozone prédites par CHIMERE et celles mesurées dans un réseau de stations. Ces méthodes permettent de modifier les sorties de CHIMERE et d'obtenir des cartes de pollution plus représentatives du champ de pollution réel.

Plutôt que de corriger les sorties d'un modèle déterministe tel que CHIMERE, on peut vouloir corriger sa conception. Les erreurs d'un modèle tel que CHIMERE peuvent être dues à des équations trop simplistes ou encore à des entrées non ou mal prises en compte. Identifier des raisons pour lesquelles un modèle fait des erreurs peut aider à son amélioration. Mon stage de DEA, supervisé par Michael Stein au département de statistique de l'université de Chicago, a porté sur l'analyse du lien entre les erreurs du modèle CHIMERE et des données météorologiques. Disposant des concentrations d'ozone prédites par CHIMERE sur la région parisienne d'une part, et mesurées dans 20 stations Airparif d'autre part, un processus résiduel a été construit (le logarithme du rapport entre prédiction et mesure pris en chaque station Airparif et à chaque pas de temps). Ce processus spatio-temporel a ensuite été modélisé en fonction de covariables météorologiques. Le but était d'éventuellement identifier des covariables jouant un rôle significatif dans le processus résiduel. Mettre en évidence de telles covariables aurait aidé à modifier CHIMERE. Les premiers résultats semblent indiquer que le processus résiduel est significativement fonction de certaines covariables météorologiques. Si ces résultats sont confirmés (une exploration plus approfondie doit être effectuée), cela montre que des effets météorologiques non ou mal pris en compte dans CHIMERE transparaissent dans le processus résiduel.

6.2.3 Que faire de l'information obtenue par l'analyse du processus résiduel

Les exemples précédents montrent qu'un processus résiduel contient de l'information sur des éléments ignorés dans la modélisation. Ils montrent également que l'étude du processus résiduel permet d'identifier certains des éléments ignorés, que ce soit pour un modèle statistique, un processus ponctuel spatial ou même un modèle déterministe. Dans certains cas (simples), l'étude du processus résiduel suggère même une modification évidente du modèle. Par exemple, en régression linéaire, un graphe tels que celui de la figure 6.1 (b) suggère que la variance du bruit est une fonction croissante d'une covariable jouant positivement sur la variable modélisée. Autre exemple, si pour le modèle hiérarchique de Hodges l'analyse des résidus suggère qu'une variable ignorée doit être intégrée dans le modèle, alors ils suffit de l'intégrer (éventuellement à une transformation près). Mais d'autres cas sont plus complexes. Dans le cas de l'expansion du pin noir, comment modifier le modèle ? La fonction de dispersion pourrait par exemple dépendre de la position du parent. Mais quel modèle

choisir pour cette dépendance ? Dans le cas de la propagation de la rouille brune, quelle distribution spatiale choisir pour les fragilités ? La question a été posée et commentée dans les sections 3.7.2 et 3.7.3. Dans le cas du modèle CHIMERE, le pas à franchir est grand entre d'une part l'identification de covariables météorologiques expliquant une partie des erreurs du modèle et d'autre part la modification des équations ou des conditions aux limites du modèle.

La question suivante est donc posée :

Question 6.2. Pour des modèles relativement complexes, peut-on développer une méthode d'analyse de résidus permettant de fournir de l'information directement exploitable pour la modification du modèle utilisé ?

6.3 Analyse de résidus dédiée à la spécification du second niveau d'un modèle hiérarchique

6.3.1 Construire un processus résiduel sous le modèle de base

Considérons un modèle hiérarchique dont on veut spécifier le second niveau. Ce second niveau, en l'absence de connaissance théorique et de données observées s'y rapportant, fait partie de ce qu'ignore le concepteur du modèle (section 6.1). Associer ainsi 'second niveau' à 'ignorance' nous place dans la problématique de la section 6.2 et suggère l'idée suivante : construire et analyser un processus résiduel qui contiendrait de l'information sur le second niveau. Tenant compte de cette idée, la question 6.1 devient :

Question 6.3. Peut-on développer une méthode d'analyse de résidus permettant de spécifier le second niveau d'un modèle hiérarchique, c'est-à-dire la distribution de ses effets aléatoires ?

Nous développons une telle méthode dans les chapitres 7 et 8. Dans cette méthode, le processus résiduel qui est analysé est construit à partir du modèle de base. Indiquons ce qui sous-tend cette approche.

- Le modèle hiérarchique est supposé mimer le phénomène étudié.
- Le second niveau du modèle hiérarchique fait partie de ce qui est ignoré par le modèle de base puisque ce dernier est la version du modèle hiérarchique dans laquelle les effets aléatoires sont supposés égaux.
- Un processus résiduel construit à partir du modèle de base peut donc contenir de l'information sur le second niveau du modèle hiérarchique (i.e. sur ce qui a été ignoré).
- L'information éventuellement obtenue sur le second niveau peut ensuite être exploitée pour le spécifier.

6.3.2 Intégrer le second niveau dans l'analyse de résidus

Qu'est-ce qui diffère entre l'analyse de résidus de Hodges qui suggère une modification évidente du modèle (ajout d'une covariable), et celles faites pour le pin noir, la rouille brune

et CHIMERE qui identifient des lacunes du modèle mais ne suggèrent pas une modification évidente du modèle ? Selon nous, c'est que l'analyse de résidus de Hodges est faite sous une hypothèse nulle (le modèle sans la covariable candidate), qu'une hypothèse alternative est précisée (le modèle avec la covariable candidate), et que la différence entre les deux hypothèses (la covariable candidate) est explicitement intégrée dans l'analyse de résidus. Ainsi, faire de l'analyse de résidus

– sous une hypothèse nulle,
– en précisant une hypothèse alternative, et
– en intégrant explicitement dans l'analyse la différence entre les deux hypothèses,

est une approche pour répondre à la question 6.2.

Notre méthode, esquissée dans la section 6.3.1, correspond à cette approche : l'hypothèse nulle est le modèle de base (effets aléatoires supposés égaux), l'hypothèse alternative est le modèle hiérarchique. Reste à intégrer explicitement dans l'analyse de résidus la différence entre les deux hypothèses, cette différence étant le second niveau du modèle hiérarchique. Pour ce faire, les résidus sont décomposés de manière à faire apparaître les valeurs réalisées des effets aléatoires (cf. les chapitres 7 et 8).

Intégrer explicitement dans l'analyse de résidus la différence entre les deux hypothèses nous permet d'ajouter un second niveau à notre modèle de base, tout comme cela permet à Hodges d'ajouter une covariable à son modèle. Ajouter un second niveau au modèle de base revient à spécifier le second niveau du modèle hiérarchique.

6.3.3 Cadres de développement de l'analyse de résidus dédiée à la spécification du second niveau d'un modèle hiérarchique

L'analyse de résidus dédiée à la spécification du second niveau d'un modèle hiérarchique est développée et appliquée dans les chapitres 7 et 8.
Le chapitre 7 expose la méthode pour des données groupées (ou en clusters) et des modèles hiérarchiques à effets aléatoires partagés. La méthode est appliqée à un GLMM, à un modèle de fragilités partagées et à un modèle décrivant la propagation spatiale à courte distance de la rouille brune du blé. Cette dernière application prolonge l'analyse du jeu de données traité au chapitre 3.
Le chapitre 8 expose la méthode pour des données géostatistiques et des modèles spatiaux hiérarchiques à effets dépendants mais non partagés, c'est-à-dire des modèles hiérarchiques intégrant un processus aléatoire caché. La méthode est appliquée à quatre modèles spatiaux dont les GLMMs spatiaux de Diggle et al. (1998) et de Desassis et al. (2005) et un modèle décrivant la propagation spatiale à longue distance de la rouille brune du blé. Cette dernière application prolonge l'analyse du jeu de données traité au chapitre 4.

7

Modèle hiérarchique intégrant des effets aléatoires partagés

Au chapitre 3, nous avons développé un modèle de propagation spatiale qui a été ajusté à des données de propagation à courte distance de la rouille brune du blé. Les données sont des nombres de lésions par feuille, les feuilles étant localisées dans des quadrats (données groupées ou en clusters). Une étude du variogramme des résidus entre le modèle et les données a indiqué qu'il existe une légère structuration spatiale qui n'a pas été prise en compte dans le modèle. Dans ce chapitre, pour prendre en compte cette structuration, nous proposons d'intégrer au modèle des effets aléatoires partagés : les feuilles d'un même quadrat sont soumis à un même effet aléatoire. Les effets aléatoires servent à refléter les éléments 'structuration spatiale de l'hétérogénéité des feuilles face à l'infection', et 'perturbations locales du potentiel infectieux'. Mais n'ayant pas de connaissance quantitative sur ces deux éléments, comment spécifier la distribution des effets aléatoires ?

Dans ce chapitre, nous proposons une méthode pour spécifier la distribution des effets aléatoires partagés d'un modèle hiérarchique qui décrit le comportement de variables regroupées en clusters. Nous procédons comme suit. Le modèle de base (modèle hiérarchique dans lequel les effets aléatoires sont égaux) est estimé, et un résidu est calculé pour chaque cluster (*cluster residual*). Le lien asymptotique entre résidus et réalisations des effets aléatoires est ensuite déterminé. A partir de ce lien, les effets aléatoires sont estimés (ou restaurés). Puis on sélectionne une spécification pour la distribution des effets aléatoires en utilisant les effets restaurés comme données.

La méthode est appliquée dans trois contextes. Le premier contexte permet de voir comment la méthode fonctionne dans un cas simple : des variables aléatoires suivent des lois de Bernouilli avec des probabilités aléatoires partagées (c'est un GLMM). Le deuxième contexte permet d'évaluer l'efficacité de notre méthode en utilisant des données simulées selon un modèle de durées de vie (modèle de fragilités partagées). Le troisième contexte est celui de la propagation spatiale de la rouille brune du blé : appliquer la méthode à ce contexte nous permet de proposer une spécification pour la distribution des effets aléatoires partagés, effets qui permettent de mieux refléter la variabilité des données.

Residual-based specification of the random-effects distribution for cluster data

By Samuel Soubeyrand, Joël Chadœuf, Ivan Sache and Christian Lannou

We propose a method to help in specifying the distribution of random effects included in a model for cluster data. The class of models we consider includes mixed models and frailty models whose random effects and explanatory variables are constant within clusters. The method is based on cluster residuals obtained by assuming the random effects are equal between clusters. We exhibit an asymptotic relationship between the cluster residuals and variations of the random effects as the number of observations increases and the dispersion of the random effects decreases. The asymptotic relationship is used to estimate functions of the random effects. The method is applied for a simple generalized linear mixed model, for a shared frailty model and for a third kind of model used in botanical epidemiology to describe disease spread.

7.1 Introduction

Cluster or group data arise when repeated measurements are made at each unit of a sample of experimental units. For instance, the following situations result on cluster data. (i) Assessing, for a sample of clinics and for several patients per clinic, the effectiveness of a surgical procedure in clinical surveys (McCulloch and Searle, 2001). (ii) Measuring, at a sample of stations and at several dates per station, the concentration of an air pollutant in environmental surveys (Huerta et al., 2004). (iii) Determining, for a sample of corn ears and for several grains per ear, the presence of a particular genetic mark to assess pollen dispersal (Klein et al., 2003). In general, the distribution of data varies from cluster to cluster. If the variation of the distribution is not entirely explained by observed explanatory variables then data are generally overdispersed.

Neglecting overdispersion can lead to underestimated standard errors (Hinde and Demétrio, 1998; McCullagh and Nelder, 1989) and biased estimators (Henderson and Oman, 1999) for the regression parameters and, consequently, can result on misleading inference. To take account of the overdispersion, random effects can be used. For instance, linear mixed model, generalized linear mixed models (McCulloch and Searle, 2001) and frailty models (Nielsen et al., 1992) include random effects and can be designed to model overdispersed cluster data. In such models, there is one random effect per cluster and the random effects are generally assumed to follow a parametric probability distribution $\mathcal{D}$.

When random effects correspond to poorly known mechanisms as in chapter 3, there is a risk of misspecification of $\mathcal{D}$ which can lead the analyst to draw wrong conclusions. Indeed, Neuhaus et al. (1992) study the misspecification of $\mathcal{D}$ for a Bernouilli-logistic model with random effects. They show that biases of estimators and standard errors are small for the regression parameters but can be high for the intercept and the variance of the random effects. Therefore, if the variability of the studied phenomenon is of interest then the misspecification of $\mathcal{D}$ can affect conclusions.

That is the reason why we propose a method based on data to help in specifying the random-effects distribution $\mathcal{D}$. The method consists in learning information about features of $\mathcal{D}$ which would be used to specify $\mathcal{D}$. The targeted features are, for instance, the expected value, the variance, the skewness, the covariance structure, the form of the marginal distribution of random effects.

More precisely, the method to learn about the random-effects distribution $\mathcal{D}$ is as follows. Consider a parametric model describing the behavior of variables observable over a finite set of clusters. Suppose the value of one of the parameters varies from cluster to cluster. These values can be viewed as realizations of random effects and the model can be viewed as a hierarchical model. Suppose the random-effects distribution $\mathcal{D}$ is unspecified. Our approach to learn about $\mathcal{D}$ through data consists in (i) estimating parameters of the base model, i.e. the model assuming random effects equal, and computing cluster residuals, (ii) exhibiting an asymptotic relationship between the cluster residuals and the realized values of random effects, and (iii) estimating functions of random effects which will be used to specify $\mathcal{D}$.

The method is developed in a likelihood framework. The observable variables can be discrete or continuous. The model can contain explanatory variables constant within clusters. The relationship we exhibit states that, when the number of observations tends to infinity and the dispersion of random effects tends to 0, the vector of cluster residuals is an affine transformation of the realized values of random effects, plus a term asymptotically normal and tending to 0, plus a negligible term. Therefore, the cluster residuals can be approximated by the affine transformation of the realized values of random effects. This transformation being not invertible, we use generalized inverses to determine estimable functions of random effects (Searle, 1982).

The paper is organized as follows. Section 7.2 presents the class of models with random effects that we consider, and the method to learn about the random-effects distribution. Section 7.3 gives a first insight in the method through a simple example. Section 7.4 assesses the efficiency of the method on simulated datasets drawn from a model in survival analysis. Section 7.5 applies the method to a real case : the dataset analyzed in chapter 3 which deals with spatial spread of brown rust of wheat. This real case has motivated the development of the method presented here. The paper is ended by a discussion. In particular, section 7.6.2 places our method in the context of residual analysis.

7.2 Model and method

7.2.1 Model

Let $\mathcal{I} = \{1, \ldots, I\}$, $I \in \mathbb{N}^*$, be a finite set of clusters. Let $Y_1, \ldots, Y_N$ be response variables in $\mathbb{R}$ observable at clusters $C_1, \ldots, C_N$ in $\mathcal{I}$. Conditionally on $\theta = (\theta_1, \ldots, \theta_I)$ in $\Theta^I \subset \mathbb{R}^{dI}$, $d \in \mathbb{N}^*$, couples $(C_1, Y_1), \ldots, (C_N, Y_N)$ are assumed to be independent and identically distributed random vectors in $\mathcal{I} \times \mathbb{R}$ with distribution $\mathbb{P}_\theta$ satisfying

$$\mathbb{P}_\theta(i, dy) = \pi(y|i, \theta_i)\mu(dy)\nu_i, \quad \forall (i, y) \in \mathcal{I} \times \mathbb{R}. \tag{7.1}$$

In expression (7.1), $\nu = (\nu_1, \ldots, \nu_I)^T$ and μ are probability measures on their supports $\mathcal{I}$ and $\mathbb{R}$ respectively ; $\pi(\cdot|i, \theta_i)$ is the conditional distribution with respect to μ of Y_n given $C_n = i \in \mathcal{I}$ and θ_i ; $\theta_i \in \Theta$ is a vector of parameters lying in cluster i and is assumed to be splitted in two components

$$\theta_i = \begin{pmatrix} \alpha_i \\ \beta \end{pmatrix},$$

where $\alpha_i \in \mathbb{R}^{d_1}$ varies from cluster to cluster and $\beta \in \mathbb{R}^{d_2}$ is common to all clusters ($d_1 + d_2 = d$). In fact, the set of parameters used to describe data is $(\alpha_1, \ldots, \alpha_I, \beta)$.

Throughout the paper, β is viewed as a vector of unknown regression parameters. In contrast, $\alpha_1, \ldots, \alpha_I$ are initially viewed as unknown parameters or fixed effects, then they are viewed as random effects whose distribution is unspecified and is investigated. In the following, the term "effects" refers to $\alpha_1, \ldots, \alpha_I$.

π, μ and ν are assumed to be known. Note that as the cluster is an argument of the distribution π, the nature of the distribution can vary from cluster to cluster, and the distribution can depend on explanatory variables constant within clusters. Moreover, response variables can be discrete or continuous depending on μ. The class of models described above includes linear and generalized mixed models and frailty models with shared random effects and shared explanatory variables.

7.2.2 Notations

Set

$$\bar{\alpha} = \sum_{i=1}^{I} \alpha_i \nu_i \qquad \text{and} \qquad \bar{\theta} = \begin{pmatrix} \bar{\alpha} \\ \beta \end{pmatrix}. \tag{7.2}$$

$\bar{\alpha}$ is the mean effect. Let

$$\begin{aligned} \varepsilon &= \sup_{1 \le i \le I} ||\alpha_i - \bar{\alpha}||_\infty \\ &= \sup_{1 \le i \le I} ||\theta_i - \bar{\theta}||_\infty \end{aligned} \tag{7.3}$$

measure the variation of effects, where $||\cdot||_\infty$ is the supremum norm of a vector. The base model, in which random effects are assumed to be equal, corresponds to condition $\varepsilon = 0$ or, equivalently, $\alpha_1 = \cdots = \alpha_I = \bar{\alpha}$.

For all θ in Θ^I, let $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote the expected value and the variance under $\mathbb{P}_\theta$. For all λ in Θ, let $\lambda^{\otimes I}$ in Θ^I denote $(\lambda, \ldots, \lambda)$. Let ∇ and $\mathbf{H}$ respectively be the gradient operator and the Hessian operator with respect to the vector λ applied to functions like $\lambda \mapsto g\{\pi(y|i,\lambda)\}$. Let ∇_k be the k-th component of ∇. Let $\mathcal{D}$ denote the unspecified probability distribution of θ on Θ^I, and $\mathbb{P}_\mathcal{D}$ the joint distribution of (C_n, Y_n, θ)

$$\mathbb{P}_\mathcal{D}(i, dy, d\theta) = \mathbb{P}_\theta(i, dy)\mathcal{D}(d\theta), \quad \forall (i, y, \theta) \in \mathcal{I} \times \mathbb{R} \times \Theta^I.$$

Set $\mathbf{C} = (C_1, \ldots, C_N)$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)$.

7.2.3 Estimation under the base model

Under the base model, $\theta_1 = \cdots = \theta_I = \bar{\theta}$ and the loglikelihood function for the parameter $\lambda \in \Theta$ is

$$l_N(\lambda, \mathbf{C}, \mathbf{Y}) = \frac{1}{N} \sum_{n=1}^{N} \log \pi(Y_n|C_n, \lambda) \tag{7.4}$$

The maximizer $\hat{\lambda}_N$ of $l_N(\cdot, \mathbf{C}, \mathbf{Y})$ exists and is finite under assumptions of regularity (B.b) and (B.c) and concavity (B.d) (assumptions are detailed in appendix B.4). The statistics $\hat{\lambda}_N$, which can be computed from data, is viewed as an estimator of $\bar{\theta}$ whose expression is given by (7.2).

7.2.4 Cluster residuals under the base model

For observation n in $\{1, \ldots, N\}$, let $r_N(n)$ be the ordinary residual estimated under the base model

$$r_N(n) = Y_n - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n). \tag{7.5}$$

For any cluster i in $\mathcal{I}$, define the cluster residual $\hat{R}_{N,i}$ as the mean of ordinary residuals $r_N(n)$ such that $C_n = i$:

$$\hat{R}_{N,i} = \sum_{n=1}^{N} \frac{\delta_{C_n=i}}{\sum_{m=1}^{N} \delta_{C_m=i}} \, r_N(n), \tag{7.6}$$

where $\delta_{c=i} = 1$ if $c = i$, 0 else.

7.2.5 Expression of cluster residuals

Under $\mathbb{P}_\theta$, $\theta \in \Theta^I$, the cluster residuals obtained under the base model $\mathbb{P}_{\lambda^{\otimes I}}$, $\lambda \in \Theta$, are functions of θ. Indeed, we show (appendix B.2) that the cluster residual $\hat{R}_{N,i}$ is the sum of four terms. The first term depends on variations of θ, i.e. $\theta_i - \bar{\theta}$, $i \in \mathcal{I}$. The second and third terms are, up to multiplicative elements, the differences between sample means and their expected values. The fourth term is a negligible term as $N \to \infty$ and $\varepsilon \to 0$ (ε measures variations of θ, see equation (7.3)).

More precisely, by using Taylor expansions and the expression of the bias $\bar{\theta} - \hat{\lambda}_N$ (appendix B.1, corollary B.2), and under assumptions required in corollary B.2 and mentioned in appendix B.2, under $\mathbb{P}_\theta$

$$\begin{aligned}\hat{R}_{N,i} =& A_i(\hat{\lambda}_N)^T \left\{ (\theta_i - \bar{\theta}) - F(\lambda^*, \bar{\theta})^{-1} \sum_{j=1}^{I} F_j(\lambda^*, \bar{\theta})(\theta_j - \bar{\theta})\nu_j \right\} \\ &+ \frac{1}{Q_{N,i}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \delta_{C_n=i} Y_n - Q_{N,i} \mathbb{E}_\theta(Y_n | C_n = i) \right\} \\ &+ A_i(\lambda^*)^T F(\lambda^*, \bar{\theta})^{-1} \left[\frac{1}{N} \sum_{n=1}^{N} \nabla \log \pi(Y_n | C_n, \lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_n | C_n, \lambda^*)\} \right] \\ &+ \xi_{N,\varepsilon,i},\end{aligned} \tag{7.7}$$

where λ^* is the limit of $\hat{\lambda}_N$ as $N \to \infty$; $A_i(\lambda) = \int_{\mathbb{R}} y \nabla \pi(y|i,\lambda)\mu(dy)$; $F_j(\lambda^*, \bar{\theta})$ and $F(\lambda^*, \bar{\theta})$, whose expressions are provided in equations (B.3) and (B.5), are Fisher informations ; $Q_{N,i}$ is the proportion of observations made at cluster i ; and $\xi_{N,\varepsilon,i} = o_{\mathbb{P}_\mathcal{D}\text{-a.s.}}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta} - \lambda^*||_\infty)$ as $N \to \infty$ and $\varepsilon \to 0$.

Equation (7.7) exhibits the link between a cluster residual and variations of effects $\alpha_1, \ldots, \alpha_I$ appearing in the first term at the right side of the equality sign since $\theta_i - \bar{\theta} = \binom{\alpha_i - \bar{\alpha}}{0}$. The next section provides the asymptotic behavior of the second and third terms which appear in equation (7.7).

7.2.6 Asymptotic relationship between cluster residuals and effects $\alpha_1, \ldots, \alpha_I$

Let $\hat{\mathbf{R}}_N$ be the vector of cluster residuals and Δ_α be the vector of variations of effects $\alpha_1, \ldots, \alpha_I$

$$\hat{\mathbf{R}}_N = \begin{pmatrix} \hat{R}_{N,1} \\ \vdots \\ \hat{R}_{N,I} \end{pmatrix} \qquad \text{and} \qquad \Delta_\alpha = \begin{pmatrix} \alpha_1 - \bar{\alpha} \\ \vdots \\ \alpha_I - \bar{\alpha} \end{pmatrix}.$$

In appendix B.3 we state

Theorem 7.1. *Suppose assumptions (B.a)-(B.i) are satisfied. Then*

$$\hat{\mathbf{R}}_N = B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon}. \tag{7.8}$$

where matrix $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ is specified in appendix B.3.1 ; under $\mathbb{P}_\theta$, $\theta \in \Theta^I$,

$$\sqrt{N}\psi_{N,\theta} \xrightarrow{d} \mathcal{N}\{0, \Sigma(\nu, \theta)\} \quad as \quad N \to \infty,$$

where $\Sigma(\nu, \theta)$ is specified by equation (B.23) and the convergence is uniform on Θ^I in the following sense

$$\sup_{\theta \in \Theta^I} \left| \mathbb{E}_\theta \left(e^{\mathbf{i} t^T \sqrt{N} \psi_{N,\theta}} \right) - \mathbb{E}_\theta \left(e^{\mathbf{i} t^T X_\theta} \right) \right| \to 0 \quad as \quad N \to \infty, \quad \forall t \in \mathbb{R}^I,$$

where $X_\theta \sim \mathcal{N}\{0, \Sigma(\nu, \theta)\}$; $\xi_{N,\varepsilon}$ is a vector whose components are $o_{\mathbb{P}_\mathcal{D} - a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta} - \lambda^*||_\infty)$ as $N \to \infty$ and $\varepsilon \to 0$.

7.2.7 Estimable functions of effects $\alpha_1, \ldots, \alpha_I$

When N is large and ε is small, $\hat{\mathbf{R}}_N$ can be approximated by an affine transformation of effects $\alpha_1, \ldots, \alpha_I$

$$\hat{\mathbf{R}}_N \approx \hat{B}\Delta_\alpha, \tag{7.9}$$

where $\hat{B} = B(\hat{\lambda}_N, \hat{\lambda}_N, \hat{\lambda}_N)$ because of theorem 7.1 and because $\hat{\lambda}_N$ approximates λ^* and $\bar{\theta}$ (see appendix B.1). So, solving equation

$$\hat{\mathbf{R}}_N = \hat{B}z \tag{7.10}$$

yields information about variations Δ_α and, consequently, about effects $\alpha_1, \ldots, \alpha_I$. However, $\hat{B}$ is not invertible (see section B.3.1). To overcome this problem, we can use generalized inverses (Searle, 1982).

The Moore-Penrose inverse provides the shortest least squares solution to equation (7.10). We propose this solution, say Δ_α^{MP}, as an estimator of Δ_α. We also look for generalized inverses of $\hat{B}$ which allow us to formally determine estimable functions of random effects, as it is done in analysis of variance (Searle, 1982). The next two paragraphs present this approach in two particular cases.

Case 1

Assume effects are unidimensional ($d_1 = 1$) and there is no regression parameter β ($d_2 = 0$). We can exhibit a matrix $\hat{B}^-$ such that

$$\hat{B}^-\hat{B} = \left(\begin{array}{c|c} \mathbf{I}_{I-1} & -\mathbf{1}_{I-1} \\ \hline 0 \cdots 0 & 0 \end{array}\right), \tag{7.11}$$

where $\mathbf{I}_{I-1}$ and $\mathbf{1}_{I-1}$ are, respectively, the identity matrix and the unit vector of dimension $I-1$. Premultiplying equation (7.9) by $\hat{B}^-$ leads to $\hat{B}^-\hat{\mathbf{R}}_N \approx \hat{B}^-\hat{B}\Delta_\alpha$. Thus, looking for vectors v in $\mathbb{R}^I$ such that $v^T\hat{B}^-\hat{B} = v^T$ allows to estimate contrasts $v^T\Delta_\alpha$ by $v^T\hat{B}^-\hat{\mathbf{R}}_N$. Examples of estimable contrasts are

$$v_{1,i}^T\Delta_\alpha = \alpha_i - \bar{\alpha}, \quad i = 1, \ldots, I \tag{7.12}$$

where $v_{1,i} = 1_i - \nu$, and 1_i is the vector of dimension I whose components are 0 except its i-th component which is 1. Let V_1 be the $I \times I$-matrix whose rows are $v_{1,i}^T$, $i = 1, \ldots, I$, then an estimator of Δ_α is

$$V_1\hat{B}^-\hat{R}_N,$$

and an estimator of $\alpha = (\alpha_1, \ldots, \alpha_I)^T$ is

$$\hat{\bar{\alpha}}\mathbf{1}_I + V_1\hat{B}^-\hat{R}_N,$$

where $\hat{\bar{\alpha}}$ is the estimator of $\bar{\alpha}$ and equals $\hat{\lambda}_N$ as $d_2 = 0$.

Case 2

Assume effects are unidimensional ($d_1 = 1$) and there is one regression parameter ($d_2 = 1$). The presence of a regression parameter decreases the rank of $\hat{B}$ and there does not exist a generalized inverse of $\hat{B}$ satisfying an expression as simple as expression (7.11). Nevertheless we can exhibit a matrix $\hat{B}^-$ such that

$$\hat{B}^-\hat{B} = \left(\begin{array}{c|cc} \mathbf{I}_{I-2} & -\mathbf{w}_0 & \mathbf{w}_0 - \mathbf{1}_{I-2} \\ \hline 0 \cdots 0 & 0 & 0 \\ 0 \cdots 0 & 0 & 0 \end{array}\right), \tag{7.13}$$

where $\mathbf{w}_0$ is a known vector in $\mathbb{R}^{I-2}$. Set $\mathbf{w}^T = (\mathbf{w}_0^T, 1, 0) = (w_1, \ldots, w_I)^T$ and $\bar{w} = \sum_{i=1}^{I} w_i \nu_i$. From expression (7.13), we can show that

$$v_{2,i}^T \Delta_\alpha = \alpha_i - \bar{\alpha} + (\bar{w} - w_i)\{\alpha_{I-1} - \alpha_I\}, \quad i = 1, \ldots, I-2, \tag{7.14}$$

where $v_{2,i} = v_{1,i} + (\bar{w} - w_i)(1_{I-1} + 1_I)$, are estimable contrasts. If two effects, say α_{I-1} and α_I, are known to be equal to 0, let V_2 be the $I \times I$-matrix whose $I-2$ first rows are $v_{2,i}^T$, $i = 1, \ldots, I-2$, and whose 2 last rows are zero, then an estimator of Δ_α is

$$V_2 \hat{B}^- \hat{R}_N,$$

and an estimator of $\alpha = (\alpha_1, \ldots, \alpha_I)^T$ is

$$\hat{\bar{\alpha}} \mathbf{1}_I + V_2 \hat{B}^- \hat{R}_N.$$

where $\hat{\bar{\alpha}}$ is the estimator of $\bar{\alpha}$ picked out from the estimator $\hat{\lambda}_N$.

From this section, we can estimate various functions of effects $\alpha_1, \ldots, \alpha_I$. Moreover, we can use the limiting distribution of $\psi_{N,\theta}$ provided in theorem 7.1 to assess uncertainty of estimators.

7.2.8 Specification of the random-effects distribution

Now, consider effects $\alpha_1, \ldots, \alpha_I$ as random. The previous section allows to compute estimates of their realized values. These estimates can then be used, for example, to estimate the expected value, the standard deviation and the covariance structure of the random effects. They also can be used to test the adequation of the random-effects distribution to a given distribution. Knowledge about these features can help in specifying the random-effects distribution.

Once the model is entirely specified, its parameters must be estimated. If initial values are required in the estimating process, our estimates can provide these initial values.

7.3 A simple example : Bernouilli data with random probabilities

Assume that Y_n follows a Bernouilli distribution with probability $\alpha_i \in [0,1]$ if its cluster of observation C_n is cluster i in $\{1,\ldots,I\}$. Assume in addition that clusters have same probabilities ($\nu_i = 1/I$, $i = 1,\ldots,I$). We aim to specify the distribution of probabilities $\alpha_1,\ldots,\alpha_I$. The general approach consists in estimating the probabilities and deducing their distribution. Here we compare three methods for estimating the probabilities.

Method 1

Considering the probabilities as random effects, the best linear predictor of α_i given the quantity N_i of variables observed at cluster i is (McCulloch and Searle, 2001)

$$\mathbb{E}(\alpha_i) + \frac{n_i \mathbb{V}(\alpha_i)}{\mathbb{E}(\alpha_i) - \mathbb{E}(\alpha_i)^2 + (n_i - 1)\mathbb{V}(\alpha_i)} \{\bar{Y}^{(i)} - \mathbb{E}(\alpha_i)\}.$$

For this estimator, $\mathbb{E}(\alpha_i)$ and $\mathbb{V}(\alpha_i)$ are required. However, it is precisely the kind of features that we investigate. Consequently, this method is unadapted to the context we consider.

Method 2

Considering the probabilities as fixed effects (McCulloch and Searle, 2001), the maximum likelihood estimator of α_i is

$$\bar{Y}^{(i)} = \frac{1}{N_i} \sum_{n=1}^{N} \delta_{C_n=i} Y_n.$$

These estimators can be computed and are, therefore, adapted to the context we consider.

Method 3

Apply our method. The estimator $\hat{\bar{\alpha}}$ of $\bar{\alpha}$ obtained under the base model is $\bar{Y} = N^{-1}\sum_{n=1}^{N} Y_n$. The vector $\hat{R}_N$ is $(\bar{Y}^{(1)} - \bar{Y}, \ldots, \bar{Y}^{(I)} - \bar{Y})^T$. The matrix $\hat{B}$ is $\mathbf{I}_I - I^{-1}\mathbf{J}_I$ where $\mathbf{J}_I$ is the $I \times I$ matrix whose elements are 1. It follows that the estimator of Δ_α is $(\bar{Y}^{(1)} - \bar{Y}, \ldots, \bar{Y}^{(I)} - \bar{Y})^T$, and the estimator of α_i is $\bar{Y}^{(i)}$. So our method and method 2 provide the same estimators of the probabilities.

Method 2 or 3 has provided estimates of the probabilities. This sample can now be used, for example, to estimate moments of the probabilities and to test its adequation to a given parametric distribution. These results will help in specifying the distribution of the probabilities, and parameters of the entirely specified model will be estimated with the generalized-linear-mixed-model methodology (McCulloch and Searle, 2001).

In this simple example, our approach to estimate the probabilities can be replaced by considering the probabilities as fixed and estimating them by maximum of likelihood. But in more complicated situations, considering the effects as fixed and solving the maximum likelihood equations can be cumbersome. For instance, if the maximum likelihood equations are analytically untractable and if there are too many random effects, neither analytical nor numerical solutions can be found. In such cases, our approach to estimate the effects can still be used.

7.4 A simulated case-study in survival analysis

We simulated survival data from a model including shared random effects (or frailties) and a regression parameter for an explanatory variable constant within cluster. In this simulation study, to assess the efficiency of our method, we made vary the number of observations and the variance of the random effects. More precisely, we assessed the ability of our method in estimating the expected value μ_α and the standard deviation σ_α of random effects, and in identifying the random-effects distribution.

7.4.1 Simulated model

It is a model for survival data with exponential distributions (McCullagh and Nelder, 1989) including shared frailties (Nielsen et al., 1992). $Y_1, \ldots, Y_N$ are survival times observable at clusters $C_1, \ldots, C_N$ in the set of clusters $\mathcal{I} = \{1, \ldots, I\}$. Given clusters $C_1, \ldots, C_N$ and random effects $\alpha_1, \ldots, \alpha_I$, survival times $Y_1, \ldots, Y_N$ are assumed to be mutually independent. Given $C_n = i$ and α_i, time Y_n is assumed to be exponentially distributed with mean parameter

$$\alpha_i \exp(\beta x_i),$$

where β is an unknown regression parameter and x_i is the observed value of an explanatory variable constant within cluster i. In the context of survival analysis, random effects $\alpha_1, \ldots, \alpha_I$ are called frailties.

In survival analysis, it is common to use Gamma-distributed frailties because of their mathematical convenience (Nielsen et al., 1992). So, frailties $\alpha_1, \ldots, \alpha_I$ are independently drawn from a truncated gamma distribution with scale parameter 1 and shape parameter σ. The gamma distribution is truncated over its quantile of order 0.95 to satisfy the bounded support assumption (B.a). Explanatory variables $x_1, \ldots, x_I$ are independently and uniformly distributed in $[0, 1]$, and the regression parameter β equals 2.

In this simulation study, there are $I = 25$ clusters, the number J of survival times per cluster varies in $\{4, 10, 20, 40, 100, 200, 1000\}$, and the shape parameter σ varies in $\{0.1, 0.5, 1.0\}$. J controls the number of observations and σ controls the variance of frailties (large σ corresponds to large variance). As a benchmark, in the leukemia survey presented in Henderson et al. (2002), there are 24 clusters and, in average, 43 observations per cluster.

7.4.2 Estimation

Consider $N = I * J$ survival times drawn from the model, and fit the base model assuming frailties $\alpha_1, \ldots, \alpha_I$ all equal to $\bar{\alpha}$. Let $\hat{\bar{\alpha}}$ and $\hat{\beta}$ be the estimates of $\bar{\alpha}$ and β. The frailty vector $\alpha = (\alpha_1, \ldots, \alpha_I)^T$ is estimated by $\alpha^{MP} = \hat{\bar{\alpha}} + \Delta_\alpha^{MP}$ based on the use of the Moore-Penrose matrix (see section 7.2.7). α is also estimated by $\alpha^{EC} = \hat{\bar{\alpha}} + V_2 \hat{B}^- \hat{R}_N$ based on estimable contrasts (7.14). To compute this estimate, clusters are sorted to keep $(\bar{w} - w_i)\{\alpha_{I-1} - \alpha_I\}$, $i = 1, \ldots, I-2$, small. The sorting is based on the expectation that a small cluster residual

corresponds to a small deviation of the random effect from its mean $\bar{\alpha}$: we make correspond cluster I to the smallest cluster residual and cluster $I-1$ to one of the five other smallest cluster residuals which allows to keep small the coefficients $(\bar{w}-w_i)$, $i=1,\ldots,I-2$. Remark that in α^{EC}, estimates of α_{I-1} and α_I are constrained to be $\hat{\bar{\alpha}}$.

7.4.3 Analysis tools

To assess the ability of our method in informing us about the expected value μ_α and the standard deviation σ_α of random effects, we study the relative bias for μ_α

$$\mathrm{RB}(\mu_\alpha) = 100\frac{m(\alpha^X)-\mu_\alpha}{\mu_\alpha},$$

and the relative bias for σ_α

$$\mathrm{RB}(\sigma_\alpha) = 100\frac{sd(\alpha^X)-\sigma_\alpha}{\sigma_\alpha},$$

where X denotes either MP or EC, and $m(\alpha^X)$ and $sd(\alpha^X)$ are the sample mean and the sample standard deviation of components of α^X. Moreover, we study the sample correlation $\mathrm{cor}(\alpha,\alpha^X)$ between frailties and their estimates. Furthermore, by performing Kolmogorov-Smirnov tests, we assess the adequation of the empirical distribution function of estimated frailties α^X (i) to the true truncated gamma distribution and (ii) to the truncated gamma distribution with parameters determined by the method of moments.

7.4.4 Analysis

For each couple (J,σ), 2000 simulations were performed and the averages of the relative biases and the correlations were computed. Results are presented in Tables 7.1 and 7.2.
First, estimates based on the Moore-Penrose matrix are slightly more accurate than estimates based on estimable contrasts (7.14) because, for the latter estimates, constraints $(\bar{w}-w_i)\{\alpha_{I-1}-\alpha_I\}=0$, $i=1,\ldots,I-2$, are only approximately satisfied in the simulations.
Second, when the number of observations $N = I*J$ increases, the correlation between simulated and estimated frailties increases and the relative biases for μ_α and σ_α decrease. So, accuracy of estimated frailties increases with N. It is logical since the approximation $\hat{\mathbf{R}}_N \approx \hat{B}\Delta_\alpha$ is valid as N is large (see sections 7.2.6 and 7.2.7).
Third, for small numbers of observations, the larger the variance of the frailties is (larger σ), the larger the correlation between simulated and estimated frailties is and the larger the relative bias for the expected value is. So, for small N, large variations are more easily detected than small variations, but they are detected with bias.
Fourth, except for the relative bias for σ_α and for small N, the relative biases for μ_α and σ_α increase with the variance of frailties, i.e. with σ. It reflects that the approximation

$\hat{\mathbf{R}}_N \approx \hat{B}\Delta_\alpha$ is valid as the variance of frailties is small (see sections 7.2.6 and 7.2.7). However, as previously noted, the relative bias for σ_α is large for small N. It reflects that in the relationship

$$\hat{\mathbf{R}}_N = B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon}$$

in theorem 7.1, for N not enough large, variations in $\psi_{N,\theta} + \xi_{N,\varepsilon}$ hide variations in $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha$, all the more as ε is small. Heuristically, if $\hat{\mathbf{R}}_N$ is based on few observations, its variations around $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha$ are large and, consequently, variations Δ_α of frailties are hardly detected.

Tab. 7.1. Relative biases for the expected value μ_α (top table) and for the standard deviation σ_α (bottom table) of frailties.

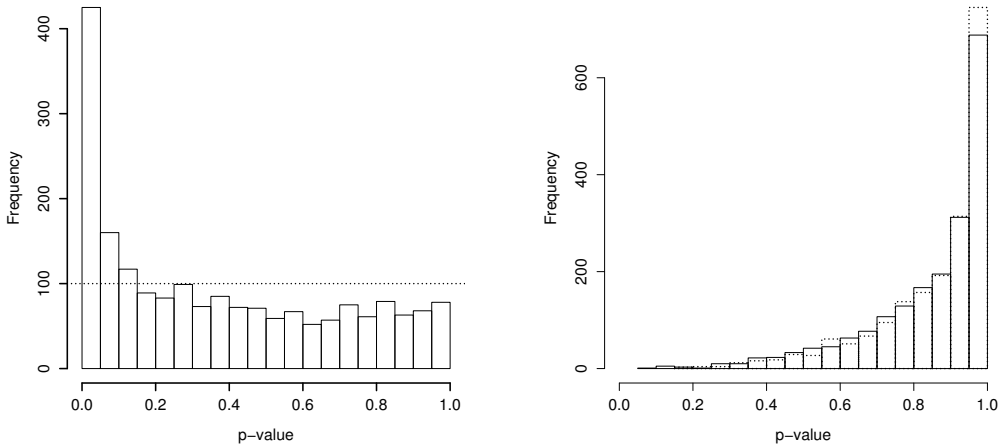
σ	Estimator	Average relative bias for μ_α (%)						
		Number J of observations per cluster						
		4	10	20	40	100	200	1000
0.1	MP	2.6	1.3	0.4	0.4	0.5	0.9	0.3
	EC	2.6	1.3	0.4	0.4	0.5	0.9	0.3
0.5	MP	6.7	4.1	4.5	3.9	2.6	3.3	3.3
	EC	6.8	4.3	4.6	4.0	2.7	3.6	3.4
1.0	MP	12.2	10.1	9.6	9.0	9.0	9.7	9.4
	EC	12.5	10.3	9.7	9.2	9.2	9.8	9.5

σ	Estimator	Average relative bias for σ_α (%)						
		Number J of observations per cluster						
		4	10	20	40	100	200	1000
0.1	MP	105.2	49.8	25.6	13.0	4.1	1.7	-1.9
	EC	110.6	53.1	28.8	15.9	7.0	4.2	0.1
0.5	MP	38.4	17.1	10.0	5.3	2.0	1.5	1.0
	EC	41.8	20.0	12.7	8.3	4.8	4.3	3.6
1.0	MP	31.6	16.3	11.1	8.4	7.7	7.6	7.0
	EC	34.5	19.7	14.7	11.7	11.0	10.3	10.3

Tab. 7.2. Correlation between simulated and estimated frailties.

σ	Estimator	Average sample correlation						
		Number J of observations per cluster						
		4	10	20	40	100	200	1000
0.1	MP	0.47	0.63	0.75	0.84	0.92	0.95	0.97
	EC	0.46	0.62	0.73	0.82	0.89	0.92	0.95
0.5	MP	0.72	0.84	0.90	0.94	0.96	0.97	0.97
	EC	0.70	0.82	0.88	0.91	0.93	0.94	0.95
1.0	MP	0.79	0.88	0.93	0.95	0.96	0.97	0.97
	EC	0.77	0.86	0.90	0.92	0.93	0.94	0.94

Kolmogorov-Smirnov tests of adequation to the true truncated gamma distribution were performed for the 2000 samples of estimated frailties α^{MP} in the case $(J, \sigma) = (40, 0.5)$. Figure 7.1(a) shows the histogram of the p-values of these tests (solid line). If the samples of estimated frailties were drawn from the true law, the histogram should be uniform over [0,1] (dotted line). Actually, there is an excess of small p-values, in particular p-values less than 0.05 : adequation of estimated frailties to the true truncated gamma distribution is rejected for 22% of the samples, instead of 5%, at the risk level 0.05. It is due to the biases for the expected value and the standard deviation. However, as the aim is to specify the distribution of the frailties, we are interested in identifying the form of the distribution even if there is bias for the parameters of the distribution. So, Kolmogorov-Smirnov tests of adequation to the truncated gamma distribution with parameters estimated by the method of moments were performed for the 2000 samples of estimated frailties α^{MP} as well as for the 2000 samples of simulated frailties α. Figure 7.1(b) shows the histograms of the p-values of these tests. Both histograms look similar and the null hypothesis is never rejected at the risk level 0.05. Consequently, even if the true distribution of the frailties is excessively rejected, the right form of the distribution is well identified.



(a) (b)

Fig. 7.1. Distribution of p-values of Kolmogorov-Smirnov tests of adequation. (a) Adequation of the estimated frailties to the true truncated gamma distribution (solid line). (b) Adequation of the estimated frailties (solid line) and of the simulated frailties (dotted line) to the truncated gamma distribution with parameters determined by the method of moments.

7.5 A real study : spread of brown rust of wheat

Chapter 3 presents a model to describe spatial spread of airborne plant diseases. Discussion 3.7 suggests the development of a more sophisticated model which would include spatially structured random effects. However, no information is available to specify the

random-effects distribution. In this section, we apply our method to learn about this distribution and then specify it.

7.5.1 Context and models

Schematically, airborne plant diseases spread as follows : infected plants produce spores which may be released, transported and deposited on susceptible plants ; then, when conditions are conducive, a spore deposited on a plant succeeds in infecting the plant and generates a lesion.

Section 3.2 presents field experiments conducted to assess spatial spread at short distances of the brown rust of wheat. An experiment consists in infecting a plant at the origin of the rectangle $[-27.6, 27.6] \times [-40, 40]$ (in centimeters) and, after the spread, counting lesions for all leaves located in the rectangle partitioned in $I = 36$ quadrats (or clusters). In the experiment we analyze here, there are $N = 256$ leaves in the rectangle (7.1 leaves per quadrat in average). For leaf n in $\{1, \ldots, N\}$, data are the number of lesions Y_n and the quadrat of location $C_n \in \{1, \ldots, I\}$.

The model in chapter 3 is built as follows. Leaf n in $\{1, \ldots, N\}$ is characterized by its location X_n in the rectangle and by an unknown individual frailty Z_n in $[0, 1]$ which determines the propensity of the leaf to be infected by spores. Given $(X_1, Z_1), \ldots, (X_N, Z_N)$, counts of lesions $Y_1, \ldots, Y_N$ are mutually independent and Poisson distributed with means

$$Z_n \frac{\bar{\alpha}}{2\pi\beta_1^2} \exp\left(-\frac{|X_n|}{\beta_1}\right), \quad n = 1, \ldots, N,$$

where $\bar{\alpha}$ and β_1 are positive parameters and $|\cdot|$ is the $\mathbb{R}^2$-Euclidean norm. The term $\bar{\alpha}\exp(-|x|/\beta_1)/2\pi\beta_1^2$ is called the infectious potential at location $x \in \mathbb{R}^2$. Locations $X_1, \ldots, X_N$ are assumed to be independent and uniformly distributed in their respective quadrats $C_1, \ldots, C_N$. Individual frailties $Z_1, \ldots, Z_N$ are assumed to be independent and identically drawn from a distribution parameterized by a two-dimensional vector β_2. From this setting, the conditional distribution of lesion counts given location quadrats can be written and parameters $\bar{\alpha}$ and $\beta = \binom{\beta_1}{\beta_2}$ can be estimated by maximum of likelihood.

In this model, the frailties are assumed to be independent and identically distributed, and the infectious potential is assumed to be isotropic and exponentially decreasing with distance from the source of spores. However, a model check and qualitative knowledge in biology related with heterogeneities in vigor of plants, in 3D-geometry of plants, in fertilization, suggest to question these assumptions (see section 3.6). So, we propose to replace the real parameter $\bar{\alpha}$ by random effects varying from cluster to cluster. The conditional mean of Y_n given the location X_n, the frailty Z_n and the random effects $\alpha_1, \ldots, \alpha_I$ becomes

$$Z_n \frac{\sum_{i=1}^{I} \delta_i(X_n)\alpha_i}{2\pi\beta_1^2} \exp\left(-\frac{|X_n|}{\beta_1}\right),$$

where $\delta_i(X_n) = 1$ if X_n is in quadrat i and 0 else. Thus, frailties can be locally rescaled and the infectious potential can show local departures from the exponential form.

No exogenous quantitative information is available to specify the random-effects distribution. Therefore, we apply our method. Note that in this setting the base model is the model proposed in chapter 3.

7.5.2 Application of the method

First, the maximum likelihood estimate of parameters $(\bar{\alpha}, \beta^T)$ of the base model was computed. We obtained $(\hat{\bar{\alpha}}, \hat{\beta}^T) = (0.89\text{E}^6, 17.0, 8.00, -10.29)$. Second, cluster residuals $\hat{R}_N$, matrix $\hat{B}$ and its Moore-Penrose generalized inverse matrix were computed. Then, using these elements, the estimate $\alpha^{MP} = (\alpha_1^{MP}, \ldots, \alpha_I^{MP})^T$ of $\alpha = (\alpha_1, \ldots, \alpha_I)^T$ was obtained. Figure 7.2 shows the sample distribution of estimated effects α^{MP}. The distribution is skewed. Figure 7.3(a) shows ratios $\alpha_i^{MP}/\hat{\bar{\alpha}}$ for all quadrats $i = 1, \ldots, I$. Quadrats for which ratios are significantly greater or less than 1 at the risk 5% are colored in dark grey and light grey respectively. Significance is decided by using the limiting distribution of $\psi_{N,\theta}$ (see theorem 7.1). As a general rule, we expect more lesions than predicted by the base model at the top of the plot and less lesions at the bottom even if there are exceptions to this rule.

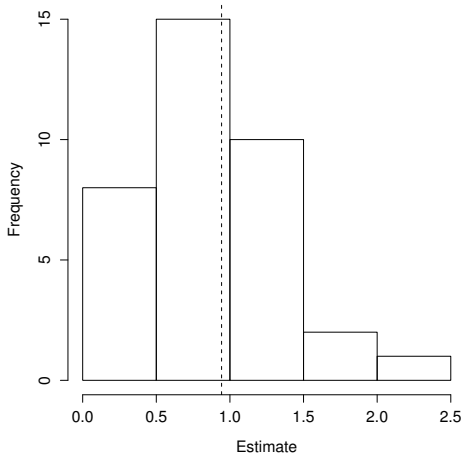
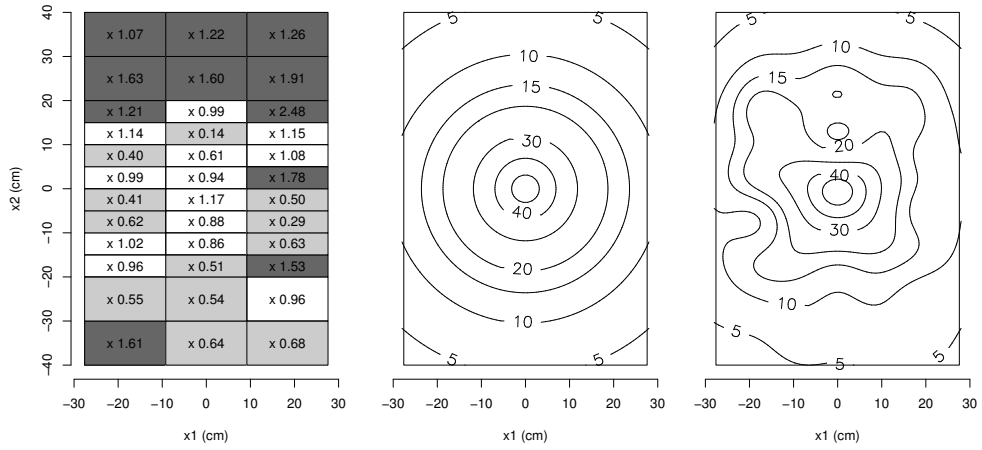


Fig. 7.2. Sample distribution of estimates $\alpha_1^{MP}, \ldots, \alpha_I^{MP}$ up to the multiplicative constant 10^{-6}. Dotted line : value of $\hat{\bar{\alpha}}$.

The distribution of estimated effects was studied. We allocated the effects to their respective quadrats centers and fitted by maximum of likelihood their distribution to a Box-Cox transformed normal random field. The Box-Cox transformation parameterized by λ is defined by (McCullagh and Nelder, 1989)

$$h_\lambda(t) = \begin{cases} \frac{t^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log \lambda & \text{if } \lambda = 0. \end{cases}$$



(a) (b) (c)

Fig. 7.3. (a) Ratios between estimated frailties α^{MP} and estimate $\hat{\bar{\alpha}}$; (b) expected number of lesions per leaf estimated under the base model; (c) expected number of lesions per leaf estimated under the model integrating a smooth version of estimated random effects.

We assumed there exists λ such that $h_\lambda(\alpha_1^{MP}), \ldots, h_\lambda(\alpha_I^{MP})$, i.e. the Box-Cox transformations of estimated effects, are generated by a normal random field characterized by a linear trend, a Matérn covariance function and a nugget effect. The estimate of λ is $\hat{\lambda} = 0.53$. The linear trend is not significant at the risk level 0.05 (likelihood ratio test : $2 \log R = 4.40$, $df = 2$, $p = 0.11$). The nugget effect is significant at the risk level 0.05 (likelihood ratio test : $2 \log R = 4.56$, $df = 1$, $p = 0.03$). See Cressie (1991) and Stein (1999) for elements related with spatial random fields and the Matérn covariance function.

7.5.3 Specification of the random-effects distribution

We then propose to let $\sqrt{\alpha} = (\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_I})^T$ be a normal random vector ($\hat{\lambda} = 0.53$ is close to 0.5) whose mean vector has equal components and whose variance-covariance matrix is based on a spatial covariance structure including a Matérn covariance function and a nugget effect.

7.5.4 Illustration

To illustrate the impact of random effects on the spread of the disease, we draw the following contour plots. Figure 7.3(b) shows the contour plot of the expected number of lesions per leaf under the estimated base model. Figure 7.3(c) shows the contour plot of the expected number of lesions per leaf estimated under the model integrating a smooth version over the rectangle, say $\hat{\alpha}(\cdot)$, of α^{MP}. $\hat{\alpha}(\cdot)$ is obtained by ordinary kriging (Cressie, 1991) based on α^{MP} and the fitted covariance structure mentioned above. The spread which was isotropic around the source of spores under the base model is deformed by the random effects : it is more likely oriented to the top left corner of the plot.

These results can be used to go further in the understanding of the studied phenomenon : the amplitude and the spatial scale of the random-effects variation should help in identifying factors explaining the random effects.

7.6 Discussion

We have considered models with random effects used to describe cluster data. In these models, the random-effects distribution is generally specified before performing data analysis. To avoid misspecification when the random effects correspond to poorly known mechanisms, we have proposed in this paper a method, based on data, to help in specifying the distribution. The method exploits residuals estimated under the base model, i.e. the model assuming equal random effects.

A question arises before specifying the random-effects distribution : where the random effects must be included in the model or, differently speaking, what parameter in the base model must be replaced by random effects ? The answer to this question mainly lies on qualitative knowledge about the studied phenomenon, as in the analysis of disease spread conducted in section 7.5. Note that our method may be used to explore different possibilities for the place of the random effects in the model. However, a criterion for selection should be proposed.

7.6.1 Asymptotic

To apply the method, the number N of observations must be as large as possible and the random-effects variations, measured by ε, must be small but not too small. Indeed, the method is based on the asymptotic relationship

$$\hat{\mathbf{R}}_N = B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon},$$

where $\hat{\mathbf{R}}_N$ is a vector of cluster residuals, $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha$ is a linear function of the random-effects variations, $\psi_{N,\theta}$ is a statistical fluctuation tending to 0 as $N \to \infty$, and $\xi_{N,\varepsilon}$ is a residual error tending to 0 as $N \to \infty$ and $\varepsilon \to 0$. Large N and small ε are required to keep the residual error $\xi_{N,\varepsilon}$ small with respect to the other terms and, consequently, to consider the approximation $\hat{\mathbf{R}}_N \approx B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha + \psi_{N,\theta}$ as valid. Then, for a given N, if ε is too small, the statistical fluctuation hides the linear function of the random-effects variations $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha$. So ε must not be too small to consider the approximation $\hat{\mathbf{R}}_N \approx B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha$ as relevant.

7.6.2 Residual analysis

Our method is based on transformations of ordinary residuals. Various definitions for residuals exist : there are ordinary or raw residuals, standardized residuals, studentized residuals, Pearson residuals and others (see Cook and Weisberg (1982), McCullagh and

Nelder (1989) and Baddeley et al. (2004)). Some of them are transformations of ordinary residuals. The transformation depends on one's aim. In this paper we have transformed ordinary residuals $r_N(1), \ldots, r_N(N)$ to obtain cluster residuals $\hat{R}_{N,1}, \ldots, \hat{R}_{N,I}$. Moreover, we can consider that another transformation was applied : the vector of cluster residuals has been premultiplied by a generalized inverse (see section 7.2.7). Resulting residuals, called thereafter transformed cluster residuals, are viewed as estimators of the realized values of the random effects. Thus, the transformed cluster residuals are designed for our aim : learning about random effects.

Residuals are commonly used in methods developed (i) to detect outliers or influential observations, (ii) to assess model assumptions, (iii) to recognize forms of departure from a model, (iv) to modify a model. See Cook and Weisberg (1982) and Rao and Toutenburg (1995) for linear models, McCullagh and Nelder (1989) for generalized linear models, Box and Jenkins (1976) for time series, Baddeley et al. (2004) for spatial point processes. Our approach is between aims (iii) and (iv). Indeed, we aim to recognize the form of an unspecified random-effects distribution and to modify a base model.

Let us compare our method with the method in time series presented by Box and Jenkins (1976) and using residuals to modify a model. In Box and Jenkins (1976), first an incorrect model is fitted assuming the noise process is white, second a model for the noise process is identified by using residuals estimated under the incorrect model, third the incorrect model and the model for the noise process are combined to arrive at a new model. Our method is quite similar. Let us make the analogy : their incorrect model is our base model, their noise process is our random-effects process, their residuals are our transformed cluster residuals. However, they do not take into account that their residuals are biased estimators of the noise process whereas we do (see section 7.2.4). In addition, the limiting distribution of $\psi_{N,\theta}$ provided in theorem 7.1 allows us to assess how confident we can be in the modification of the base model.

7.6.3 Estimation of the random effects

Although the transformed cluster residuals are estimators of realized values of random effects, we have focused on estimation of features such as the expected value, the standard deviation and the covariance structure of random effects. The reason is that our final aim has been to specify the random-effects distribution and not to estimate the realized values of random-effects. If one's aim is to estimate the realized values, say α, our method could be used as follows.

Step 1 : Estimation of parameters $\bar{\alpha}$ and β of the base model,

Step 2 : Estimation of α by the transformed cluster residuals, say $\hat{\alpha}$.

However, $\hat{\alpha}$ may be a poor estimate of α especially if the variance of random effects, which is generally not known in real studies, is large. In such a situation, $\hat{\alpha}$ partially captures variations of random effects. What is not captured can be viewed as residual random effects whose variance is surely smaller than the variance of the initial random effects. So, we apply

again our method to estimate realized values of the residual random effects, say α^{res} :
Step 3 : Definition of a new base model including $\hat{\alpha}$, i.e. the part of the random effects already captured in Step 2, and estimation of parameters of the new base model conditionally on $\hat{\alpha}$.
Step 4 : Estimation of α^{res} by the transformed cluster residuals.
Steps 3 and 4 could be repeated until the residual random effects are small enough.

Note that in this iterative procedure, the bias for β should be reduced step by step. Note also that no distribution need to be specified for the random effects contrary to the EM algorithm. However, further work need to be done to determine when iterations must be stopped and to manage problems of nonidentifiability.

7.6.4 Residual-based specification of a hidden random field

To extend the method, we are studying the specification of a random-effects distribution in the context of geostatistical data (Cressie, 1991). In this context, random effects form a hidden random field as in the model of Diggle et al. (1998). The main change is that there is generally only one data per site of observation. So cluster residuals cannot be used. Instead, we define local residuals based on a kernel smoothing.

7.6.5 Influence analysis

Influence analysis consists in studying the variations in inferences and conclusions when the formulation of a problem is modified (Cook and Weisberg, 1982). An influence analysis can be performed to assess if a model is an acceptable approximation for instance. Cook (1986) and Critchley and Marriott (2004) perform influence analysis by studying the variation in the results when the used model is perturbed. However, when the space of perturbations is large, exploring it exhaustively is difficult. Therefore, Cook (1986) and Critchley and Marriott (2004) propose complementary methods to explore interesting parts of the space of perturbations. Cook (1986) looks at infinitesimal perturbations in the direction to which his likelihood displacement function is the most sensitive, whereas Critchley and Marriott (2004) look at the perturbations which have most support of the data.

Consider the approach of Critchley and Marriott. The authors assume that perturbations are independent and identically drawn from an unknown distribution $\mathcal{D}$. Then, they acquire information about moments of $\mathcal{D}$ through data. Finally, they make influence analysis by constraining $\mathcal{D}$ to satisfy information learned about its moments.

Suppose the perturbations are cluster perturbations (Xiang et al., 2003). Then, our method could be applied to learn about the distribution of the perturbations, and learned information could constrain the analysis of influential clusters as in Critchley and Marriott (2004).

Nous remercions Rachid Senoussi et Radu Stoica pour leurs commentaires sur la rédaction.

8

Modèle hiérarchique intégrant des effets aléatoires dépendants non partagés

Au chapitre 4, nous avons développé un modèle de propagation spatiale qui a été ajusté à des données de propagation à longue distance de la rouille jaune du blé. Les données sont des nombres de feuilles malades par placette (*trap plot*) d'un mètre carré réparties autour d'une source de spores (une donnée par site). La section 5.3.1 montre qu'il existe des variations locales de la concentration de maladie non prédites par le modèle. Dans ce chapitre, pour prendre en compte cette structuration, nous proposons d'intégrer au modèle un champ aléatoire caché reflétant les perturbations locales du potentiel infectieux attribuées à la dispersion de nuages de spores. Un champ aléatoire est formé d'effets aléatoires définis en chaque site du domaine spatial considéré. En général, les effets aléatoires sont dépendants spatialement (les sites proches ont tendance à avoir des effets similaires). N'ayant pas de connaissance quantitative sur l'élément 'dispersion de nuages de spores', on se demande comment spécifier le champ aléatoire caché.

Dans ce chapitre, nous proposons une méthode pour spécifier un champ aléatoire caché intégré dans un modèle hiérarchique qui décrit le comportement de variables observées dans un espace continu (données géostatistiques). Contrairement au cadre du chapitre précédent où les effets aléatoires sont partagés (les individus d'un même cluster partagent le même effet aléatoire), les effets aléatoires sont maintenant dépendants mais non partagés (un effet pour une donnée). La différence principale dans le développement de la méthode réside principalement dans la définition des résidus : on définit des résidus locaux plutôt que des résidus par cluster. Pour le reste, c'est le même principe : le modèle de base (modèle hiérarchique dans lequel le champ aléatoire est constant) est estimé, et des résidus locaux sont calculés en tous les sites d'observation (*local residual*). Le lien asymptotique entre résidus et réalisations du champ aléatoire est ensuite déterminé. A partir de ce lien, les valeurs du champ en les sites d'observations sont estimées (ou restaurées). Puis on sélectionne une spécification pour le champ aléatoire caché en utilisant les valeurs estimées comme données.

La méthode est appliquée dans quatre contextes. Le premier contexte permet d'évaluer l'efficacité de notre méthode en utilisant des données simulées selon un GLMM spatial 1D. Le deuxième contexte (radioactivité sur Rongelap Island) permet de comparer notre méthode de spécification du champ aléatoire caché à une méthode dans laquelle le champ n'est pas reconstruit. Les deux méthodes aboutissent à la même spécification. Dans les deux

contextes suivants (mortalité des vignes du Languedoc et propagation spatiale de la rouille jaune du blé), le fait de restaurer le champ caché nous conduit à nous interroger sur la non-stationarité du champ. Ce questionnement n'aurait peut-être pas eu lieu si le champ n'avait pas été restauré. C'est là l'avantage principal de notre méthode : en restaurant le champ aléatoire caché, on peut le "visualiser" et, en se basant sur cette visualisation, on peut proposer une classe adaptée de spécifications.

Residual-based specification of a hidden random field

By Samuel Soubeyrand and Rachid Senoussi

We propose a method helping to specify the second stage of a hierarchical spatial model. The hierarchical spatial model describes the behavior of geostatistical data and its second stage is a hidden random field reflecting unobserved and spatially structured heterogeneity. The method can be briefly described as follows : (i) a base model, defined as the hierarchical model assuming the hidden random field constant, is fitted to data, (ii) residuals under the base model are computed, and (iii) an asymptotic relationship between the residuals and the realization of the hidden random field is used to specify the hidden random field. The method is applied to simulated data and to data on radionuclide concentrations on Rongelap Island.

8.1 Introduction

A hidden random field (HRF) is an unobserved process which influences the realization of random variables observable throughout a spatial domain. A HRF can be of interest for itself. Modeling, estimating and mapping the HRF can be useful to go further in the explanation of the studied phenomenon. A HRF can be a nuisance process which must be included in the modeling to avoid misleading inference. In spatial statistics, various models including various kinds of HRF have been proposed. For instance, Diggle et al. (1998) propose generalized linear mixed models including hidden normal random fields to analyze campylobacter-infections and radionuclide-concentrations data ; Henderson et al. (2002) propose a frailty model including a hidden gamma random field to analyze leukemia survival data.

How to specify a hidden random field specially when it corresponds to poorly known mechanisms ? To specify the HRF means to give elements such as the forms of its point distribution, of its trend, of its covariance structure. Attention must be paid to these elements because a misspecification of the HRF can result on wrong conclusions. For example, Monestiez et al. (2005) compare the use of two models including, respectively, a distribution-free hidden random field and a lognormal hidden random field. The models describe counts of fin whales in the Mediterranean sea, and are used to predict the realization of the hidden random field which reflects the abundance of whales. The authors show that variances of predictions are strongly different between the two models. They also show that predictions are different for high values of the hidden random field. To conclude, they question the

relevancy of the logarithm transformation used in the second model, that is the lognormal assumption for the HRF.

For avoiding misspecification of a HRF, we have developed a method based on residuals to catch information about the HRF. The method is as follows. Consider a model including a hidden random field and describing the behavior of variables observable throughout a spatial domain. Suppose the HRF is unspecified whereas the rest of the model is parametrically specified. Let the base model denote the version of the model assuming the HRF constant. Our approach to learn about the HRF through data consists in (i) estimating parameters of the base model and computing local residuals, (ii) exhibiting an asymptotic relationship between the local residuals and the realization of the HRF and (iii) estimating functions of the HRF from this relationship. Targeted functions are for instance the trend, the covariance structure, the point distribution of the HRF.

The paper is organized as follows. Section 8.2 presents the class of models with hidden random fields we consider and the method to learn about the HRF. Section 8.3 assesses the efficiency of our method on simulated datasets. In section 8.4, the method is applied to the data on radionuclide concentrations on Rongelap Island for which Diggle et al. (1998) use a model with a HRF. Results of Christensen (2004) on the specification of the HRF are confirmed : the identity link has to be preferred to the log link used in Diggle et al. (1998). The paper ends by a discussion.

8.2 Model and method

8.2.1 Model

Let $Y_1, \ldots, Y_N$ be response variables in $\mathbb{R}$ observable at locations $Z_1, \ldots, Z_N$ in a spatial domain $\mathcal{Z} \subset \mathbb{R}^q$, $q \in \mathbb{N}^*$. Conditionally on a function $\theta(\cdot)$ from $\mathcal{Z}$ to $\Theta \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$, couples $(Z_1, Y_1), \ldots, (Z_N, Y_N)$ are independent and identically distributed in $\mathcal{Z} \times \mathbb{R}$ with distribution $\mathbb{P}_\theta$ satisfying

$$\mathbb{P}_\theta(dz, dy) = \pi\{y|z, \theta(z)\}\mu(dy)\nu(dz), \quad \forall (z, y) \in \mathcal{Z} \times \mathbb{R}. \tag{8.1}$$

In expression (8.1), ν and μ are probability measures on their respective supports $\mathcal{Z}$ and $\mathbb{R}$; $\pi\{\cdot|z, \theta(z)\}$ is the conditional distribution with respect to μ of Y_n given $Z_n = z \in \mathcal{Z}$ and $\theta(\cdot)$; $\theta(z)$ is a vector of parameters lying at location z and is assumed to be splitted in two components

$$\theta(z) = \begin{pmatrix} \alpha(z) \\ \beta \end{pmatrix},$$

where $\alpha(\cdot)$ is a varying function over the spatial domain $\mathcal{Z}$ and takes values in $\mathbb{R}^{d_1}$ and $\beta \in \mathbb{R}^{d_2}$ is constant over $\mathcal{Z}$ ($d_1 + d_2 = d$).

β is viewed as a vector of unknown regression parameters and $\alpha(\cdot)$ is viewed as the unknown realization of a hidden random field whose distribution is unspecified. Thereafter,

for the sake of shortness, we use the term hidden random field instead of realization of the hidden random field.

π, μ and ν are assumed to be known. Note that as the location is an argument of the distribution π, the distribution can depend on explanatory functions $x : \mathcal{Z} \to \mathbb{R}$ which are location dependent. The probability measure μ over $\mathbb{R}$ is unconstrained, so the response variable can be discrete or continuous. The class of models described above includes the generalized linear mixed models of Diggle et al. (1998) proposed to do geostatistics for non-Gaussian data.

8.2.2 Notations

Set

$$\bar{\alpha} = \int_{\mathcal{Z}} \alpha(z)\nu(dz) \qquad \text{and} \qquad \bar{\theta} = \begin{pmatrix} \bar{\alpha} \\ \beta \end{pmatrix}, \tag{8.2}$$

where $\bar{\alpha}$ is the mean of the hidden random field. Let

$$\begin{aligned} \varepsilon &= \sup_{z \in \mathcal{Z}} ||\alpha(z) - \bar{\alpha}||_\infty \\ &= \sup_{z \in \mathcal{Z}} ||\theta(z) - \bar{\theta}||_\infty \end{aligned} \tag{8.3}$$

measure the variation of the hidden random field, where $||\cdot||_\infty$ is the supremum norm of a vector. The base model, in which the hidden random field is assumed to be constant, corresponds to condition $\varepsilon = 0$ or, equivalently, to $\alpha(z) = \bar{\alpha}$ for all $z \in \mathcal{Z}$.

Let θ be in a class $\mathscr{C}(\mathcal{Z}, \Theta)$ of functions from $\mathcal{Z}$ to Θ. For all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$, let $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote the expected value and the variance under $\mathbb{P}_\theta$. For all λ in Θ, let $\lambda^{\mathcal{Z}}$ denote the function defined over $\mathcal{Z}$ such that $\lambda^{\mathcal{Z}}(z) = \lambda$ for all $z \in \mathcal{Z}$. Let ∇ and $\mathbf{H}$ respectively be the gradient operator and the Hessian operator with respect to the vector λ applied to functions like $\lambda \mapsto g\{\pi(y|z, \lambda)\}$. Let ∇_k be the k-th component of ∇. Let $\mathcal{D}$ denote the unspecified probability distribution of θ on $\mathscr{C}(\mathcal{Z}, \Theta)$, and $\mathbb{P}_{\mathcal{D}}$ the joint distribution of (Z_n, Y_n, θ)

$$\mathbb{P}_{\mathcal{D}}(dz, dy, d\theta) = \mathbb{P}_\theta(dz, dy)\mathcal{D}(d\theta), \quad \forall (z, y, \theta) \in \mathcal{Z} \times \mathbb{R} \times \mathscr{C}(\mathcal{Z}, \Theta). \tag{8.4}$$

Let $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$.

8.2.3 Estimation under the base model

Under the base model $\mathbb{P}_{\lambda^{\mathcal{Z}}}$, $\lambda \in \Theta$, the loglikelihood function is

$$l_N(\lambda, \mathbf{Z}, \mathbf{Y}) = \frac{1}{N} \sum_{n=1}^{N} \log \pi(Y_n|Z_n, \lambda) \tag{8.5}$$

The maximizer $\hat{\lambda}_N$ of $l_N(\cdot, \mathbf{Z}, \mathbf{Y})$ exists and is finite under assumptions of regularity (C.b) and (C.c) and concavity (C.d) (assumptions are detailed in appendix C.4). The statistics $\hat{\lambda}_N$, which can be computed from data, is viewed as an estimator of $\bar{\theta}$ whose expression is given by (8.2).

8.2.4 Local residuals under the base model

For observation n in $\{1,\ldots,N\}$, let $r_N(n)$ be the ordinary residual estimated under the base model

$$r_N(n) = Y_n - \mathbb{E}_{\hat{\lambda}_N^{\mathcal{Z}}}(Y_n|Z_n). \tag{8.6}$$

Define the local residual $\hat{R}_N(s)$ at location $s \in \mathcal{Z}$ as the kernel estimator of $\mathbb{E}_\theta\{Y_n - \mathbb{E}_{\hat{\lambda}_N^{\mathcal{Z}}}(Y_n|Z_n) \mid Z_n = s\}$ which is the conditional expected value of ordinary residual $r_N(n)$ given observation n is made at s. The local residual $\hat{R}_N(s)$ can be written

$$\hat{R}_N(s) = \sum_{n=1}^{N} \frac{K\left(\frac{s-Z_n}{h_N}\right)}{\sum_{m=1}^{N} K\left(\frac{s-Z_m}{h_N}\right)}\, r_N(n), \tag{8.7}$$

where K is a kernel function and h_N is a bandwidth. In the following, let $w_n(s)$ be the weight

$$w_n(s) = \frac{K\left(\frac{s-Z_n}{h_N}\right)}{\sum_{m=1}^{N} K\left(\frac{s-Z_m}{h_N}\right)} \qquad \forall s \in \mathcal{Z}. \tag{8.8}$$

8.2.5 Expression of local residuals

Under $\mathbb{P}_\theta$, $\theta \in \mathscr{C}(\mathcal{Z},\Theta)$, the local residuals obtained under the base model $\mathbb{P}_{\lambda^{\mathcal{Z}}}$, $\lambda \in \Theta$, are functions of θ. Indeed, we show (appendix C.2) that the local residual $\hat{R}_N(s)$ is the sum of five terms. The first two terms depend on variations of $\theta(\cdot)$ at the sample locations, i.e. $\theta(Z_n) - \bar{\theta}$, $n = 1,\ldots,N$. The third and fourth terms are differences between weighted means and their expected values under $\mathbb{P}_\theta$. The fifth term is a negligible term as $N \to \infty$ and $\varepsilon \to 0$ (ε measures variations of θ, see equation (8.3)).

More precisely, by using Taylor expansions and the expression of the bias $\bar{\theta} - \hat{\lambda}_N$ (appendix C.1, corollary B.2) and under assumptions required in corollary B.2 and mentioned in appendix C.2, under $\mathbb{P}_\theta$

$$
\begin{aligned}
\hat{R}_N(s) = & \sum_{n=1}^{N} w_n(s) A(Z_n, \hat{\lambda}_N)^T \{\theta(Z_n) - \bar{\theta}\} \\
& - \left\{ \sum_{n=1}^{N} w_n(s) A(Z_n, \hat{\lambda}_N)^T \right\} F(\lambda^*, \bar{\theta})^{-1} \frac{1}{N} \sum_{n=1}^{N} F(Z_n, \lambda^*, \bar{\theta}) \{\theta(Z_n) - \bar{\theta}\} \\
& + \sum_{n=1}^{N} w_n(s) \{Y_n - \mathbb{E}_\theta(Y_n | Z_n)\} \\
& + \left\{ \sum_{n=1}^{N} w_n(s) A(Z_n, \lambda^*)^T \right\} F(\lambda^*, \bar{\theta})^{-1} \\
& \quad \left[\frac{1}{N} \sum_{n=1}^{N} \nabla \log \pi(Y_n | Z_n, \lambda^*) - \mathbb{E}_\theta \{\nabla \log \pi(Y_n | Z_n, \lambda^*)\} \right. \\
& \quad \left. + \frac{1}{N} \sum_{n=1}^{N} F(Z_n, \lambda^*, \bar{\theta}) \{\theta(Z_n) - \bar{\theta}\} - \mathbb{E}_\theta [F(Z_n, \lambda^*, \bar{\theta}) \{\theta(Z_n) - \bar{\theta}\}] \right] \\
& + \xi_{N,\varepsilon}(s),
\end{aligned}
\tag{8.9}
$$

where λ^* is the limit of $\hat{\lambda}_N$ as $N \to \infty$; $A(z, \lambda) = \int_{\mathbb{R}} y \nabla \pi(y|z, \lambda) \mu(dy)$; $F(z, \lambda^*, \bar{\theta})$ and $F(\lambda^*, \bar{\theta})$, whose expressions are provided in equations (C.3) and (C.5), are Fisher informations; and $\xi_{N,\varepsilon}(s) = o_{\mathbb{P}_{\mathcal{D}}\text{-a.s.}}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta} - \lambda^*||_\infty)$ as $N \to \infty$ and $\varepsilon \to 0$.

Equation (8.9) exhibits the link between a local residual and variations of the hidden random field $\alpha(\cdot)$ which appear in the first two terms at the right side of the equality sign since $\theta(z) - \bar{\theta} = \binom{\alpha(z) - \bar{\alpha}}{0}$.

8.2.6 Asymptotic relationship between local residuals and the hidden random field

Let $s_1, \ldots, s_I$ be $I \in \mathbb{N}$ locations in $\mathcal{Z}$. Let $\hat{\mathbf{R}}_N$ be the vector of local residuals at locations $s_1, \ldots, s_I$ and Δ_α be the vector of variations of the hidden random field at sample locations $Z_1, \ldots, Z_N$

$$
\hat{\mathbf{R}}_N = \begin{pmatrix} \hat{R}_N(s_1) \\ \vdots \\ \hat{R}_N(s_I) \end{pmatrix} \qquad \text{and} \qquad \Delta_\alpha = \begin{pmatrix} \alpha(Z_1) - \bar{\alpha} \\ \vdots \\ \alpha(Z_N) - \bar{\alpha} \end{pmatrix}.
$$

In appendix C.3 we state the following theorem.

Theorem 8.1. *Suppose assumptions (C.a)-(C.i) in appendix C.4 are satisfied. Then*

$$
\hat{\mathbf{R}}_N = B(\hat{\lambda}_N, \lambda^*, \bar{\theta}) \Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon}, \tag{8.10}
$$

where matrix $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ is specified in appendix C.3.1; under $\mathbb{P}_{\mathcal{D}}$

$$
\lim_{N \to \infty} \mathbb{E}\left(e^{\mathbf{i} t^T \sqrt{N h_N^q} \psi_{N,\theta}} \right) - \mathbb{E}\left(e^{\mathbf{i} t^T X_\theta} \right) = 0, \quad \forall t \in \mathbb{R}^I, \tag{8.11}
$$

where $X_\theta | \theta \sim \mathcal{N}\{0, \Sigma(\theta)\}$ and $\Sigma(\theta)$ is diagonal and is specified by equation (C.14); $\xi_{N,\varepsilon} = o_{\mathbb{P}_{\mathcal{D}}-a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta} - \lambda^*||_\infty)$ as $N \to \infty$ and $\varepsilon \to 0$ is a $\mathbb{P}_{\mathcal{D}}$-a.s.-negligible term.

Remark on the convergence of $\psi_{N,\theta}$. In appendix C.3.2 we first show the following pointwise convergence : under $\mathbb{P}_\theta$, $\theta \in \mathscr{C}(\mathcal{Z}, \Theta)$,

$$\sqrt{Nh_N^q}\psi_{N,\theta} \xrightarrow{d} \mathcal{N}\{0, \Sigma(\theta)\} \quad \text{as} \quad N \to \infty.$$

This is the conditional weak convergence of $\psi_{N,\theta}$ given θ. Then we show the unconditional weak convergence of $\psi_{N,\theta}$. Equation (8.11) presents this result in term of characteristic functions.

8.2.7 Estimating equations

$\hat{B} = B(\hat{\lambda}_N, \hat{\lambda}_N, \hat{\lambda}_N)$ is viewed as an estimator of $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ because of equations (C.1) and (C.2) in appendix C.1. From theorem 8.1, when N is large and ε is small, $\hat{\mathbf{R}}_N$ can be approximated by the affine transformation $\hat{B}\Delta_\alpha$ of the variations of the HRF

$$\hat{\mathbf{R}}_N \approx \hat{B}\Delta_\alpha. \tag{8.12}$$

So, we solve the following equation in t

$$\hat{\mathbf{R}}_N = \hat{B}t \tag{8.13}$$

to estimate variations Δ_α. As $\hat{B}$ is not invertible (see section C.3.1), we use generalized inverses (Searle, 1982).

The Moore-Penrose inverse of $\hat{B}$, say $\hat{B}^+$ provides the shortest least squares solution to equation (8.13). We propose this solution, say

$$\hat{\Delta}_\alpha = \hat{B}^+\hat{R}_N,$$

as an estimator of Δ_α. Moreover, we propose

$$\underline{\hat{\alpha}} = \hat{\bar{\alpha}}\mathbf{1}_{d_1 N} + \hat{B}^+\hat{R}_N$$

as an estimator of $\underline{\alpha} = \{\alpha(Z_1)^T, \ldots, \alpha(Z_N)^T\}^T$, where $\hat{\bar{\alpha}}$ is the estimator of $\bar{\alpha}$ picked out from $\hat{\lambda}_N$.

$\hat{\mathbf{R}}_N$ is the vector of local residuals at sites $s_1, \ldots, s_I$ whereas Δ_α is the vector of variations of $\alpha(\cdot)$ at sample sites $Z_1, \ldots, Z_N$. Thus, $\hat{\mathbf{R}}_N = \hat{B}t$ is a system with I equations and N unknowns. Heuristically, to be able to conveniently estimate all components of Δ_α, the grid $\{s_1, \ldots, s_I\}$ must be large enough and must overlap the zone of sample sites. In the applications (sections 8.3 and 8.4) we have let sites $s_1, \ldots, s_I$ be the sample sites.

8.2.8 Specification of the hidden random field

The preceding section provides estimates of realized values of the HRF or their variations at the sample locations ; in other words, the HRF is restored at the sample locations. These estimates can be used, for example, to estimate the expected value, the standard deviation

and the covariance structure of the HRF. They can also be used to test the adequation to a given parametric distribution. Learning information about these features can help in specifying the HRF.

Now, consider that the model with the hidden random field is completely specified. An algorithm such as those proposed by Diggle et al. (1998), Christensen (2004) and Zhang (2002) can be used to estimate the parameters. Our method can provide initial values for the parameters. Moreover, if a Bayesian framework is used, one can check whether the priors are in agreement with information learned from our method.

8.3 Simulated case-study

We simulated data from a spatial overdispersed-Poisson model including a hidden random field, an explanatory variable and a regression parameter. In this simulation study, we assessed the efficiency of our method along the variations of the number of observations and the variance of the hidden random field. In particular, we assessed the ability of our method in estimating the sample mean, standard deviation and variogram of the hidden random field.

8.3.1 Simulated model

Let $Y_1, \ldots, Y_N$ be counts observable at locations $Z_1, \ldots, Z_N$ in the interval $\mathcal{Z} = [0, 1]$. Given locations $Z_1 = z_1, \ldots, Z_N = z_N$ and the hidden random field $\alpha(\cdot)$, counts $Y_1, \ldots, Y_N$ are assumed to be mutually independent and Poisson distributed with means

$$\alpha(z_n) \exp\{\beta x(z_n)\}, \quad n = 1, \ldots, N.$$

where β is an unknown regression parameter and $x(\cdot)$ is a known location-dependent explanatory function.

This model is simulated as follows. Locations $z_1, \ldots, z_N$ are equally spaced over the interval $\mathcal{Z} = [0, 1]$. Regression parameter β equals 1 and explanatory variables $x(z_1), \ldots, x(z_N)$ are independently and identically distributed from the normal distribution with mean 1 and standard deviation 0.5. Moreover, $\alpha(\cdot)$ is a lognormal random field. We use a lognormal form because, even if it does not satisfy the bounded support assumption (C.a), spatial dependences are easily defined. The mean of the logarithm of $\alpha(\cdot)$ is 1 and its covariance function is a powered exponential model $\sigma^2 \exp\{-(20z)^{1.8}\}$, for z in $[0, 1]$.

We have simulated datasets from this model for N in $\{100, 250, 500, 1000\}$ and σ in $\{0.1, 0.5, 1.0, 1.5, 2.0\}$. Realizations of $N = 250$ counts and the underlying random field are shown in Fig. 8.1(a) and (b) for $\sigma = 1.5$.

8.3.2 Estimation

Consider N counts drawn from the model described above, and fit the base model assuming $\alpha(z_1), \ldots, \alpha(z_N)$ all equal to $\bar{\alpha}$. Let $\hat{\bar{\alpha}}$ and $\hat{\beta}$ be the estimates of $\bar{\alpha}$ and β. To compute

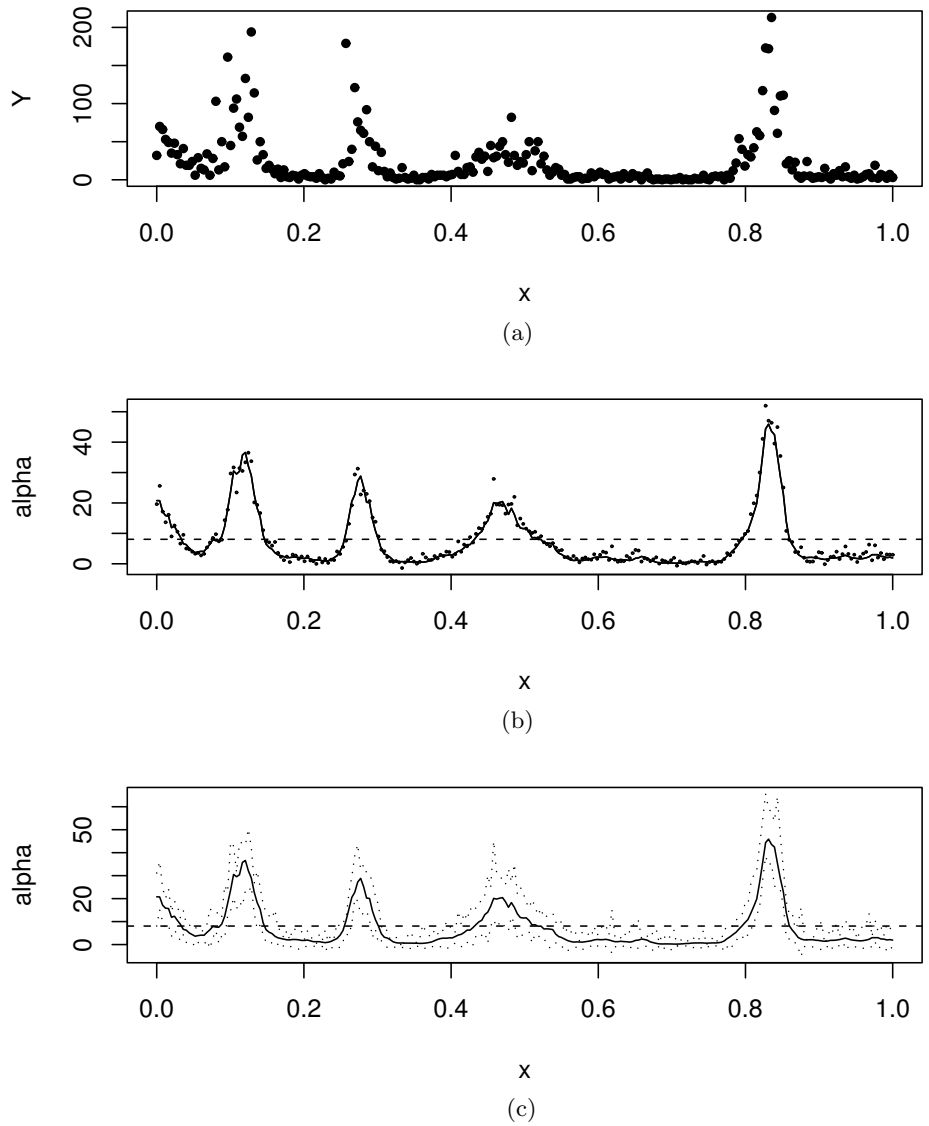


Fig. 8.1. Simulated example : (a) realizations of counts ; (b) realization of the underlying random field (solid line) and estimates of local values of the random field (dots) ; (c) realization of the underlying random field (solid line) and its estimated 95%-confidence envelopes (dotted lines). The dashed lines in plots (a) and (b) give the value of the estimate $\hat{\bar{\alpha}}$ of $\bar{\alpha}$ under the base model.

local residuals, a kernel function and a bandwidth must be chosen. We chose the normal kernel which satisfies required assumptions. We did not address the question of the choice of the bandwidth : as the true $\theta(\cdot)$ is known, we selected the bandwidth h_N which minimizes

$$\sum_{n=1}^{N} \left\{ \mathbb{E}_\theta(Y_n | Z_n = z_n) - \sum_{m=1}^{N} \frac{K\left(\frac{z_n - z_m}{h_N}\right)}{\sum_{l=1}^{N} K\left(\frac{z_n - z_l}{h_N}\right)} Y_m \right\}^2 ,$$

where the second term between the brackets is the kernel estimator of $\mathbb{E}_\theta(Y_n|Z_n = z_n)$. We considered $\hat{\underline{\alpha}} = \hat{\bar{\alpha}}\mathbf{1}_N + \hat{\Delta}_\alpha$ as an estimate of $\underline{\alpha} = \{\alpha(z_1), \ldots, \alpha(z_N)\}^T$ (see section 8.2.7).

We applied the method to simulated counts shown in Fig. 8.1(a). Fig. 8.1(b) shows the realization of the underlying random field together with the estimates $\hat{\underline{\alpha}}$. Fig. 8.1(c) shows the 95%-confidence envelopes for the realization of the hidden random field obtained from the limiting distribution of $\psi_{N,\theta}$ (see Theorem 8.1).

8.3.3 Analysis tools

For assessing the ability of our method in informing us about the sample mean $m(\underline{\alpha})$ and the sample standard deviation $sd(\underline{\alpha})$, we study the relative bias for $m(\underline{\alpha})$

$$\mathrm{RB}_m = 100\frac{m(\hat{\underline{\alpha}}) - m(\underline{\alpha})}{m(\underline{\alpha})},$$

and the relative bias for $sd(\underline{\alpha})$

$$\mathrm{RB}_{sd} = 100\frac{sd(\hat{\underline{\alpha}}) - sd(\underline{\alpha})}{sd(\underline{\alpha})}.$$

For comparing the dependence structures of the HRF $\alpha(\cdot)$ and of the estimate $\hat{\underline{\alpha}}$ of $\underline{\alpha}$, we study the relationship between the sample semivariogram (Chilès and Delfiner, 1999) based on the simulated sample $\underline{\alpha}$ and the sample semivariogram based on the estimated sample $\hat{\underline{\alpha}}$.

8.3.4 Analysis

Sample mean and standard deviation. For each couple (N, σ), 2000 simulations were performed and the averages of the relative biases were computed. Results are presented in Table 8.1.

The relative bias for the mean value decreases as the variance σ of the hidden random field decreases and as the number of observations N increases. Moreover, for variations of $\alpha(\cdot)$ not too small (σ in $\{1.0, 1.5, 2.0\}$), the relative bias for the standard deviation decreases as σ decreases and as N increases. It reflects that the approximation $\hat{\mathbf{R}}_N \approx \hat{B}\Delta_\alpha$ is valid for small σ and large N.

However, for low values of σ (0.1 and 0.5), the relative bias for the standard deviation is high. It reflects that in the relationship

$$\hat{\mathbf{R}}_N = B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon},$$

of Theorem 8.1, $\psi_{N,\theta}$ is strongly varying for small N and hides the vector $B(\lambda_N, \tilde{\lambda}, \bar{\theta})\Delta_\alpha$.

Sample semivariograms. For various couples (N, σ), we computed $\bar{\gamma}$, the average of the 2000 sample semivariograms based on the simulated samples $\underline{\alpha}$, and $\hat{\gamma}$, the average of the 2000 sample semivariograms based on the estimated samples $\hat{\underline{\alpha}}$. In Fig. 8.2 we plotted $\hat{\gamma}$ against $\bar{\gamma}$ (circles). Circles near the origin correspond to small distances and circles in the

Tab. 8.1. Relative bias for the sample mean (left) and the sample standard deviation (right).

σ	Relative bias (%)				σ	Relative bias (%)			
	Number of observations					Number of observations			
	100	250	500	1000		100	250	500	1000
0.1	0.04	0.26	0.00	0.02	0.1	296.63	310.44	316.47	330.39
0.5	1.84	0.71	0.35	0.10	0.5	29.13	29.28	28.99	29.11
1.0	3.26	2.09	1.20	0.40	1.0	8.37	7.75	7.12	6.53
1.5	10.44	4.57	1.86	1.09	1.5	10.73	6.24	3.69	2.91
2.0	18.40	8.34	4.93	2.36	2.0	16.71	8.50	5.58	3.13

top right corner correspond to large distances. The ordinate at the origin reflects a nugget effect, say $\hat{\tau}$, observed on $\hat{\gamma}$. It is due to the uncertainty of estimates $\underline{\hat{\alpha}}$. The abscissa and the ordinate of the circle at the top right are the sills of, respectively, $\bar{\gamma}$ and $\hat{\gamma}$. See Chilès and Delfiner (1999) for definitions of the sill and the nugget effect. Let the relative nugget effect be the nugget effect divided by the sill. Semivariograms $\bar{\gamma}$, whose nugget effect is 0, and $\hat{\gamma}$ can be written for all $x \in [0, 1]$

$$\bar{\gamma}(x) = \bar{\eta}\bar{\gamma}^{norm}(x) \qquad \text{and} \qquad \hat{\gamma}(x) = \hat{\tau} + \hat{\eta}\hat{\gamma}^{norm}(x),$$

where $\bar{\gamma}^{norm}$ and $\hat{\gamma}^{norm}$ are normalized semivariograms whose nugget effects are 0 and sills are 1, and $\bar{\eta}$ and $\hat{\eta}$ are positive constants. If the relationship $(\bar{\gamma}, \hat{\gamma})$ is linear, then the HRF and estimates $\underline{\hat{\alpha}}$ have the same normalized semivariogram, i.e. $\hat{\gamma}^{norm} = \bar{\gamma}^{norm}$. In other words, the dependence structure of the hidden random field and the dependence structure of estimates $\underline{\hat{\alpha}}$ are the same up to a multiplicative constant and a nugget effect. If the relationship $(\bar{\gamma}, \hat{\gamma})$ follows the first bisector, then the dependence structure of the hidden random field and the dependence structure of estimates $\underline{\hat{\alpha}}$ are simply the same.

Fig. 8.2(a), (b) and (c) show that for $N = 250$, normalized semivariograms of the HRF and estimates $\underline{\hat{\alpha}}$ are approximately the same. Indeed, even if the linearity of the relationship $(\bar{\gamma}, \hat{\gamma})$ is degraded as σ increases, the correlation is still close to 1 for $\sigma = 1.5$ (corr $= 0.9972$). More precisely, when σ is small, the dashed line and the relationship $(\bar{\gamma}, \hat{\gamma})$ are parallel $(\bar{\eta} = \hat{\eta})$. So, the semivariograms of the HRF and $\underline{\hat{\alpha}}$ are approximately the same up to the nugget effect $\hat{\tau}$. When σ increases, the relative nugget effect observed on $\hat{\gamma}$ decreases but parallelism is degraded. The relative nugget effect decreases because the part of the variance of $\underline{\hat{\alpha}}$ due to the uncertainty of estimation becomes negligeable relative to the variance of the HRF. Parallelism is degraded because the approximation $\hat{\mathbf{R}}_N \approx \hat{B}\Delta_\alpha$ is valid as the variations of $\alpha(\cdot)$ are small.

Fig. 8.2(d) show that, for $N = 1000$ and $\sigma = 1.5$, the adequation between the relationship $(\bar{\gamma}, \hat{\gamma})$ and the dashed line is quite satisfactory. So the semivariograms of the HRF and $\underline{\hat{\alpha}}$ are approximately the same.

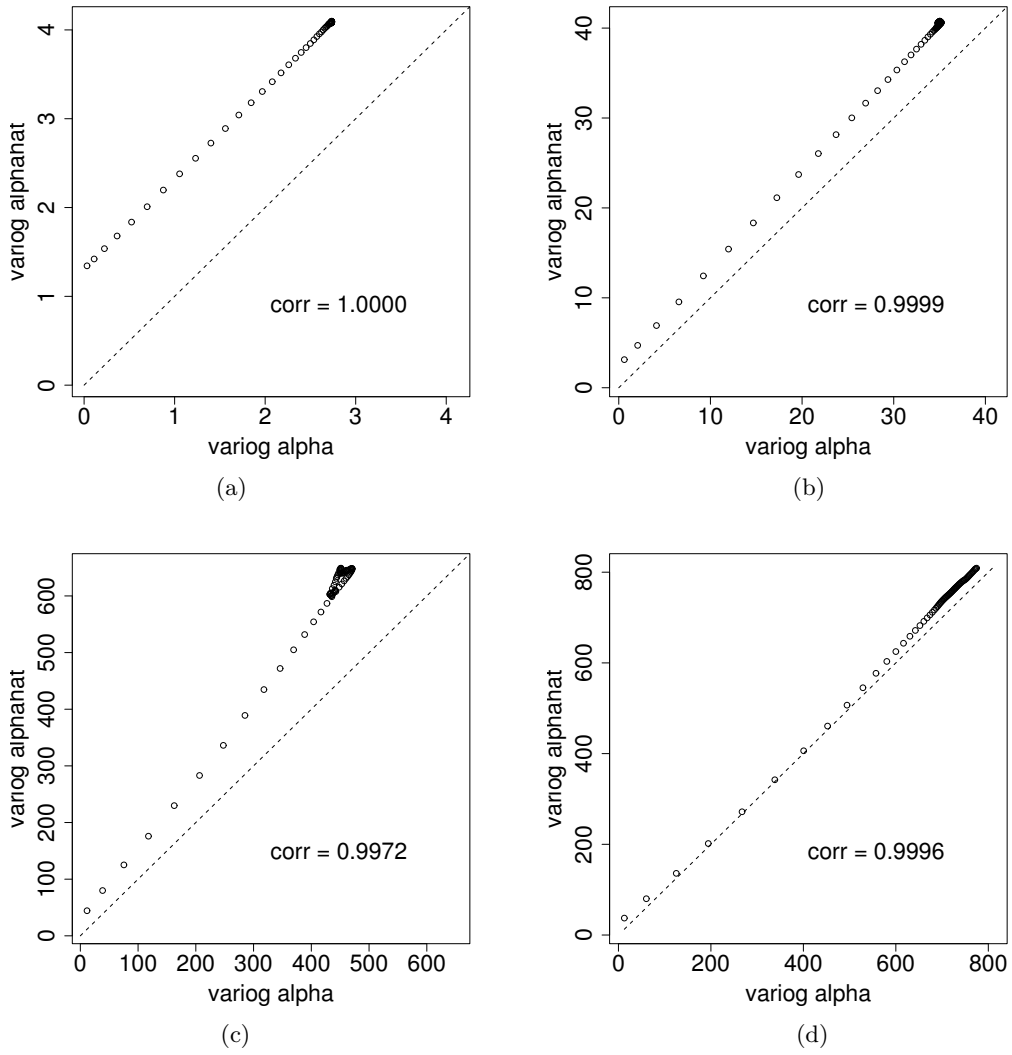


Fig. 8.2. Relationships between semivariograms $\hat{\gamma}$ and $\bar{\gamma}$ (circles) and the first bisector (dashed lines). (a) $N = 250$ and $\sigma = 0.5$; (b) $N = 250$ and $\sigma = 1.0$; (c) $N = 250$ and $\sigma = 1.5$; (d) $N = 1000$ and $\sigma = 1.5$.

8.4 Radionuclide concentrations on Rongelap Island

We applied our method to investigate the nature of a hidden random field $\alpha(\cdot)$ included in a model designed for the analysis of a dataset on radionuclide concentrations (Diggle et al., 1998). More precisely, we investigated the point distribution and the covariance structure of the hidden random field.

8.4.1 Context and Model

Diggle et al. (1998) apply their model-based geostatistics methodology to data on radionuclide concentrations on Rongelap Island in the Pacific Ocean. Data consists in γ-ray counts $y_1, \ldots, y_N$ measured at $N = 157$ locations $z_1, \ldots, z_N$ over the island. Let $x_1, \ldots, x_N$ denote the durations of measurements (the duration of measurement is a space-varying

explanatory variable). Conditionally on an unknown and space-varying intensity of the radioactivity, say $\alpha(\cdot)$, random counts $Y_1, \ldots, Y_N$ are assumed to be mutually independent and Poisson distributed with means

$$\alpha(z_n)x_n, \quad n = 1, \ldots, N.$$

Diggle et al. (1998) assume that the logarithm of $\alpha(\cdot)$ is a normal random field with a constant mean (or trend), a powered-exponential covariance function and a nugget effect ; the authors apply their methodology to predict local values of $\alpha(\cdot)$. Then, they study high values of the intensity of the radioactivity $\alpha(\cdot)$ In the discussion on the paper by Diggle et al. (1998), Ledford and Marriott (1998) ask whether normality of the logarithm of $\alpha(\cdot)$ is an appropriate assumption. The question is relevant as the study of high values of $\alpha(\cdot)$ depends on the heaviness of the tail of the distribution chosen for $\alpha(\cdot)$.

Christensen (2004) answer to Ledford and Marriott by letting $\alpha(\cdot)$ be in a wider class of random fields : he assumes that $\alpha(\cdot)$ is a Box-Cox transformed normal random field and selects the more suitable transformation. More precisely, he assumes there exists $b \geq 0$ such that $f_b\{\alpha(z_1)\}, \ldots, f_b\{\alpha(z_N)\}$ are generated by a stationary normal random field, where

$$f_b(t) = \begin{cases} \frac{t^b - 1}{b} & \text{if } b > 0 \\ \log t & \text{if } b = 0, \end{cases} \tag{8.14}$$

and he uses MCMC maximum likelihood to select b. Christensen (2004) concludes by preferring the identity link ($b = 1$) instead of the log link ($b = 0$) chosen by Diggle et al. (1998). Moreover, he selects a spatial covariance structure including an exponential covariance function and a nugget effect. Note that the optimal b is 0.84 but the author prefers $b = 1$ because the identity link provides a possible additive separation of $\alpha(\cdot)$ in two components : a measurement error component and a spatially structured component.

In the following, we investigate the nature of $\alpha(\cdot)$ using our method and we compare Christensen's results with our results.

8.4.2 Application of the method

We first fitted the base model assuming $\alpha(\cdot)$ congruent with a constant value $\bar{\alpha}$. The estimator of $\bar{\alpha}$ is $\hat{\bar{\alpha}} = N^{-1} \sum_{n=1}^{N} Y_n$ and its value is 7.49. Second, we let K be the normal kernel and selected the bandwidth h_N by generalized cross validation (Loader, 1999). Third, we computed the residuals $\hat{R}_N(z_n)$, $n = 1, \ldots, N$ (we let $s_1, \ldots, s_I$ be $z_1, \ldots, z_N$). Fourth, using a Moore-Penrose generalized inverse, we solved the system of equations (8.13), i.e. $\hat{\mathbf{R}}_N = \hat{B}t$, whose row n in $\{1, \ldots, N\}$ is

$$\hat{R}_N(z_n) = \sum_{m=1}^{N} w_m(z_n)x_m t_m - \hat{x}(z_n) \sum_{m=1}^{N} \frac{x_m}{\sum_{l=1}^{N} x_l} t_m,$$

where $t = (t_1, \ldots, t_N)^T$ and $\hat{x}(z) = \sum_{m=1}^{N} w_m(z)x_m$. Let $\hat{\Delta}_\alpha$ denote the solution of $\hat{\mathbf{R}}_N = \hat{B}t$. The estimate of $\underline{\alpha} = \{\alpha(z_1), \ldots, \alpha(z_N)\}^T$ is

$$\begin{aligned}\underline{\hat{\alpha}} &= \hat{\bar{\alpha}}\mathbf{1}_N + \hat{\Delta}_\alpha \\ &= \{\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)\}^T.\end{aligned}$$

Figure 8.3 shows the histogram (left) and the sample semivariogram (right) of estimated values of $\alpha(\cdot)$ at $z_1, \ldots, z_N$. The histogram is symmetric and unimodal. The sample semivariogram is increasing. These are informal observations since no test is done; however, these observations will allow us to propose a class for the HRF which makes sense, and the specification for the HRF will be selected within this class by performing tests (see below).

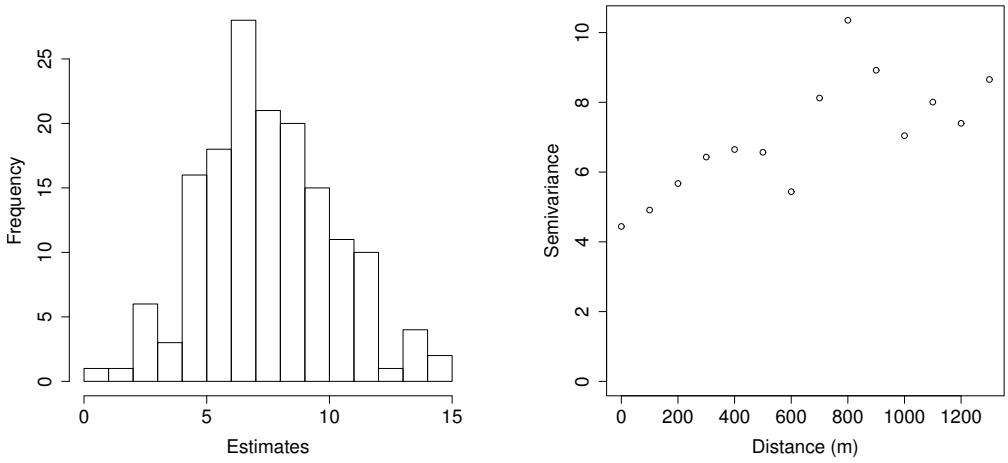


Fig. 8.3. Characteristics of the estimates $\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)$ of the radioactivity intensity. Left : sample distribution of the estimates. Right : sample semivariogram computed using the estimates.

8.4.3 Specification of the hidden random field

As the histogram of $\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)$ is unimodal (Figure 8.3, left) and as $\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)$ shows a spatial dependence decreasing with distance (Figure 8.3, right), we assumed that $\alpha(\cdot)$ is in the class of the Box-Cox transformed normal random fields with constant trends and with spatial covariance structures depending on a Matérn covariance function (Stein, 1999) and a nugget effect. Then, we selected a specification for $\alpha(\cdot)$ within this class using $\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)$ as data. The following presents the selection of the specification.

The Box-Cox transformation is given by expression (8.14). The covariance structure is given by

$$\mathrm{Cov}\{\alpha(z), \alpha(z+z_0)\} = \begin{cases} \sigma^2 + \tau & \text{if } ||z_0|| = 0 \\ \frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{||z_0||}{\phi}\right)^\kappa \mathcal{K}_\kappa \left(\frac{||z_0||}{\phi}\right) & \text{else,} \end{cases} \tag{8.15}$$

where z and z_0 are in $\mathbb{R}^2$, $\tau \geq 0$ is the nugget effect, σ^2 is the variance parameter, $\phi > 0$ is the range parameter, $\kappa > 0$ is the smoothness parameter and $\mathcal{K}_\kappa$ is the modified Bessel

function of the third kind of order κ. $\tau + \sigma^2$ is the sill. The exponential covariance structure selected by Christensen (2004) corresponds to $\kappa = 0.5$.

We first fitted the model for $\alpha(\cdot)$ letting all the parameters free and using $\underline{\hat{\alpha}}$ as data. Parameters were estimated by maximum of likelihood. We found that the maximum likelihood estimate of the transformation parameter b is 0.84, the same value than the one found as optimal by Christensen (2004).

Then we tested some submodels for $\alpha(\cdot)$. For several values of the transformation parameter b, we let all the other parameters free and we fitted the model for $\alpha(\cdot)$. The plot on left of figure 8.4 shows the evolution with b of the maximum loglikelihood. In particular, it provides the maximum loglikelihoods for $b = 0$, i.e. the link chosen by Diggle et al. (1998), and for $b = 1$, i.e. the link selected by Christensen (2004). We performed likelihood ratio tests to test the hypotheses $b = 1$ on one hand and $b = 0$ on the other. The hypothesis $b = 1$ is not rejected ($-2 \log R = 1.31$, $df = 1$ and $p = 0.25$). For $b = 0$, the likelihood ratio test statistic was parametrically bootstrapped (McLachlan, 1987) to get its null distribution because the value 0 is at the border of the domain of parameter b. We performed 1000 replications. The right plot of Figure 8.4 shows the null distribution of the likelihood ratio test statistic together with the observed likelihood ratio test statistic, i.e. 64.08. Hypothesis $b = 0$ is rejected. So we prefer the identity link ($b = 1$) instead of the log link ($b = 0$). Moreover, we performed a likelihood ratio test to test the hypothesis $b = 1$ and $\kappa = 0.5$ (values that Christensen (2004) finally selects). This hypothesis is not rejected ($-2 \log R = 1.31$, $df = 2$ and $p = 0.52$).

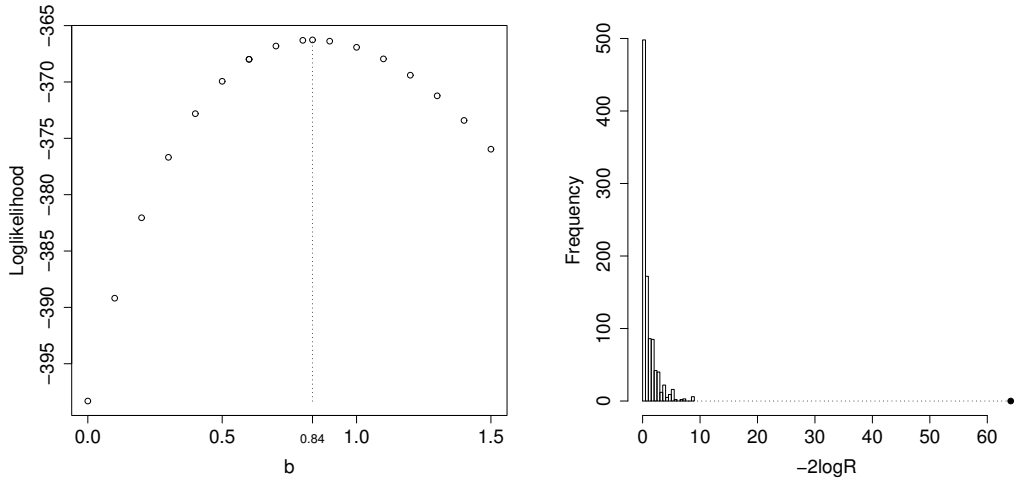


Fig. 8.4. Left : Evolution with b of the maximum loglikelihood. Right : graphical likelihood ratio test of the hypothesis $b = 0$; histogram : null distribution of the LRT statistic obtained by parametric bootstrap ; dot : observed LRT statistic.

Finally, we select the same specification for the intensity $\alpha(\cdot)$ than the one selected by Christensen (2004) :

(S) $\alpha(\cdot)$ is a normal random field with constant trend and with spatial covariance structure depending on an exponential covariance function and a nugget effect.

Note that this specification allows negative values for $\alpha(\cdot)$ although $\alpha(\cdot)$ is a positive function. However, even if the selected distribution cannot be the true distribution of $\alpha(\cdot)$, its right tail is certainly more appropriate than the right tail of the distribution chosen by Diggle et al. (1998).

Remark : table 8.2 shows the estimates for the parameters of the model for $\alpha(\cdot)$ obtained by Christensen (2004) and by us. The estimates for the covariance structure parameters (σ^2 and ϕ) obtained by Christensen and by us are nearly the same but it is not the case for the constant trend and the nugget effect (τ). This comparison is informal since uncertainties about the parameters are not taken into account. However, it shows that estimates obtained from our method must be used with care : our method is designed to specify the HRF, not to estimate the parameters.

Tab. 8.2. Estimates for the parameters of the HRF obtained by Christensen and by us.

	Trend	σ^2	ϕ	τ
Christensen	6.19	2.41	338.2	2.05
Us	7.08	2.39	340.0	4.97

8.5 Discussion

In this paper we have considered models with hidden random fields, thereafter called complete models, and studied how to specify the model for the hidden random field, thereafter called HRF model.

8.5.1 Specifying the HRF model : choice approach and data-based approaches

A first approach for specifying the HRF model consists in choosing a specification before performing data analysis. This approach requires some knowledge about the underlying mechanisms of the analyzed phenomenon. If the analyst does not have at his disposal such a knowledge, the risk of misspecification is high.

Alternatively, there are data-based approaches. For instance, the approach of Christensen (2004) consists in (i) estimating parameters of different complete models corresponding to different HRF models by using observed variables as data and (ii) selecting the more appropriate specification for the HRF model. Our approach is also a data-based approach. It consists in (i') estimating values of the HRF once by using observed variables as data, (ii') estimating parameters of different HRF models by using the estimated values $\underline{\hat{\alpha}}$ as data and (iii') selecting the more appropriate specification.

In our method a preliminary analysis, allowing to figure out what are the main characteristics of the HRF, can be performed after step (i') for proposing relevant HRF models

in step (ii'). Indeed, step (i') is a sort of restoration of the HRF since it provides estimated local values of the HRF. The histogram of these estimated values can be drawn to see whether they show one or several modes (see section 8.4.2), their sample semivariogram can be drawn to see whether they show a spatial structure or not and what kind of structure they show (see section 8.4.2), their map can be drawn to see whether they show a drift (Chilès and Delfiner, 1999) or not (not shown in this paper). Such a preliminary analysis provides guidance for proposing relevant HRF models in step (ii') (see section 8.4.3). In comparison, in step (i) of Christensen (2004), the HRF stays hidden and HRF models are proposed blindly.

Step (i) of Christensen requires to perform several times a MCMC algorithm to fit different complete models. The MCMC algorithm is used because the complete models are hierarchical models. Our step (i') is based on the fit of the base model done once, and our step (ii') is based on the fit of different HRF models. In section 8.4 for instance, the base model is a GLM and the HRF model is a Box-Cox transformed normal random field. For such non hierarchical models, the fitting procedures are simpler and quicker than fitting procedures including MCMC algorithms. Using simple and quick tools eases the exploration of the class of specifications for the HRF model.

If the variance of the HRF is too large, then local values of the HRF can be poorly estimated in our step (i') and our method can be inaccurate (see theory and simulations in sections 8.2 and 8.3). For Rongelap data, the variance of the HRF is not too large since Christensen's method and our method result on the same specification for the HRF. For other data, a method to see whether the variance of the HRF is too large or not would be useful.

8.5.2 The place of the HRF in the complete model

A question arises before specifying the hidden random field : where the hidden random field must be included in the complete model or, in other words, which parameter in the base model must be replaced by a hidden random field ? The answer mainly lies on qualitative knowledge about the studied phenomenon, but it could be based on exploration conducted with our method. However, when performing such an exploration, we must keep in mind that our approach is more efficient if the expected value of the response variable is, under the complete model, linear in the local values of the HRF (see section 8.2).

8.5.3 Residual analysis

Residuals of any model fitted to data can be marked by elements determining data but not integrated in the model. That is the reason why residuals are used in methods developed (i) to detect outliers or influential observations, (ii) to assess model assumptions and (iii) to recognize forms of departure from a model. See Cook and Weisberg (1982) and Rao and Toutenburg (1995) for linear models, McCullagh and Nelder (1989) for generalized linear models, Baddeley et al. (2004) for spatial point processes. In this paper, a phenomenon

including a HRF is considered, a base model which does not integrate the element 'HRF' is fitted to data, and the mark let by the element 'HRF' on residuals is exhibited. The mark is evidenced by the asymptotic relationship between the residuals and the realization of the HRF. This relationship shows how the HRF partly determines the residual values. In this paper this relationship has been exploited to specify the HRF.

Nous remercions Steven L. Simon pour laisser le libre accés aux données de radioactivité sur Rongelap Island, Nicolas Desassis pour ses commentaires sur l'analyse de ces données, et André Kretzschmar pour ses commentaires sur la rédaction.

8.6 Complementary application : Vine plant mortality in Languedoc

In the model for radionuclide concentrations on Rongelap Island, even if there is an explanatory variable, there is no regression parameters. In this section we consider a more complicated model including a hidden random field, explanatory variables and regression parameters. It is designed for the analysis of a dataset on vine plant mortality in Languedoc, France (Desassis et al., 2005). We applied our method to investigate the point distribution and the stationarity of the hidden random field.

8.6.1 Context and Model

Desassis et al. (2005) apply the model-based geostatistics methodology to data on vine plant mortality collected by P. Lagacherie (Institut National de la Recherche Agronomique, UMR LISAH, Montpellier, France). Response variables are counts of dead or declining vine plants $y_1, \ldots, y_N$ in $N = 192$ disks with diameters of 5m. For all the disks, we know the locations of the disk centers $z_1, \ldots, z_N$, the total numbers of vine plants $m_1, \ldots, m_N$, the hydromorphy rates and the textural types of the soil obtained from measures done at the disk centers. The hydromorphy rate and the textural type are explanatory variables with three modalities. Let their interactions be contained in vectors $x_1, \ldots, x_N$ in $\mathbb{R}^8$ (one of the 9 interaction modalities is not estimable). Sampling disks are located around four villages in Languedoc : Caux, Neffiès, Pezenas and Roujan (figure 8.5).

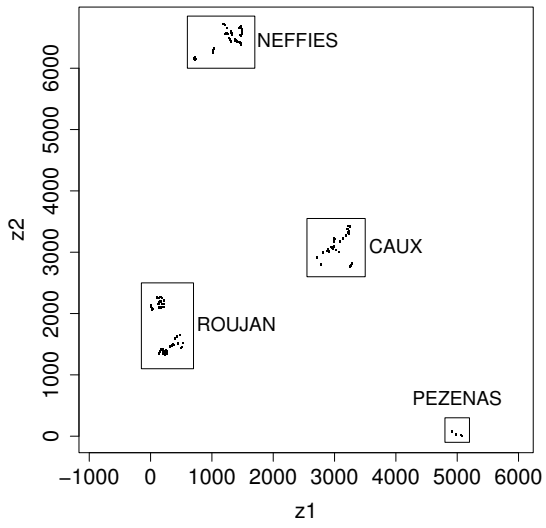


Fig. 8.5. Map with the 192 sampling locations for the vine plant mortality study. The distance scale is in meters.

Conditionaly on an unknown and space-varying intensity of mortality, say $\alpha(\cdot)$, random counts $Y_1, \ldots, Y_N$ are assumed to be mutually independent and drawn from binomial distributions

$$Y_n \sim \text{Binomial}\{m_n, p(z_n)\}, \quad n = 1, \ldots, N$$
$$\text{logit}p(z_n) = \alpha(z_n) + \beta^T x_n,$$

where $\beta \in \mathbb{R}^8$ is a vector of regression parameters and m_n and $p(z_n)$ are the sizes and the probabilities of the binomial distributions.

Desassis et al. (2005) assumes that the mortality intensity $\alpha(\cdot)$ is a normal random field with a constant trend and an exponential covariance function with nugget effect. In the following, we question these assumptions by applying our method.

8.6.2 Application of the method

The method was applied as it has been applied in the case of radionuclide concentrations on Rongelap Island (section 8.4.2). The sole difference is that there are regression parameters which are estimated under the base model and which appear in the expression of $\hat{B}$. We obtained the estimate $\underline{\hat{\alpha}} = \{\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)\}^T$ of $\underline{\alpha} = \{\alpha(z_1), \ldots, \alpha(z_N)\}^T$. Figure 8.6 shows the histogram (left), the boxplots for each village (middle) and the sample semivariogram (right) of the estimated values $\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)$. The histogram is unimodal and skewed. The boxplot for Neffiès seems to be different from the boxplots for Caux and Roujan (the boxplot for Pezenas is not compared to the others since it is based on only 6 values). The sample semivariogram, which is given for distances less than distances between villages, is increasing. These elements indicate that there may be village effects in the structure of the HRF and that there may be a spatial structure within each village. In the following, we propose a class of models for the HRF by taking into account these observations.

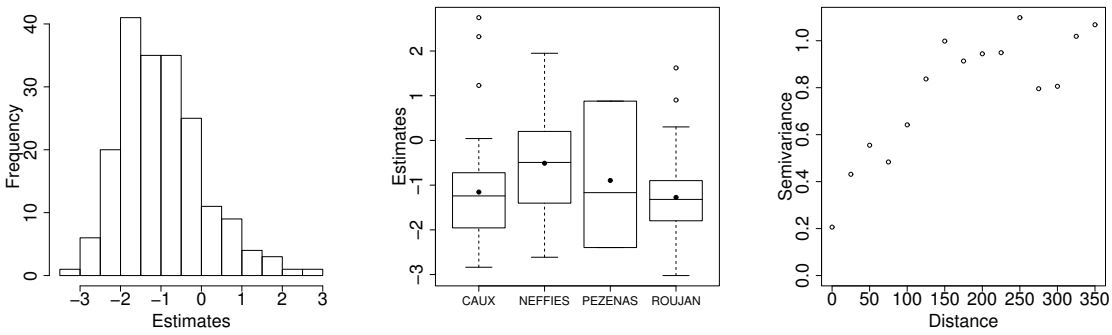


Fig. 8.6. Histogram (left), boxplots for each village (middle) and sample semivariogram (right) of estimates $\hat{\alpha}(z_1), \ldots, \hat{\alpha}(z_N)$ of the mortality intensity. On the boxplots, black points are the means of the estimates for each village.

8.6.3 Specification of the hidden random field

The class of models for the HRF

We assumed that $\alpha(\cdot)$ is a Box-Cox-Pregibon transformed normal random field (McCullagh and Nelder, 1989; Pregibon, 1980), and we used $\underline{\hat{\alpha}}$ to select a specification for $\alpha(\cdot)$ within

this class of random fields. The random field $\alpha(\cdot)$ is said to be in the Box-Cox-Pregibon class if there exist a transformation parameter $b \geq 0$ and a shift parameter $s \in \mathbb{R}$ such that, for any locations $z_1, \ldots, z_N$, $f_{b,s}\{\alpha(z_1)\}, \ldots, f_{b,s}\{\alpha(z_N)\}$ are generated by a normal random field, where $f_{b,s}$ is defined by

$$f_{b,s}(t) = \begin{cases} \frac{(t+s)^b - 1}{b} & \text{if } b > 0 \\ \log(t+s) & \text{if } b = 0. \end{cases}$$

The shift parameter is used because $\alpha(\cdot)$ can take negative values ($f_{b,s}$ is defined over $]-s, \infty[$). The specification for $\alpha(\cdot)$ chosen by Desassis et al. (2005) is in the Box-Cox-Pregibon class : it corresponds to $b = 1$. Note that when $b = 1$, the shift parameter s is no more identifiable because there is confusion with the constant component of the trend of the random field. Consequently, when $b = 1$, we fix s to 0.

We studied different specifications for $\alpha(\cdot)$.

(S_0) $b = 1$ (and $s = 0$), constant trend, exponential covariance function with nugget effect. This is the specification chosen by Desassis et al. (2005).

(S_1) b and s unconstrained, constant trend, Matérn covariance function with nugget effect (see equation (8.15)).

(S_2) b and s unconstrained, village-dependent trend, Matérn covariance function with nugget effect.

(S_3) b and s unconstrained, constant trend, Matérn covariance function with village-dependent σ^2 (equation (8.15)) and with nugget effect.

(S_4) b and s unconstrained, village-dependent trend, Matérn covariance function with village-dependent σ^2 and with nugget effect.

Specifications (S_2)-(S_4) correspond to nonstationary Box-Cox-Pregibon transformed normal random fields. Parameters of these models for $\alpha(\cdot)$ were estimated by maximum of likelihood using $\underline{\hat{\alpha}}$ as data. Likelihood ratio tests were performed to select the more appropriate specification.

Test of (S_0 : Gaussian-exponential) versus (S_1 : Box-Cox-Pregibon-Matérn)

We first tested (S_0) versus (S_1) to see whether the normal shape for the point distribution of $\alpha(\cdot)$ and the exponential covariance function are appropriate. The likelihood ratio test statistic was parametrically bootstrapped to get its null distribution because of the nonidentifiability when $b = 1$. We proceeded as McLachlan (1987) proceeds for the selection of the number of components in a normal mixture. We performed 1000 replications. Figure 8.7 shows the null distribution of the likelihood ratio test statistic together with the observed likelihood ratio test statistic, i.e. 28.74. Specification (S_0) is clearly rejected. Note that (S_0) was rejected because of both the Box-Cox-Pregibon transformation and the covariance structure (tests not shown).

For specification (S_1) as well as for the following specifications, the optimal parameter transformation is $b = 0$ (log link), and the optimal nugget effect is 0.

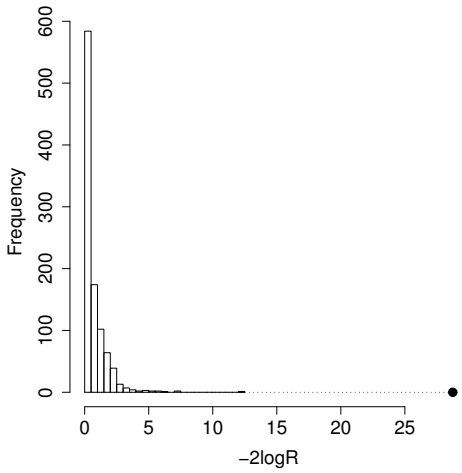


Fig. 8.7. Graphical likelihood ratio test (LRT) (S_0) versus (S_1) for the vine plant mortality study. Histogram : null distribution of the LRT statistic obtained by parametric bootstrap ; dot : observed LRT statistic.

Tests of the stationarity of the HRF

Then we tested (S_1) versus, respectively, (S_2), (S_3) and (S_4) to see whether the stationary assumptions for the trend and for the variance parameter σ^2 are appropriate. As there is no identifiability problem, we used χ^2 distributions as the null distributions for the likelihood ratio test statistics. To perform the tests, we did not used estimates corresponding to Pezenas : as there are only 6 values, the trend and the variance are not statistically robust. Results of the tests are shown in table 8.3. (S_1) is rejected at the risk level 5% when the alternative hypothesis is (S_2). This suggests that there are village effects for the trend of the HRF.

Tab. 8.3. Likelihood ratio tests of the stationarity of the mortality intensity.

	$-2\log R$	df	p
(S_1) vs (S_2)	6.31	2	0.043
(S_1) vs (S_3)	3.60	2	0.165
(S_1) vs (S_4)	9.23	4	0.056

Selected specification

From this analysis, we propose the following specification for $\alpha(\cdot)$. It is, up to an additive constant, a lognormal random field, i.e. there exists a shift parameter $s \in \mathbb{R}$ such that $\log\{\alpha(z_1) + s\}, \ldots, \log\{\alpha(z_N) + s\}$ are generated by a normal random field. The trend of the normal random field is constant within villages but varies from village to village. The covariance function of the normal random field is a Matérn covariance function. It corresponds to specification (S_2) with transformation parameter $b = 0$ and without the nugget effect.

For assessing the effects of the explanatory variables (the hydromorphy rate and the textural type of the soil) and for estimating the parameters of the hidden random field $\alpha(\cdot)$ (the mortality intensity), the model should be fitted as in Desassis et al. (2005).

Je remercie Philippe Lagacherie pour m'avoir permis d'étudier les données de la mortalité des vignes du Languedoc, et Nicolas Desassis pour ses commentaires sur l'analyse de ces données.

8.7 Application complémentaire : propagation spatiale de la rouille jaune

Le chapitre 4 présente des données (figures 4.1 et 4.2) et un modèle paramétrique (équation (4.7)) de propagation spatiale à longue distance pour la rouille jaune du blé. Dans la section 5.3.1, nous avons émis l'hypothèse que les spores pourraient être dispersées en nuages (la figure 5.1 appuie cette hypothèse). Dans cette section, (i) nous présentons un modèle hiérarchique dont le second niveau intègre l'élèment 'dispersion de nuages de spores' mais n'est pas spécifié, et (ii) nous appliquons notre méthode d'analyse de résidus à ce modèle et aux données de propagation à longue distance de la rouille jaune. Appliquer notre méthode nous permet de restaurer le champ aléatoire caché. A partir de ces données restaurées, nous proposons des pistes pour spécifier le champ aléatoire.

8.7.1 Construction d'un modèle

Nous construisons un modèle décrivant les nombres de feuilles malades $Y_1, \ldots, Y_I$ dans I placettes (*trap plots*). $n_1, \ldots, n_I$ et $x_1, \ldots, x_I$ sont, respectivement, les nombres totaux de feuilles et les centres des I placettes.

Zones de dépôt des nuages de spores

La source (point noir sur le schéma de gauche de la figure 8.8) émet, aux temps $t = 1, \ldots, T$, T nuages de spores qui sont déposées dans T zones de dépôt. La zone de dépôt au temps t (en noir sur le schéma de gauche de la figure 8.8) est notée

$$X^t + \Xi^t,$$

où X^t est la position aléatoire dans $\mathbb{R}^2$ du barycentre de la zone de dépôt, et Ξ^t est un sous-espace aléatoire de $\mathbb{R}^2$ dont le barycentre est l'origine[1].

Les barycentres X^t sont supposés indépendants et identiquement distribués selon la densité $h(\cdot)$ donnée au chapitre 5 par la formule 5.2

$$h(x) = \frac{f(\phi_x)}{g(\phi_x)^2} \exp\left\{-\frac{||x||}{g(\phi_x)}\right\}.$$

$h(\cdot)$ est la fonction de dispersion. Ainsi, on suppose que la quantité de zones de dépôt décroit avec la distance et qu'il y a anisotropie de la direction et de la distance de dispersion des barycentres.

Les sous-espaces Ξ^t sont supposés indépendants et identiquement distribués selon une loi $\mathbb{P}_\Xi$ non spécifiée définie sur l'ensemble $\mathfrak{X}$ des sous-espaces de $\mathbb{R}^2$. De plus, les Ξ^t sont supposés indépendants des barycentres X^t.

[1] X^t est appelé 'germe' et Ξ^t est appelé 'grain' dans le contexte des modèles germes-grains (Stoyan et al., 1995; Molchanov, 1997). Ces modèles sont des modèles booléens : tout point x du plan est soit recouvert soit non recouvert par les $X^t + \Xi^t$.

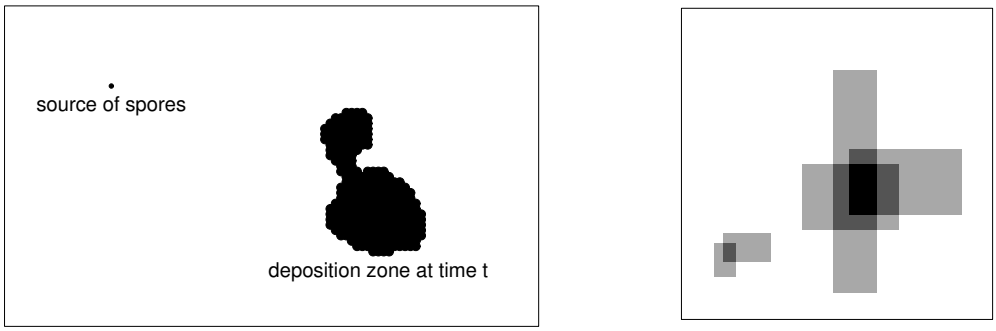


Fig. 8.8. A gauche : zone de dépôt d'un nuage de spores émis par une source ponctuelle à un instant donné. A droite : réalisation d'un modèle à jetons aléatoires ; blanc : zone non recouverte, gris clair : zones recouvertes 1 fois, gris fonçé : zones recouvertes 2 fois, noir : zone recouverte 3 fois.

Processus ponctuel des positions de dépôt des spores

Conditionnellement à la zone de dépôt $X^t + \Xi^t$, les positions de dépôt des spores émises aux temps t forment un processus ponctuel poissonien dont l'intensité est

– constante et égale à λ dans la zone de dépôt (en noir sur la figure 8.8), et

– nulle hors de la zone de dépôt (en blanc sur la figure 8.8).

La valeur λ est supposée indépendante du temps t. La superposition de processus poissoniens indépendants étant un processus poissonien d'intensité la somme des intensités des processus initiaux, les positions de dépôt des spores émises aux temps $1, \ldots, T$ forment, conditionnellement aux zones de dépôt $X^1 + \Xi^1, \ldots, X^T + \Xi^T$, un processus ponctuel poissonien dont l'intensité est

$$\Lambda(x) = \lambda Q(x),$$

où $Q(x)$ est le nombre de zones de dépôt contenant le point x. $Q(\cdot)$ est défini par

$$Q(x) = \sum_{t=1}^{T} \delta_{x \in X^t + \Xi^t}, \quad x \in \mathbb{R}^2. \tag{8.16}$$

$Q(\cdot)$ est appelé modèle à jetons aléatoires (*random token model*, Lantuéjoul, 2002). Le schéma de droite de la figure 8.8 montre une réalisation d'un tel modèle.

Le processus ponctuel des positions de dépôt des spores (non conditionné par les zones de dépôt) est un processus ponctuel de Cox (Diggle, 1983; Stoyan et al., 1995) également appelé processus poissonien doublement stochastique puisque d'une part l'intensité $\Lambda(\cdot)$ est aléatoire et d'autre part, conditionnellement à cette intensité, la répartition des points est aléatoire (et poissonienne).

Loi conditionnelle du nombre de feuilles malades dans une placette

Considérons une sous-espace infinitésimal dx du plan $\mathbb{R}^2$ centré en x ; le nombre de spores déposées dans dx suit, conditionnellement à l'intensité $\Lambda(\cdot)$, une loi de Poisson de moyenne

$\Lambda(x)|dx|$ où $|dx|$ est l'aire de dx. L'espace de dépôt des spores qui nous intéresse n'est pas $\mathbb{R}^2$ mais l'ensemble des surfaces des feuilles. De plus nous nous intéressons non au nombre de spores déposées sur une feuille mais au nombre de lésions sur la feuille. Considérons donc une feuille située en x. Le nombre de lésions créées sur cette feuille suit, conditionnellement à l'intensité $\Lambda(\cdot)$, une loi de Poisson de moyenne $c\Lambda(x)$ où c est un coefficient prenant en compte la capacité de capture de la feuille et la probabilité qu'une spore déposée sur la feuille réussisse à l'infecter. Le coefficient c est supposé indépendant de la feuille considérée.

Conditionnellement aux zones de dépôt, la probabilité qu'une feuille située en x soit infectée, c'est-à-dire qu'au moins une lésion y soit installée, est donc

$$1 - \exp\{-c\Lambda(x)\} = 1 - \exp\{-c\lambda Q(x)\}.$$

Par conséquent, le nombre Y_i de feuilles infectées parmi n_i dans la placette i centrée en x_i suit, conditionnellement à $Q(x_i)$ et en supposant que $Q(\cdot)$ est constante dans la placette, une loi binomiale

$$Y_i|Q(x_i) \sim \text{Binomiale}[n_i, 1 - \exp\{-c\lambda Q(x_i)\}]. \tag{8.17}$$

Conditionnellement à $Q(\cdot)$, les $Y_1, \ldots, Y_I$ sont mutuellement indépendants.

Donnons la loi marginale de $Q(x)$ (équation (8.16)). Les sous-espaces Ξ^t étant indépendants et identiquement distribués, le nombre $Q(x)$ de zones de dépôt parmi T contenant x, suit une loi binomiale de taille T et de probabilité $p(x) = \mathbb{P}(x \in X^t + \Xi^t)$

$$Q(x) \sim \text{Binomiale}\{T, p(x)\}.$$

La fonction $h(\cdot)$ étant la densité de X^t, on montre que

$$\begin{aligned} p(x) &= \int_{\mathcal{X}} \int_{x-\xi} h(s) ds \mathbb{P}_{\Xi}(d\xi) \\ &= h(x)\mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|) + \mathbb{E}_{\mathbb{P}_{\Xi}}\left[\int_{\Xi^t} \{h(x) - h(x-s)\} ds\right], \end{aligned}$$

où $\mathbb{P}_{\Xi}$ est, rappelons-le, la distribution de Ξ^t dans l'ensemble $\mathcal{X}$ des sous-espaces de $\mathbb{R}^2$, $\mathbb{E}_{\mathbb{P}_{\Xi}}(\cdot)$ est l'espérance sous $\mathbb{P}_{\Xi}$, et $|\Xi^t|$ est l'aire de Ξ^t. La distribution $\mathbb{P}_{\Xi}$ n'étant pas spécifiée, $p(x)$ ne peut être calculé. Néanmoins, en supposant que $\mathbb{E}_{\mathbb{P}_{\Xi}}[|\Xi^t|\{h(x)-h(x-s)\}]$ est négligeable devant $\mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|)h(x)$ ($h(\cdot)$ varie peu à l'échelle des Ξ^t probables), $p(x)$ peut être approximé par

$$p(x) \approx \mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|)h(x). \tag{8.18}$$

La loi marginale de $Q(x_i)$ étant une loi de binomiale de taille T et de probabilité $p(x_i) \approx \mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|)h(x_i)$ (cf. équation (8.18)), $Q(x_i)$ fluctue autour de son espérance $T\mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|)h(x_i)$. Il existe donc une variable aléatoire $A(x_i)$ d'espérance 1 telle que $Q(x_i) = A(x_i)T\mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|)h(x_i)$. De plus, définissons $\alpha(\cdot)$ par

$$\alpha(x_i) = c\lambda T\mathbb{E}_{\mathbb{P}_{\Xi}}(|\Xi^t|)A(x_i).$$

D'après cette équation et la loi de $Y_i|Q(x_i)$ (équation (8.17)), on a

$$Y_i|\alpha(x_i) \sim \text{Binomiale}[n_i, 1 - \exp\{-\alpha(x_i)h(x_i)\}], \tag{8.19}$$

et conditionnellement aux $\alpha(x_1), \ldots, \alpha(x_I)$, les $Y_1, \ldots, Y_I$ sont mutuellement indépendants.

Le modèle obtenu pour $Y_1, \ldots, Y_I$ est un modèle hiérarchique à deux niveaux. Le premier niveau décrit la loi de $Y_1, \ldots, Y_I$ conditionnellement au champ aléatoire $\alpha(\cdot)$. Le second niveau correspond à $\alpha(\cdot)$ qui est un champ aléatoire défini sur $\mathbb{R}^2$ et à valeurs dans un espace discret (parce que $Q(\cdot)$ l'est et $\alpha(x_i) = c\lambda Q(x_i)/h(x_i)$). La distribution ponctuelle (en $x \in \mathbb{R}^2$) de $\alpha(\cdot)$ est spécifiée du fait de la relation entre $\alpha(\cdot)$ et $Q(\cdot)$ et parce que $Q(\cdot)$ suit une loi binomiale. En revanche, la structure de covariance de $\alpha(\cdot)$ n'est pas spécifiée car la distribution des zones de dépôt $\mathbb{P}_{\Xi^t}$ ne l'est pas.

8.7.2 Généralisation du modèle

La construction précédente permet d'appréhender ce qu'est l'élément 'dispersion de nuages de spores' et comment cet élément peut être intégré dans un modèle de propagation. Cette construction repose sur des hypothèses fortes (taille réduite des zones de dépôt Ξ^t, intensité λ constante sur les zones de dépôt et indépendante de t). Mais ces hypothèses permettent de déteminer la loi conditionnelle (8.19) du nombre de feuilles malades Y_i dans la placette i.

Comparons le modèle (8.19) au modèle développé au chapitre 4 qui décrit le comportement de Y_i en ignorant l'élément 'dispersion de nuages de spores' et qui peut être écrit

$$Y_i \sim \text{Binomiale}[n_i, 1 - \exp\{-\bar{\alpha}h(x_i)\}], \tag{8.20}$$

où $\bar{\alpha}$ représente la force de la source de spores et la fonction $S(\cdot) = \bar{\alpha}h(\cdot)$ la fonction potentiel infectieux (cf. équations (4.7) et (4.8)). Dans le modèle (8.19), la variable aléatoire $\alpha(x_i)$ remplace la constante $\bar{\alpha}$. Le champ aléatoire $\alpha(\cdot)$ ne décrit plus seulement la force de la source puisque, en plus, il perturbe localement le potentiel infectieux $S(\cdot)$.

Au vu de cette interprétation de $\alpha(\cdot)$, généralisons le modèle (8.19) en considérant que $\alpha(\cdot)$ n'est plus le résultat de la construction présentée dans les paragraphes précédents, mais que c'est un champ aléatoire (positif) très général qui perturbe localement le potentiel infectieux. Le modèle devient : conditionnellement aux $\alpha(x_1), \ldots, \alpha(x_I)$, les $Y_1, \ldots, Y_I$ sont mutuellement indépendants et

$$Y_i|\alpha(x_i) \sim \text{Binomiale}[n_i, 1 - \exp\{-\alpha(x_i)h(x_i)\}], \tag{8.21}$$

où $\alpha(\cdot)$ est un champ aléatoire positif dont la structure de covariance et la distribution ponctuelle ne sont pas spécifiées. Afin de spécifier ce champ aléatoire caché, nous appliquons dans les sections suivantes notre méthode d'analyse de résidus.

8.7.3 Application de la méthode : restauration des effets aléatoires et exploration des effets restaurés

La méthode décrite dans la section 8.2 a été appliquée aux données de propagation de la rouille jaune et au modèle hiérarchique (8.21). Notons que le modèle de base associé au modèle hiérarchique (8.21) correspond au modèle (8.20) dont les paramètres ont été estimés par maximum de vraisemblance au chapitre 4. Notons $\underline{\hat{\alpha}} = \{\hat{\alpha}(x_1), \ldots, \hat{\alpha}(x_I)\}^T$ l'estimation du vecteur $\underline{\alpha} = \{\alpha(x_1), \ldots, \alpha(x_I)\}^T$.

La figure 8.9 montre l'histogramme (en haut à gauche) des estimations $\hat{\alpha}(x_1), \ldots, \hat{\alpha}(x_I)$ qui est unimodal, symétrique et à queues lourdes. Cette figure montre également le semi-variogramme des estimations. Il semble y avoir une structuration spatiale jusqu'à cent mètres. Le graphe de droite de la figure 8.9 représente la carte des estimations. La structuration décelée dans le variogramme est visible sur cette carte : il y a des zones où de grands cercles se côtoient, des zones où de petits cercles se cotoient, mais surtout des zones où des cercles de taile moyenne se côtoient (ces cercles correspondent au mode de l'histogramme).

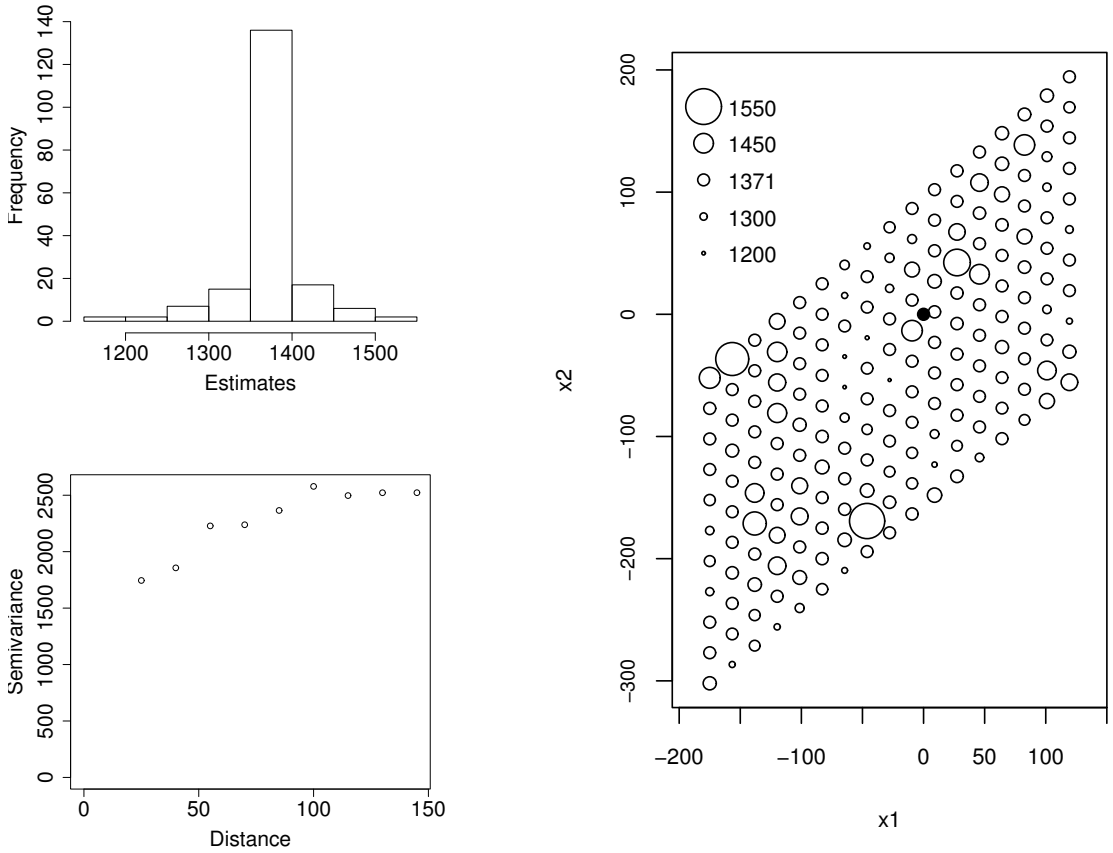


Fig. 8.9. Histogramme (en haut à gauche), semi-variogramme empirique (en bas à gauche) et carte (à droite) des estimations $\hat{\alpha}(x_1), \ldots, \hat{\alpha}(x_I)$. Pour indication, la valeur estimée de $\bar{\alpha}$ sous le modèle de base est $\hat{\bar{\alpha}} = 1371$.

8.7.4 Pistes pour la spécification du champ aléatoire caché qui perturbe le potentiel infectieux

Nous n'avons pas encore trouvé une spécification satisfaisante pour le champ aléatoire caché $\alpha(\cdot)$ qui perturbe localement le potentiel infectieux. Cependant les commentaires faits sur la figure 8.9 donnent des pistes que nous n'aurions pas si les effets aléatoires $\alpha(x_1), \ldots, \alpha(x_I)$ n'avaient pas été restaurés. En effet, au vu du semi-variogramme et de la carte, $\alpha(\cdot)$ semble être un champ aléatoire structuré spatialement. Si l'on regarde plus en détail, le fait qu'il y ait de grandes zones où $\alpha(\cdot)$ est quasiment constant et qu'il y ait des zones où $\alpha(\cdot)$ est variable semble indiquer que $\alpha(\cdot)$ a une structure de covariance non stationnaire[2]. Un cadre pour construire des champ aléatoire gaussien avec des covariances non-stationaires est proposé par Pintore and Holmes (2004). Ce cadre pourrait être utilisé afin de trouver une spécification satisfaisante pour notre champ aléatoire caché $\alpha(\cdot)$.

8.7.5 Bilan de l'analyse de résidus effectuées sur les données de propagation à longue distance de la rouille jaune

L'analyse menée dans les sections précédentes indique que le potentiel infectieux est effectivement perturbé. Nous avons attribué ces perturbations à l'élément 'dispersion de nuages de spores'. D'autres éléments pourraient être la cause de ces perturbations, parmi lesquels l'hétérogénéité des placettes en terme de propension à être infectées et la présence de zones d'accumulation des spores[3]. Nous avons attribué les perturbations à l'élément 'dispersion de nuages de spores' parce que rien ne semblait jouer en faveur des autres hypothèses. Ainsi, spécifier le champ aléatoire $\alpha(\cdot)$ et estimer les paramètres du modèle hiérarchique l'incluant devrait permettre, par exemple, de caractériser ce que nous appelons 'nuages de spores' car la structuration de $\alpha(\cdot)$ est liée à la distribution de la taille et de la densité des nuages de spores.

[2] Nous avons essayer d'ajuster un modèle de Box-Cox avec une covariance stationnaire, mais un test d'ajustement a rejeté ce modèle à cause d'une sous-estimation de la lourdeur des queues de la distribution.

[3] La présence d'une haie au bord du champ pourrait jouer le rôle de barrage à spores, et ces spores qui auraient dû aller plus loin pourraient être déposées juste avant la haie à cause des turbulences que celle-ci crée. La zone avant la haie serait alors une zone d'accumulation des spores.

9

Améliorations et extensions possibles de la méthode de spécification du second niveau d'un modèle hiérarchique

La méthode développée dans les deux chapitres précédents permet de spécifier le second niveau d'un modèle hiérarchique intégrant des effets aléatoires. Deux cadres sont traités : dans le premier, les données sont groupées (en clusters) et les effets aléatoires sont partagés au sein des groupes (chapitre 7) ; dans le second, les données sont des données géostatistiques (récoltées sur un support continu) et les effets aléatoires forment un champ aléatoire caché (chapitre 8). Nous retraçons dans la section 9.1 du présent chapitre les grandes lignes de la méthode de spécification. Puis nous nous proposons de discuter des améliorations possibles de la méthode (réduction du biais dans la décomposition des résidus, restauration des effets aléatoires sous contrainte) et des extensions possibles de la méthode à d'autres cadres de modélisation (autres modèles hiérarchiques intégrant des effets aléatoires et modèles déterministes). Ce chapitre de synthèse s'ajoute aux discussions 7.6 et 8.5 des chapitres 7 et 8.

9.1 Résumé de la méthode de spécification

Décomposition linéaire des résidus

Les résidus (*cluster residuals* ou *local residuals*) obtenus en estimant le modèle de base sont, sous le modèle hiérarchique (à effets aléatoires), la somme

- d'une fonction linéaire des effets aléatoires,
- d'une fluctuation stochastique tendant vers 0 quand le nombre de données croît, et
- d'un biais tendant vers 0 (plus vite que les deux termes précédents) quand le nombre de données croît et la variance des effets aléatoires tend vers 0

$$\begin{aligned} \text{résidus} &= \text{fonction linéaire des effets aléatoires} \\ &\quad + \text{fluctuation stochastique} + \text{biais} \end{aligned} \tag{9.1}$$

(cf. équations (7.8) et (8.10)). Cette décomposition des résidus est dite linéaire.

On peut concevoir trois niveaux de réalisme dans la représentation des effets aléatoires ; ces niveaux de réalisme, illustrés par la figure 9.1, aident à comprendre comment est obtenue la décomposition des résidus. Au premier niveau de réalisme, les effets aléatoires sont

constants : on retrouve le modèle de base sous lequel sont calculés les résidus. Au deuxième niveau, les effets aléatoires du modèle hiérarchique sont 'peu variables' : cela permet de linéariser les résidus calculés sous le modèle de base et d'obtenir la décomposition énoncée plus haut. Au troisième niveau, les effets aléatoires sont non contraints : on retrouve le modèle hiérarchique qui est sensé mimer le phénomène étudié.

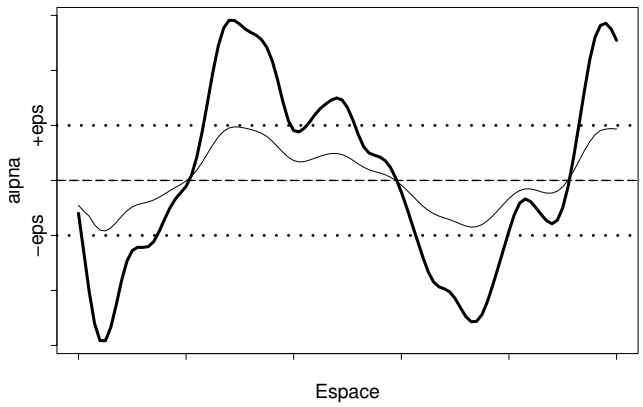


Fig. 9.1. Les trois niveaux de réalisme dans le cas où les effets aléatoires forment un champ aléatoire $\alpha(\cdot)$ sur un espace à une dimension. Tirets : $\alpha(\cdot)$ constant et égale à 0 ; trait fin : $\alpha(\cdot)$ contraint dans la bande $[-\varepsilon, \varepsilon]$ (pointillés) ; trait épais : $\alpha(\cdot)$ non contraint.

Approximation linéaire des résidus

La décomposition linéaire des résidus permet d'obtenir l'approximation suivante

$$\text{résidus} \;\approx\; \text{fonction linéaire des effets aléatoires} \tag{9.2}$$

La validité de cette approximation dépend de la quantité de données et de la variabilité des effets aléatoires. En effet, la fluctuation stochastique est négligeable si la quantité de données est suffisamment grande. Le biais est négligeable si la quantité de données est suffisamment grande et si les effets aléatoires sont 'peu variables', c'est-à-dire si les deuxième et troisième niveaux de réalisme sont quasiment confondus. Nous revenons plus loin sur le lien entre validité de l'approximation (9.2) et variabilité des effets aléatoires.

Estimation des effets aléatoires

L'approximation (9.2) s'écrit

$$\hat{R}_N \approx \hat{B}\Delta_\alpha,$$

où $\hat{R}_N$ est le vecteur des résidus, Δ_α est le vecteur des variations des effets aléatoires et $\hat{B}$ est une matrice. L'estimation des effets aléatoires passe par la résolution en t du système d'équations

$$\hat{R}_N = \hat{B}t.$$

Ce système est résolu en utilisant une inverse généralisée de $\hat{B}$. Estimer les effets aléatoires revient à restaurer le processus caché (ou sous-jacent) influant la variable observée au premier niveau du modèle hiérarchique.

Spécification de la distribution des effets aléatoires

Estimer les effets aléatoires permet d'estimer leurs moments, leur semivariogramme, de tracer leur histogramme ou encore de les cartographier dans l'espace. Obtenir de l'information sur ce type de caractéristiques oriente le choix d'une classe de modèles, ou de spécifications, pour la distribution des effets aléatoires (thème développé à la section 8.5.1). Ainsi, c'est en partie parce que les effets aléatoires sont restaurés que notre méthode de spécification est 'non aveugle' (cf. section 6.1.3 et question 6.1).

Une fois qu'une classe de modèles pour la distribution des effets aléatoires est choisie, un modèle à l'intérieur de cette classe peut être sélectionné. La sélection est opérée en utilisant les estimations des effets aléatoires comme données (i.e. conditionnellement aux valeurs estimées des effets aléatoires).

Dans la suite, nous expliquons comment la méthode de spécification pourrait être améliorée et étendue. Pour ce faire, nous discutons de la validité de l'approximation (9.2), de la résolution sous contrainte du système d'équations (9.6), et de l'extension de la méthode à d'autres types de modèles que ceux traités dans les deux précédents chapitres.

9.2 Réduction du biais dans la décomposition des résidus pour obtenir une approximation des résidus moins sensible à la variabilité des effets aléatoires

Variabilité des effets aléatoires et validité de l'approximation linéaire des résidus

L'approximation linéaire (9.2)

$$\text{résidus} \approx \text{fonction linéaire des effets aléatoires}$$

est obtenue sous le modèle hiérarchique à effets aléatoires 'peu variables' (deuxième niveau de réalisme). Si les effets aléatoires sont 'trop variables' dans le troisième niveau de réalisme, la linéarisation peut rendre le biais de la décomposition linéaire (9.1) non négligeable, l'approximation (9.2) invalide, les estimateurs des effets aléatoires basés sur cette approximation biaisés (cf. les études sur données simulées des sections 7.4 et 8.3) et, par voie de conséquence, la méthode de spécification inefficace. Développer un moyen permettant de détecter quand une telle situation advient s'avèrerait utile pour savoir quand la méthode de spécification est efficace et quand elle ne l'est pas. Cela permettrait également de définir les notions de 'peu variable' et 'trop variable'.

Plutôt que de se demander si les effets aléatoires sont trop variables pour appliquer notre méthode, nous pouvons chercher à obtenir une décomposition moins sensible à la variance des effets aléatoires. Pour ce faire, nous devons chercher à réduire le biais de la décomposition linéaire (9.1). Nous discutons ci-dessous deux solutions : la décomposition non-linéaire des résidus et la reparamétrisation du modèle.

Décomposition non linéaire des résidus

Dans ce paragraphe, nous explicitons comment nous avons obtenu la décomposition linéaire des résidus, et nous montrons comment nous pourrions en obtenir une décomposition non linéaire dans laquelle le biais serait réduit. Adoptons les notations du chapitre 7. Le résidu ordinaire pour la donnée n

$$r_N(n) = Y_n - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)$$

est décomposé en

$$\begin{aligned} r_N(n) =& \{Y_n - \mathbb{E}_\theta(Y_n|C_n)\} \\ &+ \{\mathbb{E}_\theta(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\} \\ &+ \{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}. \end{aligned}$$

Le terme $\{Y_n - \mathbb{E}_\theta(Y_n|C_n)\}$ est intégré à la fluctuation stochastique et donc ne nous importe point (cf. annexes B.2, B.3.1, C.2 et C.3.1). Le terme $\{\mathbb{E}_\theta(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\}$, apparaît parce que le résidu est calculé à partir du modèle de base ; il est linéarisé en

$$A_{C_n}(\hat{\lambda}_N)^T(\theta_{C_n} - \bar{\theta}). \tag{9.3}$$

Le terme $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ apparaît parce que les paramètres du modèle de base ont été estimés pour pouvoir calculer le résidu ; il est linéarisé en

$$-A_{C_n}(\hat{\lambda}_N)^T F(\lambda^*, \bar{\theta})^{-1} \sum_{i=1}^{I} F_i(\lambda^*, \bar{\theta})(\theta_i - \bar{\theta})\nu_i. \tag{9.4}$$

Les restes de ces linéarisations sont intégrés au biais de la décomposition linéaire (9.1).

Afin de réduire le biais, les deux termes $\{\mathbb{E}_\theta(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\}$ et $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ pourraient ne pas être linéarisés.
La linéarisation (9.3) est effectuée principalement parce que le modèle que l'on considère est général (équation (7.1)). Mais elle peut être évitée dans des cas particuliers. Par exemple, si $Y_n|C_n \sim \text{Exponentielle}\{\exp(\alpha_{C_n} + \beta x_{C_n}\}$, on peut écrire l'expression exacte de $\{\mathbb{E}_\theta(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\}$ en fonction de $\theta_{C_n} = (\alpha_{C_n}, \beta)^T$ et de $\bar{\theta} = (\bar{\alpha}, \beta)^T$:

$$\mathbb{E}_\theta(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) = \{\exp(\alpha_{C_n}) - \exp(\bar{\alpha})\}\exp(\beta x_{C_n}).$$

La linéarisation (9.4) provient de l'expression du biais $\bar{\theta} - \hat{\lambda}_N$ qui est obtenue à partir des équations du maximum de vraisemblance écrites sous le modèle hiérarchique et sous le

modèle de base. Dans ce cas, obtenir l'expression exacte de $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ en fonction des $\theta_i - \bar{\theta}$, $i = 1, \ldots, I$, semble difficile. Toutefois, on peut aller plus loin que l'ordre 1 dans le développement limité de $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ afin d'obtenir une approximation plus précise.

Ainsi, pour obtenir une décomposition des résidus dans laquelle le biais serait réduit par rapport au biais de la décomposition linéaire (9.1), on peut ne pas linéariser les termes $\{\mathbb{E}_{\theta}(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\}$ et $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ intervenant dans l'expression du résidu ordinaire $r_N(n)$. Ce faisant, l'approximation (9.2) est remplacée par une approximation non linéaire

$$\text{résidus} \;\approx\; \text{fonction non linéaire des effets aléatoires.} \tag{9.5}$$

C'est maintenant ce système d'équations qui doit être résolu pour estimer les effets aléatoires. Or, résoudre un système dont les équations sont non linéaires et nombreuses (autant que d'effets aléatoires) peut ne pas être chose aisée selon le type de non-linéarité.

Reparamétrisation du modèle

Quand on écrit la décomposition linéaire des résidus, les restes de la linéarisation de $\{\mathbb{E}_{\theta}(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\}$ et de $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ sont intégrés au biais. Pour réduire le biais, on peut donc chercher à réduire les restes de la linéarisation par un choix judicieux de la paramétrisation.

Par exemple, considérons le modèle de la section 7.4 où la variable aléatoire Y_n suit, conditionnellement au cluster C_n, une loi exponentielle de moyenne $\alpha_{C_n} \exp\{\beta x_{C_n}\}$ où x_{C_n} est une covariable et β un paramètre. Avec cette paramétrisation

$$\mathbb{E}_{\theta}(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) = (\alpha_{C_n} - \bar{\alpha}) \exp\{\beta x_{C_n}\}$$

alors qu'avec la paramétrisation $\exp\{\alpha_{C_n} + \beta x_{C_n}\}$,

$$\mathbb{E}_{\theta}(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) = \{\exp(\alpha_{C_n}) - \exp(\bar{\alpha})\} \exp\{\beta x_{C_n}\}.$$

Avec la première paramétrisation, le terme $\{\mathbb{E}_{\theta}(Y_n|C_n) - \mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n)\}$ n'a pas à être linéarisé puisque déjà linéaire. En revanche avec la deuxième paramétrisation, il doit l'être.

Trouver une paramétrisation qui évite d'avoir à linéariser le terme $\{\mathbb{E}_{\bar{\theta}^{\otimes I}}(Y_n|C_n) - \mathbb{E}_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}$ est plus difficile puisqu'on ne connaît pas l'expression exacte de ce terme (cf. paragraphe précédent).

Au final, un choix judicieux de la paramétrisation permet de réduire le biais de la décomposition linéaire et donc de rendre moins sensible la décomposition linéaire des résidus à la variabilité des effets aléatoires.

9.3 Restauration des effets aléatoires sous contrainte

L'estimation (ou la restauration) des effets aléatoires passe par la résolution en t du système d'équations

$$\hat{R}_N = \hat{B}t. \tag{9.6}$$

Nous avons résolu ce système en utilisant une inverse généralisée de $\hat{B}$ (cf. sections 7.2.7 et 8.2.7). Cependant nous n'avons pas pris en compte d'éventuelles contraintes (e.g. positivité, nombre fini d'états) auxquelles peuvent être soumis les effets aléatoires. Trouver un moyen de résoudre le système (9.6) sous ces contraintes permettrait d'améliorer la méthode de spécification.

Positivité des effets aléatoires

A la section 7.4 par exemple, les moyennes $\alpha_{C_n} \exp\{\beta x_{C_n}\}$, $n = 1, \ldots, N$, des lois exponentielles doivent être positives et donc les effets aléatoires α_{C_n} doivent l'être également. Notre mode de résolution du système d'équations (9.6) ne permet pas de contraindre les estimateurs des effets aléatoires à être positifs. Utiliser la paramétrisation $\exp\{\alpha_{C_n} + \beta x_{C_n}\}$ par exemple pourrait être une solution puisque dans cette expression les effets aléatoires peuvent être aussi bien négatifs que positifs. Cependant cette paramétrisation pose le problème de non linéarité discuté plus haut. La question de la résolution de (9.6) sous la contrainte de positivité des effets aléatoires reste donc posée.

Nombre fini d'états pour les effets aléatoires

Dans les exemples traités aux chapitres 7 et 8, les effets aléatoires sont à valeurs dans un espace continu. Cependant, les effets aléatoires peuvent prendre un nombre fini de valeurs ou d'états (Besag et al., 1991, les états sont 1 pour 'site anciennement actif' et 0 pour 'site non anciennement actif'). Encore une fois, notre mode de résolution du système d'équations (9.6) ne permet pas de contraindre les estimateurs des effets aléatoires à être dans un espace fini. Une voie peut être envisagée : estimer les effets aléatoires comme nous le faisons déjà, puis les classifier. Le mode de classification dépendra de notre connaissance au sujet du nombre d'états et des valeurs des états. Mais cette solution reste imparfaite.

9.4 Etendre la méthode de spécification à d'autres modèles

9.4.1 Autres modèles hiérarchiques intégrant des effets aléatoires

Une réponse à la question 6.3

> "Peut-on développer une méthode d'analyse de résidus permettant de spécifier le second niveau d'un modèle hiérarchique, c'est-à-dire la distribution de ses effets aléatoires ?"

a été apportée dans les chapitres 7 et 8 pour des modèles hiérarchiques à effets partagés et à effets dépendants mais non partagés. Ces modèles, bien que généraux[1], ne représentent pas toute la diversité des modèles hiérarchiques dont une partie a été exposée dans la section 6.1. Les paragraphes suivants discutent l'extension de notre méthode de spécification à d'autres modèles hiérarchiques que ceux traités dans les chapitres 7 et 8.

[1] Les classes de modèles traitées aux chapitres 7 et 8 contiennent notamment des GLMMs, des GLMMs spatiaux, des modèles de fragilités.

Modèles à effets indépendants et identiquement distribués (i.i.d.)

Notre méthode ne permet pas de spécifier la distribution d'effets i.i.d. En effet, afin de contrôler la fluctuation stochastique des résidus (*cluster residuals* ou *local residuals*), chaque résidu est défini comme une moyenne pondérée des résidus ordinaires (cf. équations (7.6) et (8.7)). La pondération est choisie de manière à ce que chaque résidu prenne une valeur caractéristique du cluster ou du site en lequel il est calculé. Avec des effets aléatoires i.i.d. (absence de covariable structurante), les résidus ordinaires ont tous le même comportement probabiliste. On ne peut donc pas calculer des résidus qui, à la fois, (i) permettent de contrôler la fluctuation stochastique et (ii) prennent des valeurs caractéristiques (différentes).

La méthode développée par Critchley and Marriott (2004) dans le cadre de l'analyse d'influence permet d'estimer certains moments de la distribution d'effets aléatoires i.i.d. Une telle information peut ensuite être utilisée pour spécifier la distribution des effets aléatoires. Notons que la méthode de Critchley and Marriott (2004), contrairement à notre méthode, ne passe pas par la restauration des effets aléatoires. Or, restaurer les effets aléatoires permet (i) l'estimation de leurs moments mais aussi (ii) leur visualisation (sous forme d'histogramme, sous forme de nuage de points 'effets restaurés $\times$ covariable') ; et (i) et (ii) aide, plus que (i) tout seul, aux choix des spécifications possibles pour la distribution des effets aléatoires.

Modèles sur grille

Considérons un modèle hiérarchique sur grille tel que les modèles de Besag et al. (1991), Green and Richardson (2002) et Hrafnkelsson and Cressie (2003) dans lequels sont intégrés des champs aléatoires de Markov cachés (on se place dans le cas où il n'y a qu'une observation par noeud de la grille, ce n'est donc pas le contexte du chapitre 7 où les observations peuvent être faites sur une grille mais où il y a plusieurs données par noeud de la grille). D'un point de vue théorique, la méthode de spécification développée au chapitre 8 ne peut être appliquée à ce type de modèles car les positions des observations doivent être à densité (cf. section C.3.2 et hypothèse (i2)) et donc ne peuvent pas être faites sur une grille. Néanmoins, d'un point de vue pratique, l'étude sur simulations de la section 8.3 semblent indiquer que la méthode fonctionne quand les observations sont faites sur grille. On pourrait envisager de démontrer ce résultat théoriquement.

Processus ponctuels de Cox (processus poissoniens doublement stochastiques)

Les processus ponctuels de Cox (Diggle, 1983; Stoyan et al., 1995) sont des modèles hiérarchiques intégrant un champ aléatoire caché. Dans ces modèles, les points sont aléatoirement et indépendemment distribués dans l'espace conditionnellement à la réalisation d'une intensité aléatoire. L'intensité aléatoire est un champ aléatoire caché défini sur l'espace. Afin de spécifier l'intensité aléatoire, une méthode similaire à celle proposée dans le chapitre 8 pourrait être développée. La principale différence entre la méthode du chapitre 8 et celle qui serait développée pour les processus de Cox est la définition des résidus. En effet,

qu'est-ce qu'un résidu ordinaire pour un processus ponctuel ? Baddeley et al. (2004) apporte des réponses à cette question, réponses qui pourraient être exploitées pour développer une méthode de spécification de l'intensité aléatoire d'un processus de Cox.

Modèles avec covariables individuelles

Dans la méthode de spécification que nous avons proposé, les covariables prennent des valeurs qui dépendent de la covariable structurante : une covariable prend la même valeur pour les individus d'un même cluster (ou situés au même site). Cependant, dans des études telles que celle d'épidémiologie humaine analysée par Henderson et al. (2002), certaines covariables prennent des valeurs différentes pour des individus différents mais appartenant à un même cluster (ou situés dans un même site). On parle de covariables individuelles. Etendre notre méthode à des modèles hiérarchiques contenant des covariables individuelles permettrait, par exemple, de spécifier le second niveau des modèles proposés par Henderson et al. (2002).

Modèles dont les modèles de base associés ne sont pas ajustés par maximum de vraisemblance

Notre méthode de spécification a été développée dans un cadre d'estimation par maximum de vraisemblance : les modèles de base associés aux modèles hiérarchiques sont paramétriques et leurs paramètres sont estimés par maximum de vraisemblance. Nous avons choisi ce cadre parce que les paramètres de tous les modèles de base que nous avons utilisés pouvaient être estimés par maximum de vraisemblance. Cependant, on peut trouver des cas où les paramètres du modèle de base sont estimés en utilisant un autre critère que le maximum de vraisemblance. Par exemple, le modèle de base associé au modèle de fragilités dépendantes de Henderson et al. (2002) est un modèle de Cox (à hasard proportionnel). Les modèles de Cox sont estimés en utilisant une vraisemblance partielle (McCullagh and Nelder, 1989; Box-Steffensmeier and Jones, 2004). Il serait donc judicieux d'étendre notre méthode de spécification à des modèles hiérarchiques dont les modèles de base associés sont estimés en utilisant un contraste (Senoussi, 1990) ou une quasi-vraisemblance (Heyde, 1997) qui englobent, par exemple, la vraisemblance partielle et les moindres carrés.

9.4.2 Modèles déterministes

A la fin de la section 6.2, nous avons posé la question 6.2

> "Pour des modèles relativement complexes, peut-on développer une méthode d'analyse de résidus permettant de fournir de l'information directement exploitable pour la modification du modèle utilisé ?".

Nous avons répondu à cette question pour des modèles (de base) qui sont ajustés à des données et que l'on modifie en ajoutant des effets aléatoires. Mais qu'en est-il pour des modèles déterministes tels que le modèle CHIMERE (section 6.2.2) et les modèles de simulation décrivant la propagation des maladies aériennes des végétaux (section 2.2.2) ? Nous

pensons qu'une méthode d'analyse de résidus permettant de modifier ces modèles peut être développée, mais des différences entre le cadre que nous avons traité et le cadre de ce type de modèles déterministes doivent être prises en compte :

- les paramètres de ces modèles déterministes ne sont pas estimés sur les données que modélisent ces modèles mais sont mesurés expérimentalement, et
- modifier un modèle déterministe ne signifie pas forcément ajouter des effets aléatoires, mais plutôt ajouter ou modifier une équation, ajouter ou modifier des covariables, ajouter ou modifier des conditions initiales ou des conditions aux bornes du domaine d'étude.

Le premier point implique qu'il n'y a pas de biais d'estimation. Par exemple, si l'un des paramètres d'un modèle de propagation est la vitesse du vent nécessaire à la libération des spores, la mesure de ce paramètre sera inchangée que le modèle soit une bonne approximation ou non. En revanche, il peut y avoir des erreurs de mesure qui doivent être prises en compte dans la décomposition des résidus.

Le second point implique que le lien entre résidus et processus aléatoire caché doit être remplacé par un lien entre résidus et équation manquante par exemple. Ce changement est le verrou principal auquel on doit faire face si l'on veut développer une méthode d'analyse de résidus permettant de fournir de l'information directement exploitable pour modifier un modèle de type CHIMERE.

Troisième partie

Conclusion

10

Vers une compréhension multi-échelle de la propagation des maladies aériennes des végétaux

Nous avons présenté deux outils (un cadre multi-échelle de modélisation et une méthode d'analyse de résidus) utiles dans une approche visant à

– détecter des processus impliqués dans la propagation d'une maladie et

– mesurer leurs variabilités et leurs structures

à partir de données expérimentales de propagation obtenues à différentes échelles (cf. question 1.2 de la section 1.1). Le cadre multi-échelle de modélisation permet de décrire des expériences de propagation spatiale à partir d'une source ponctuelle. Ce cadre a été mis en oeuvre aux chapitres 3 et 4 et a été explicité au chapitre 5. La méthode d'analyse de résidus aide à la spécification du second niveau de modèles hiérarchiques. Cette méthode a été appliquée à des données de propagation dans les sections 7.5 et 8.7. Elle a permis de spécifier un modèle hiérarchique décrivant la propagation spatiale à courte distance de la rouille brune du blé et intégrant l'élément 'structuration de l'hétérogénéité des feuilles face à l'infection'. Elle a également permis de proposer des pistes pour un modèle hiérarchique décrivant la propagation spatiale à longue distance de la rouille jaune du blé et intégrant l'élément 'dispersion de nuages de spores'. Dans ce chapitre de conclusion, nous nous proposons de voir comment l'utilisation des deux outils mentionnés ci-dessus peut favoriser la compréhension multi-échelle de la propagation des maladies aériennes des végétaux.

10.1 Le concept d'échelle

Inférer sur un phénomène nécessite de disposer d'un échantillon et d'un outil d'analyse. Partant de ce constat, Dungan et al. (2002) décomposent le concept d'échelle en trois dimensions : le phénomène étudié, l'échantillonnage et l'analyse des données. Décrivons les trois dimensions, qui sont représentées sur la figure 10.1, en prenant pour exemple la propagation des rouilles du blé.

10.1.1 Echelle du phénomène étudié

Axe horizontal sur la figure 10.1.

Le phénomène étudié est caractérisé par des structures physiques et des processus agissant sur ces structures. Pour la propagation des rouilles, les structures physiques sont, entre

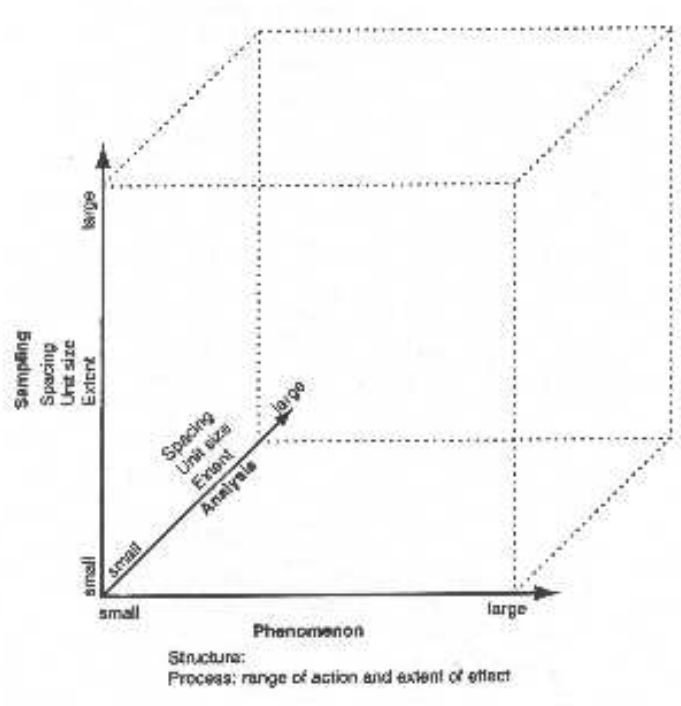


Fig. 10.1. Le concept d'échelles vu sous trois dimensions. Source : Dungan et al. (2002).

autres, la feuille, la plante, la placette (*trap plot*), le champ de blé, la lésion, le groupe de lésions, la spore (ou l'amas de spore), le nuage de spores. Ces structures physiques sont caractérisées par des tailles différentes, par des distances deux à deux différentes. Les processus agissant sur ces structures physiques sont, par exemple, la dispersion par le vent, le frottement entre feuilles, l'infection, la croissance des lésions, les processus qui rendent hétérogènes les feuilles et les placettes. Ces processus sont caractérisés par des distances d'actions différentes.

S'intéresser à une échelle donnée d'un phénomène, c'est s'intéresser au devenir d'un ensemble de structures physiques et aux conséquences d'un ensemble de processus. Ces structures et ces processus sont généralement caractérisés par des distances du même ordre. Par exemple, dans l'étude de la propagation à courte distance de la rouille brune (chapitre 3), les structures physiques d'intérêt sont les feuilles, l'ensemble des spores déposées dans le mètre carré entourant la source ; les processus d'intérêt sont notamment les turbulences locales, les frottements entre feuilles, le *splashing* et l'hétérogénéité des feuilles. Ces structures et ces processus mettent en jeu des distances du même ordre.

Changer d'échelle, c'est changer à la fois de structures d'intérêt et de processus. Ainsi, quand on passe de la propagation sur de courtes distances (chapitre 3) à la propagation sur de longues distances (chapitre 4), l'intérêt est porté non plus sur les frottements entre feuilles par exemple, mais sur l'effet du vent qui rend anisotrope la dispersion.

10.1.2 Echelle de l'échantillonnage

Axe vertical sur la figure 10.1.

L'unité d'échantillonnage, la distance entre unités et le type de mesure définissent l'échelle de l'échantillonnage. Par exemple, dans l'étude de la propagation à longue distance de la rouille jaune (chapitre 4), l'unité d'échantillonnage est la placette d'un mètre carré (*trap plot*), les placettes voisines sont distantes d'une vingtaine de mètres, et la mesure est la quantité de feuilles malades dans la placette.

Notons qu'il peut y avoir des niveaux de hiérarchie dans l'échantillonnage réalisé pour une même expérience. Par exemple, dans l'étude de la propagation à courte distance de la rouille brune (chapitre 3), les feuilles sont les unités d'échantillonnage pour la mesure nombre de lésions, mais les feuilles sont localisées dans d'autres unités d'échantillonnage que sont les quadrats.

10.1.3 Echelle de l'analyse statistique

Axe oblique sur la figure 10.1.

Supposons que l'outil d'analyse est un modèle décrivant le comportement d'une variable d'intérêt. C'est principalement la variable modélisée qui définit l'échelle de l'analyse statistique. Par exemple, dans le modèle de fragilités du chapitre 3, le nombre de lésions par feuille est modélisé. Dans le modèle paramétrique du chapitre 4, le nombre de feuilles malades par mètre carré est modélisé.

L'échelle d'analyse ne correspond pas systématiquement à l'échelle d'échantillonnage. Par exemple, pour évaluer la propagation spatiale d'une maladie des plantes, les mesures sont généralement faites sur feuilles puis elles sont agrégées localement pour ajuster un gradient de maladie et avoir peu de variabilité autour de ce gradient. Cet usage entraîne une perte d'information (cf. chapitre 3 et annexe A.4). Le cadre de modélisation proposé à la section 5.1 permet de dériver des modèles adaptés à l'unité d'échantillonnage et à la mesure de la maladie. Ainsi on peut faire correspondre l'échelle d'échantillonnage et l'échelle d'analyse (c'est pourquoi nous parlons dans la suite de couple d'échelles d'échantillonnage et d'analyse).

10.2 Information et échelles

Le concept d'échelle défini ci-dessus n'est pas analogue au concept d'échelle associé à une carte routière ou à une maquette d'avion : il ne se réduit pas à un chiffre (e.g. échelle au 1/200 000). En fait, il est défini d'un point de vue relativiste, point de vue selon lequel l'espace est perçu à travers des structures et des processus (Meentemeyer, 1989). Le concept d'échelle ainsi défini perd en concision, mais gagne en richesse car il permet d'expliciter ce que changer d'échelle signifie. Expliciter ce que changer d'échelle signifie est important car c'est en changeant d'échelle que nous visons une meilleure compréhension de la propagation.

Dungan et al. (2002) expliquent que *"changing observation and analysis scales influence inference" about the studied phenomenon*. Cependant, on ne peux pas Ci-dessous, nous étudions ce que changer les échelles d'échantillonnage et d'analyse peut impliquer : une dégradation de la quantité d'information (section 10.2.1) et un changement de la nature de l'information (section 10.2.2).

10.2.1 Dégradation de l'information sur le potentiel infectieux en fonction des échelles d'échantillonnage et d'analyse

La présente section illustre par un exemple simple comment l'information peut être dégradée par un changement des échelles d'échantillonnage et d'analyse

Contexte

Considérons une feuille soumise à un potentiel infectieux s et caractérisée par une fragilité Z (observée ou non). Pour inférer sur le paramètre s, deux types de mesure de la maladie sont envisagés : le nombre N de lésions sur la feuille ou l'absence/présence de lésions sur la feuille $Y = \delta_{N>0}$ ($Y = 1$ si présence de lésions, 0 sinon). Selon le cadre de modélisation proposé à la section 5.1, les distributions conditionnelles de N et Y sachant la fragilité Z sont

$$N|Z \sim \text{Poisson}(Zs)$$
$$Y|Z \sim \text{Bernouilli}(1 - e^{-Zs})$$

Ainsi, afin d'inférer sur s, deux couples d'échelles d'échantillonnage et d'analyse sont envisagés :

- on mesure et modélise N, ou bien
- on mesure et modélise Y.

Coût de l'échantillonnage et précision de l'information

Le coût de l'échantillonnage et la précision de l'information apportée par les données varient selon le couple d'échelles d'échantillonnage et d'analyse choisi : la mesure N est plus informative sur s mais aussi plus chère à acquérir. Les choix concernant l'échantillonnage pour une expérience de propagation doivent donc être faits en arbitrant entre coût et précision. Dans la suite, nous ne nous intéressons pas au coût de l'échantillonnage mais étudions seulement comment se dégrade l'information quand on mesure et modélise Y plutot que N.

Information de Fisher

L'information de Fisher permet de quantifier la précision de l'information apportée par une variable donnée sur un paramètre donné ; elle est reliée à la variance d'estimation du paramètre (Dacunha-Castelle and Duflo, 1982; Serfling, 2002). Une information de Fisher élevée reflète une précision élevée. Nous quantifions donc l'information apportée par N ou Y

sur le paramètre s en utilisant l'information de Fisher conditionnelle à Z. Les informations de Fisher conditionnelles apportées sur s par N et Y sont, sachant $Z = z$

$$I_{N|Z}(s;z) = \frac{z}{s}$$
$$I_{Y|Z}(s;z) = \frac{z^2}{e^{-zs}-1}.$$

Comparaison entre les informations apportées sur s par N et Y

Notons $R(s,z)$ le ratio entre l'information de Fisher apportée sur s par Y et l'information de Fisher apportée sur s par N :

$$R(s,z) = \frac{I_{Y|Z}(s;z)}{I_{N|Z}(s;z)} = \frac{zs}{e^{-zs}-1}.$$

Le ratio $R(s,z)$ est une fonction du produit zs qui est l'espérance du nombre de lésions pour la feuille considérée. La figure 10.2 montre l'évolution du ratio $R(s,z)$ en fonction de zs.

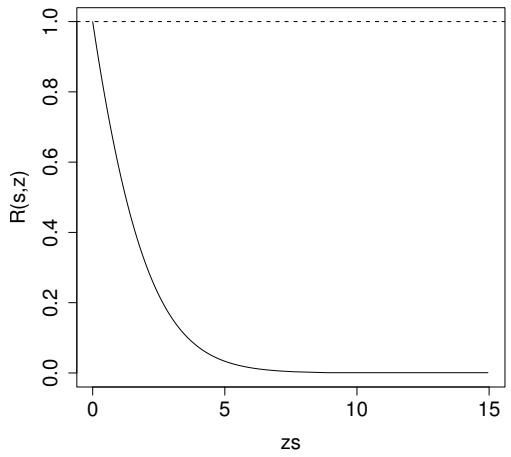


Fig. 10.2. Ratio des informations de Fisher en fonction de l'espérance du nombre de lésions.

La quantité d'information apportée par N est toujours plus grande que celle apportée par Y puisque $R(s,z) \leq 1$. De plus, quand le nombre espéré de lésions zs tend vers 0, alors les informations apportées par N et Y tendent à être équivalentes puisque $\lim_{zs\to 0} R(s,z) = 1$. Ceci s'explique comme suit : quand $zs \to 0$,

$$\begin{aligned}
\mathbb{P}_s(N=0|Z=z) &= \mathbb{P}_s(Y=0|Z=z)\\
\mathbb{P}_s(N=1|Z=z) &= \mathbb{P}_s(Y=1|Z=z)\{1+o(1)\}\\
\mathbb{P}_s(N>1|Z=z) &= o\{\mathbb{P}_s(Y=0|Z=z)\}\\
&= o\{\mathbb{P}_s(Y=1|Z=z)\},
\end{aligned}$$

c'est-à-dire N et Y ont asymptotiquement la même distribution et donc apportent la même information. Quand zs croît, la quantité d'information apportée par Y devient négligeable par rapport à la quantité d'information apportée par N puisque $\lim_{zs\to\infty} R(s,z) = 0$. Ceci s'explique comme suit :

$$\lim_{zs\to\infty} \mathbb{P}_s(Y = 1|Z = z) = 1,$$

c'est-à-dire Y vaut presque-sûrement 1 et donc n'apporte aucune information.

Bilan et perspectives de l'étude sur l'information de Fisher

Cet exemple simple montre qu'échantillonner et modéliser Y plutôt que N conduit à une dégradation de l'information de Fisher plus ou moins sensible selon le nombre espéré de lésions. Cette dégradation a pour conséquence une diminution de la précision de l'inférence. Cela confirme le propos de Dungan et al. (2002) qui expliquent que changer d'échelles d'échantillonnage et d'analyse modifie l'inférence faite sur le phénomène étudié. Ce constat nous amène à donner des pistes en ce qui concerne la manière de concevoir les expériences de propagation.

La manière de concevoir les expériences de propagation pourrait être améliorée en étudiant l'information de Fisher en fonction des échelles d'échantillonnage et d'analyse. Par exemple, si l'on veut estimer les paramètres de la fonction potentiel infectieux (chapitres 3 et 4), on pourrait évaluer l'information de Fisher apportée sur les paramètres pour différents schémas d'échantillonnage et pour différents modèles intégrant le potentiel infectieux. Autre exemple, si l'on s'intéresse aux paramètres liés à l'hétérogénéité des feuilles face à l'infection, on pourrait évaluer l'information de Fisher apportée sur les paramètres pour différents schémas d'échantillonnage et pour différents modèles intégrant l'hétérogénéité des feuilles[1]. Ainsi, le plan d'expérience pourrait être choisi en tenant compte du coût de l'échantillonnage mais aussi de l'information apportée sur les paramètres d'intérêt par les données issues des expériences.

10.2.2 Changement de la nature de l'information en fonction des échelles d'échantillonnage et d'analyse

La section précédente montre comment l'information sur un paramètre peut être dégradée quand on utilise un couple d'échelles d'échantillonnage et d'analyse "grossier". Cependant, quand on change d'échelle, on ne s'intéresse pas forcément au même paramètre ou au même élément du phénomène étudié. La présente section montre, au regard des analyses menées dans les chapitres 3 et 4 et les sections 7.5 et 8.7, comment la nature de l'information sur la dispersion des spores et sur l'hétérogénéité des plantes face à l'infection évolue lorsque les échelles d'échantillonnage et d'analyse changent. Cette section est également l'occasion de récapituler ce qu'apporte mon travail sur le plan de l'épidémiologie végétale.

[1] La section 5.2.2 a évoqué la dégradation de l'information sur l'hétérogénéité des feuilles quand on passe des échelles d'échantillonnage et d'analyse du chapitre 3 (propagation à courte distance) à celles du chapitre 4 (propagation à longue distance).

Expériences de propagation spatiale à courte distance

A partir des expériences de propagation à courte distance et à l'aide d'une analyse combinant modèles et résidus (chapitre 3 et section 7.5), nous avons étudié les éléments suivants :

- Le potentiel infectieux entre 0 et 50 centimètres. Cet élément à été modélisé par une fonction déterministe combinant un paramètre de force de la source de spores et une fonction de dispersion des spores (section 3.3.1), et a été estimé en prenant en compte la sur-dispersion des données.
- Les mécanismes de dispersion à très courte distance tels que l'impaction et le *splashing* dûs à la pluie (mécanismes connus), et le frottement entre feuilles (mécanisme moins bien connu). Ces mécanismes pourraient expliquer la présence de feuilles contenant beaucoup de lésions à proximité de la source de spores (section 3.6.2 et annexe A.2).
- L'hétérogénéité des feuilles en terme de propensions à être infectées. Cet élément étant non observé, il a été pris en compte dans la modélisation en introduisant des fragilités (effets aléatoires) caractérisant les feuilles (section 3.3.2). Ceci a permis de modéliser une grande part de la variabilité des données (section 3.6.1) et d'obtenir des estimateurs plus précis des paramètres de la fonction de dispersion (annexe A.4).
- La structuration spatiale de l'hétérogénéité des feuilles (due par exemple aux variations spatiales de la quantité de nutriments), et les perturbations locales du potentiel infectieux (dues par exemple à la structure 3D du couvert végétal). Ces éléments[2] ont été pris en compte dans la modélisation en introduisant des effets aléatoires partagés par les feuilles situées dans des zones de quelques décimètres carrés (les quadrats). La méthode d'analyse de résidus a indiqué que la variabilité des données autour du potentiel infectieux est structurée spatialement.

Expériences de propagation spatiale à longue distance

A partir des expériences de propagation à longue distance et à l'aide d'une analyse combinant modèles et résidus (chapitre 4 et section 8.7), nous avons étudié les éléments suivants :

- Le potentiel infectieux entre 10 et 200 mètres environ. Cet élément a été modélisé par une fonction déterministe combinant un paramètre de force de la source de spores et une fonction de dispersion des spores (section 4.2.4).
- L'anisotropie de la dispersion des spores en terme de direction et en terme de distance. Cet élément a été modélisé par deux fonctions d'anisotropie : la *directional density function* et la *mean distance function* (chapitre 4). Ceci a permis de modéliser la variabilité des données quand la direction change.
- Les perturbations du potentiel infectieux dues certainement à la dispersion de nuages de spores. Cet élément a été pris en compte dans la modélisation en introduisant un

[2] La structuration de l'hétérogénéité des feuilles et les perturbations du potentiel infectieux sont des éléments confondus au regard des données dont nous disposons.

champ aléatoire caché (effets aléatoires corrélés spatialement) qui déforme le potentiel infectieux (section 8.7). La méthode d'analyse de résidus a indiqué que la variabilité des données autour du potentiel infectieux est structurée spatialement.

La mise en parallèle de ces études menées à deux échelles différentes montre que changer d'échelles d'échantillonnage et d'analyse ne consiste pas seulement à dégrader ou améliorer l'information, mais consiste également à changer la nature de l'information. Par exemple, alors qu'à la courte distance on s'intéresse aux perturbations du potentiel infectieux dues certainement à la structure 3D du couvert végétal, à la longue distance, on s'intéresse plutôt aux perturbations du potentiel infectieux dues à la dispersion de nuages de spores. Disposer d'un cadre de modélisation (section 5.1), qui permet de construire des modèles ayant des composantes communes et des composantes différentes, aide à analyser ce qui ne change pas et ce qui change entre échelles.

10.3 Perspective 1 : estimer la fonction de dispersion des spores de 0 à 10 km

Dans cette section nous discutons d'un problème, l'estimation de la fonction de dispersion des spores [3] sur une large gamme de distances, que notre cadre multi-échelle de modélisation et notre méthode de spécification des processus cachés pourraient aider à résoudre.

10.3.1 Pourquoi estimer la fonction de dispersion des spores de 0 à 10 km ?

L'estimation de la fonction de dispersion des spores sur une large gamme de distances doit permettre d'évaluer et comparer les rôles, dans la dynamique épidémique, (i) de la propagation de proche en proche et (ii) des événements de dispersion à longue distance. La propagation de proche en proche représente le développement d'un foyer primaire de maladie. Les évènements de dispersion à longue distances sont à l'origine du développement des foyers secondaires. Comparer (i) et (ii) revient à comparer d'une part la quantité des spores qui restent à proximité de la source et d'autre part la quantité des spores qui partent loin, c'est-à-dire à évaluer la lourdeur de la queue de dispersion. Comparer (i) et (ii) doit donc permettre de quantifier le potentiel d'une maladie à générer des foyers secondaires et de quantifier les risques de contamination à différentes distances.

Dans l'optique de la lutte contre la maladie, quantifier les risques de contamination à différentes distances est essentiel pour évaluer l'efficacité d'une méthode de lutte telle que la gestion multi-parcellaire des résistances (cf. section 2.1.4). Plus exactement, connaître les risques de contamination à différentes distances doit permettre de déterminer quelle répartition spatiale des résistances dans une région agricole de plusieurs dizaines

[3] La fonction de dispersion est la densité de probabilité de la variable 'position de dépôt d'une spore libérée en l'origine' (cf. section 2.2.2).

de kilomètres carrés est optimale, sachant que l'objectif est de ralentir la maladie. Cette répartition optimale pourrait être déterminée en utilisant un modèle de simulation des épidémies qui intègrerait la fonction de dispersion des spores sur une large gamme de distances (de 0 à 10 km).

10.3.2 Utiliser des expériences réalisées à plusieurs échelles pour estimer la fonction de dispersion de 0 à 10 km

Une expérience unique fournissant les données nécessaires à l'estimation de la fonction de dispersion entre 0 cm et 10 km n'est pas réalisable. En effet, le chapitre 4 montre que pour avoir des plantes malades à seulement 200 m de la source de spores, cette dernière doit être puissante[4]. Mais avec une source puissante, aucune différence dans l'infection des plantes les plus proches de la source n'est décelable, et donc la fonction de dispersion ne peut être estimée à proximité de la source.

En réalisant des expériences à différentes échelles, on doit pouvoir estimer la fonction de dispersion des spores de 0 à 10 km. L'idée est que chaque expérience contient de l'information sur la fonction de dispersion dans une gamme de distances donnée. En combinant les informations apportées sur des gammes de distances successives, on doit pouvoir obtenir une estimation globale de la fonction de dispersion.

Cependant, chaque expérience est caractérisée par des processus qui sont propres à l'échelle considérée (hétérogénéité des feuilles, anisotropie de la dispersion, hétérogénéité des résistances, relief). Ces processus peuvent être vus comme des processus de nuisance qui doivent être intégrés à la modélisation dans le but de filtrer l'information sur la fonction de dispersion. Le cadre de modélisation que l'on a proposé doit permettre d'intégrer ces processus de nuisance. La méthode de spécification que l'on a proposée doit permettre de sépcifier ces processus s'ils sont non observés.

La section suivante montre comment les données issues d'expériences réalisées à différentes échelles pourraient être combinées dans une sorte de méta-analyse visant à estimer la fonction de dispersion des spores sur une large gamme de distances.

10.3.3 Principe de la méta-analyse combinant les expériences réalisées à différentes échelles

Nous explicitons ici le principe d'une méta-analyse visant à estimer la fonction de dispersion des spores sur une large gamme de distances.

[4] Notons que réaliser une expérience où l'on observerait des lésions filles dues à des spores émises par une source située à une distance de l'ordre de 5 ou 10 km semble difficile. En effet, pour espérer trouver des lésions filles formées après un cycle épidémique, il faudrait une source démeusurément forte (car la dispersion à 5 ou 10 km est un évènement rare). L'objet 'lésion' est un objet trop petit pour être détecté. En revanche, on peut penser que l'objet 'foyer secondaire qui se serait développer durant quelques cycles épidémiques' est un objet plus facilement détectable. Toutefois, des difficultés persistent. En particulier, comment différencier les foyers secondaires des foyers tertiaires ? Développer un modèle spatio-temporel pourrait permettre de contourner cette difficulté. La section 10.4 donne une piste pour développer un tel modèle

- Des données issues d'expériences de propagation à différentes échelles pourraient être récoltées.
- Pour chaque expérience (et donc chaque jeu de données), un modèle pourrait être construit à partir d'un cadre de modélisation englobant celui de la section 5.1 ; le modèle intègrerait
  - la fonction de dispersion,
  - des fonctions de cofacteurs observés (spécifiques ou non de l'échelle considérée),
  - des processus aléatoires cachés (spécifiques ou non de l'échelle considérée) reflétant la variabilité non expliquée par les cofacteurs. Les processus cachés pourraient être spécifiés en utilisant notre méthode d'analyse de résidus.
- Un méta-modèle, réunion des modèles construits pour toutes les expériences, serait alors obtenu. A priori, il y aurait un jeu de paramètres par modèle. Afin de réduire le nombre de paramètres, ceux-ci pourraient être vus comme des paramètres aléatoires distribués selon des lois qui pourraient être spécifiées en utilisant notre méthode d'analyse de résidus. En ajustant ce méta-modèle aux données, on obtiendrait une estimation de la fonction de dispersion sur la gamme de distance couverte par l'ensemble des expériences.

Un exemple schématique d'une telle méta-analyse est donné dans l'annexe D.

10.4 Perspective 2 : analyser la variabilité dans les expériences de propagation spatio-temporelle

Dans cette section nous discutons d'un problème, l'analyse de la variabilité dans les expériences de propagation spatio-temporelle d'une maladie aérienne, que notre méthode de spécification des processus cachés pourrait aider à résoudre.

10.4.1 S'intéresser à de nouveaux éléments de la propagation en analysant des expériences de propagation spatio-temporelle

Les expériences de propagation uniquement spatiale permettent d'inférer sur des éléments causant de la variabilité au niveau de la dispersion des spores et au niveau de l'infection. Mais ces expériences ne permettent pas d'inférer sur des éléments causant de la variabilité au niveau de la production des spores. Elles ne permettent pas non plus d'évaluer l'évolution dans le temps des éléments tels que l'anisotropie de la dispersion ou encore l'hétérogénéité des feuilles. Des questions posées au sujet de ces éléments peuvent trouver leurs réponses dans l'analyse de données obtenues à partir d'expériences de propagation dans l'espace et le temps (cf. figure 1.1 : propagation de la rouille jaune du blé dans une parcelle expérimentale durant plusieurs semaines). En effet, dans ce type d'expériences, il y a plusieurs cycles production-dispersion-infection et, par conséquent, (i) le nombre de sources de spores est important et une analyse statistique de la production des spores est possible, et (ii) les éléments tels que l'anisotropie de la dispersion et l'hétérogénéité des feuilles peuvent être analysés sur différentes périodes.

10.4.2 Utilisation de la méthode de spécification pour construire un modèle spatio-temporel

L'information contenue dans une expérience de propagation spatio-temporelle est plus riche que celle contenue dans l'expérience de propagation spatiale qui se limiterait au premier cycle épidémique de l'expérience spatio-temporelle. Cette richesse d'information doit permettre d'inférer sur de nouveaux éléments comme mentionné précédemment. Mais, à cause du plus grand nombre de processus impliqués, la construction du modèle permettant d'analyser les données est plus complexe. Par exemple, l'hétérogénéité des feuilles en terme de propensions à être infectées n'est plus seulement un processus spatial mais un processus spatio-temporel (une feuille de blé qui dépérit est de moins en moins réceptive aux rouilles).

Mais, face aux données issues de l'expérience de propagation spatio-temporelle illustrée par la figure 1.1, quels éléments introduire dans le modèle ? Comment spécifier les composantes du modèle, en particulier celles qui correspondent à des processus non observés ?

– Partir d'un modèle de base relativement simple intégrant les éléments prépondérants de la propagation spatio-temporelle,
– et le complexifier petit-à-petit en utilisant notre méthode d'analyse des résidus

devraient permettre d'identifier, modéliser et quantifier les éléments jouant un rôle important dans la propagation spatio-temporelle à l'échelle considérée. Ainsi, l'approche pas-à-pas que nous avons adoptée pour étudier la propagation spatiale pourrait être également adoptée pour étudier la propagation dans l'espace et le temps d'une maladie aérienne.

Quatrième partie

Annexes

A

Annexe du chapitre 3

A.1 Consistency and asymptotic normality : sketch of proof

Preliminaries.

We provide assumptions under which inequalities used to prove consistency and asymptotic normality are satisfied. We use notations of Sections 3.3 and 3.4. Moreover, let ∂_u and ∂_{uv}, $1 \leq u, v \leq 4$, denote the partial derivative operators of first and second order with respect to components of θ. Let i be in $\{1, \ldots, I\}$. $\theta \mapsto p_\theta^{ij}(n)$, $j = 1, \ldots, J_i$ and $n \in \mathbb{N}$, is infinitely differentiable on $\Theta = \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \{(s,t) \in \mathbb{R}^2 : \Delta(s,t) > 0\}$.
Assume $\theta \in \Theta_0 = [\underline{a}, \bar{a}] \times [\underline{b}, \bar{b}] \times \{(s,t) \in \mathbb{R}^2 : \Delta(s,t) \geq 0\} \subset \mathbb{R}^4$, where $0 < \underline{a} < \bar{a} < \infty$ and $0 < \underline{b} < \bar{b} < \infty$. Assume there exists $0 < r < \infty$ such that for all j in $\{1, \ldots, J_i\}$, $A_{ij} \subset \mathscr{B}(0, r)$ (fixed-domain asymptotic). Then, there exist constants $0 < \beta_1, \beta_2, \gamma_1, \gamma_2 < \infty$ depending on r such that for all $\theta \in \Theta$, $j \in \{1, \ldots, J_i\}$, $n \in \mathbb{N}$ and $(u, v) \in \{1, \ldots, 4\}^2$,

$$\beta_1 \frac{\gamma_1^n}{n!} \leq \ p_\theta^{ij}(n) \leq \frac{\bar{a}^n}{n!} \tag{A.1}$$

$$\frac{|\partial_u p_\theta^{ij}(n)|}{p_\theta^{ij}(n)} \leq \beta_2 (1+n) \gamma_2^n \tag{A.2}$$

$$\frac{|\partial_{uv} p_\theta^{ij}(n)|}{p_\theta^{ij}(n)} \leq \beta_2 (1 + 4n + n^2) \gamma_2^n. \tag{A.3}$$

In the following, we set $\Theta_\delta = [\underline{a}, \bar{a}] \times [\underline{b}, \bar{b}] \times \{(s,t) \in \mathbb{R}^2 : \Delta(s,t) \geq \delta\}$ where $0 \leq \delta \leq 4 = \max\{\Delta(s,t) : (s,t) \in \mathbb{R}^2\}$.

Consistency of $\hat{\theta}_i$.

Here, we only give elements allowing to satisfy the assumptions of a Hoadley's theorem (Hoadley, 1971). First, the model is identifiable on Θ_δ for all $\delta \in]0, 4]$. Second, let $\theta_0 \in \Theta_\delta$, $0 < \delta \leq 4$, and $R_{ijk}(\theta)$ equals $\log\{p_\theta^{ij}(N_{ijk}) / p_{\theta_0}^{ij}(N_{ijk})\}$ if $p_{\theta_0}^{ij}(N_{ijk}) > 0$, 0 otherwise. Because of inequation (A.1), there exists $0 < R < \infty$ such that $R_{ijk}(\theta) \leq R$, and $\mathbb{E}_{\theta_0}\{\sup_{\theta \in \Theta_\delta} R_{ijk}(\theta)^2\} < R^2 < \infty$. This proves that $\hat{\theta}_i$ converges in probability p_{θ_0} to θ_0.

Asymptotic normality of $\hat{\theta}_i$.

We now use the theorem of Philippou and Roussas (1973). Assume that θ is in the interior of Θ_δ and $0 < \delta \leq 4$. The support of p_θ^{ij} does not depend on θ, and $\hat{\theta}_i$ is consistent. Assumptions on the derivatives of the log-likelihood have to be satisfied. Set $\phi_u^{ijk}(\theta) = \partial_u \log p_\theta^{ij}(N_{ijk})$ and $\psi_{uv}^{ijk}(\theta) = \partial_{uv} \log p_\theta^{ij}(N_{ijk})$. For all $\theta \in \Theta$, $j \in \{1, \ldots, J_i\}$, $k \in \{1, \ldots, K_{ij}\}$ and $(u, v) \in \{1, \ldots, 4\}^2$,

$$\mathbb{E}_\theta\{|\phi_u^{ijk}|^3\} \leq \beta_2^3 \sum_{n=0}^{\infty} (1+n)^3 \frac{(\gamma_2^3 \bar{a})^n}{n!} = M_1 < \infty \tag{A.4}$$

$$\mathbb{V}_\theta\{\psi_{uv}^{ijk}\} \leq \beta_2^2 \sum_{n=0}^{\infty} (1+4n+n^2)^2 \frac{(\gamma_2^2 \bar{a})^n}{n!} = M_2 < \infty, \tag{A.5}$$

because of inequations (A.1), (A.2) and (A.3). Inequation (A.5) leads to

$$\lim_{K_i \to \infty} \frac{1}{K_i^2} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \mathbb{V}_\theta\{\psi_{uv}^{ijk}\} = 0. \tag{A.6}$$

Because of (A.4) and (A.6), the assumptions on the log-likelihood are satisfied. Assume that the Fisher information matrix $I_{K_i}(\theta)$ is invertible and $(1/K_i) I_{K_i}(\theta)$ converges to $\mathrm{i}_i(\theta)$ when $K_i \to \infty$. Then, under θ, the asymptotic distribution of $\sqrt{K_i}(\hat{\theta}_i - \theta)$ is the normal distribution $\mathcal{N}(0, \mathrm{i}_i(\theta)^{-1})$. We use $(1/K_i) I_{K_i}(\hat{\theta}_i)$ to estimate $\mathrm{i}_i(\theta)$.

A.2 Check of the dispersal function

Let $(i, \mathcal{J})$ denote a set of quadrats in subexperiment i which are symmetric with respect to the spore source location of subexperiment i. Let $\mu_{i\mathcal{J}}$ and $\sigma_{i\mathcal{J}}$ respectively denote the sample mean and the sample standard deviation of the lesion count for the leaves in the quadrats belonging to $(i, \mathcal{J})$

$$\mu_{i\mathcal{J}} = \frac{1}{\sum_{j \in \mathcal{J}} K_{ij}} \sum_{j \in \mathcal{J}} \sum_{k=1}^{K_{ij}} N_{ijk}$$

$$\sigma_{i\mathcal{J}} = \left\{ \frac{1}{(\sum_{j \in \mathcal{J}} K_{ij}) - 1} \sum_{j \in \mathcal{J}} \sum_{k=1}^{K_{ij}} (N_{ijk} - \mu_{i\mathcal{J}})^2 \right\}^{1/2}.$$

Their simulated analogues $\mu_{i\mathcal{J}}^*$ and $\sigma_{i\mathcal{J}}^*$ are obtained by replacing N_{ijk} by N_{ijk}^* which is a simulated count of lesions on leaf (i, j, k) under the estimated model. For all $(i, \mathcal{J})$, we compare $\mu_{i\mathcal{J}}$ and $\sigma_{i\mathcal{J}}$ to the distributions of $\mu_{i\mathcal{J}}^*$ and $\sigma_{i\mathcal{J}}^*$. Figure A.2.1 plots the evolution of the sample mean and standard deviation $\mu_{i\mathcal{J}}$ and $\sigma_{i\mathcal{J}}$ (black dots) with distance from the source. In addition, bootstrap 95%-confidence intervals (circles) and simulated 95%-confidence intervals (triangles) for the true means and standard deviations are drawn. The bootstrap procedure is chosen to avoid any distribution assumption and is carried out by

performing 2000 resampling for each $(i, \jmath)$. The simulated intervals are obtained by performing 499 Monte-Carlo simulations under the model. Plots are drawn for subexperiments 2 and 3 (plots of subexp. 4 look like those of subexp. 2, and plots of subexp. 1 and 5 look like those of subexp. 3). Bootstrap confidence intervals and simulated confidence intervals overlap. However, for subexperiments 1, 3 and 5, the sample mean and standard deviation of the quadrats in contact with the source are over the corresponding simulated confidence intervals. Thus, the level of disease is underestimated near the source for subexperiments 1, 3 and 5.

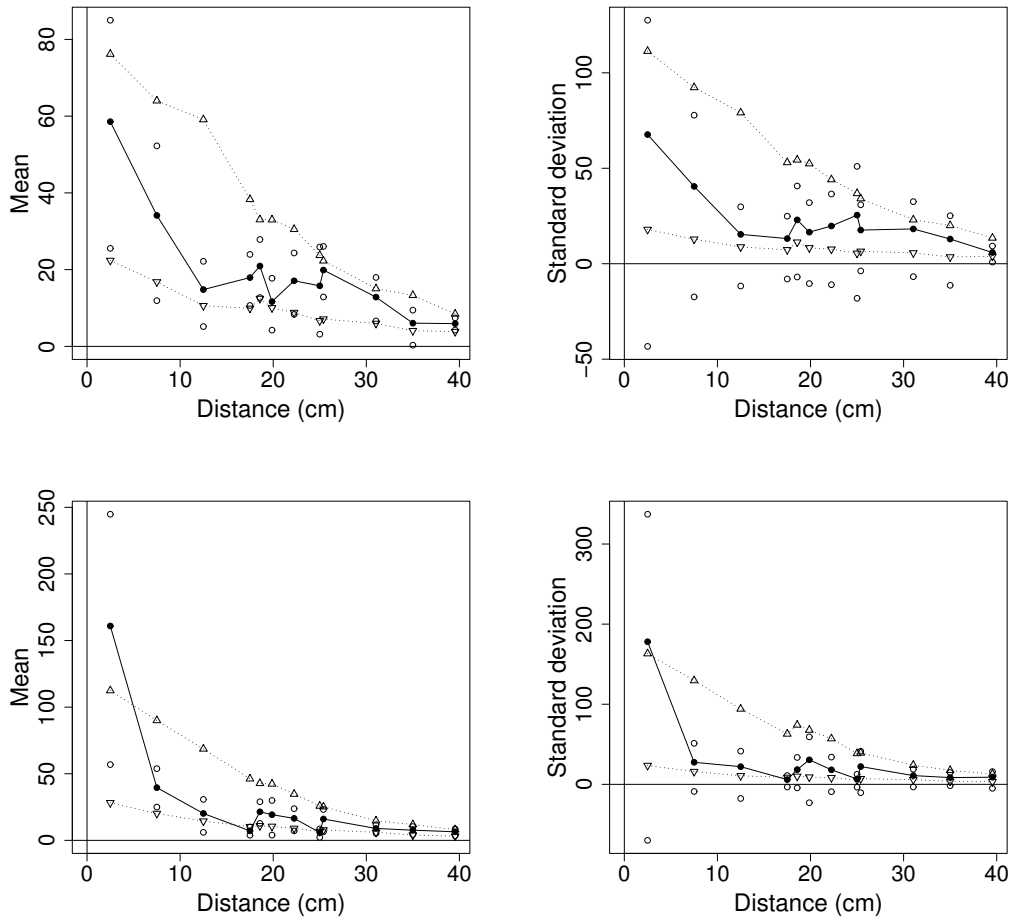


Fig. A.1. Evolution, with distance, of the sample mean (left) and standard deviation (right) of the lesion count per leaf for subexperiments 2 (top) and 3 (bottom). The sample statistics (black dots) are drawn together with bootstrap 95%-confidence intervals (circles) and simulated 95%-confidence intervals (triangles) under the estimated model.

A.3 Expected value and variance of the lesion count

Consider leaves of subexperiment i in $\{1, \ldots, I\}$. Under $\theta = (a, b, c, d)^T$,

$$\mathbb{E}_\theta(N_{ijk}) = \mathbb{E}_\theta(Z_{ijk})\frac{1}{|A_{ij}|}\int_{A_{ij}} S_{ab}(x)dx$$
$$\mathbb{V}_\theta(N_{ijk}) = \mathbb{E}_\theta(N_{ijk}) + \left\{\mathbb{E}_\theta(Z_{ijk}^2)\frac{1}{|A_{ij}|}\int_{A_{ij}} S_{ab}(x)^2dx - \mathbb{E}_\theta(N_{ijk})^2\right\}.$$

where $\mathbb{E}_\theta\{Z_{ijk}\} = \mathbb{E}_{c,d}\{Z_{ijk}\} = \int_0^1 z f_{cd}(z)dz$ is the expected value of the frailty and $\mathbb{E}_\theta\{Z_{ijk}^2\} = \int_0^1 z^2 f_{cd}(z)dz$ is its second moment about the origin.

A.4 Individual data versus aggregated data

In this section we show that using individual data rather than aggregated data allows to more accurately assess disease spread. As mentioned in introduction 3.1, assessing disease spread is commonly done in botanical epidemiology by aggregating data and fitting a gradient curve (Aylor, 1987; Fitt et al., 1987). The 2D-version of the gradient curve is the conditional expectation of the number of lesions N on a leaf, given the leaf location X, that is $\mathbb{E}_\theta(N|X)$. Under our frailty model

$$\begin{aligned}\mathbb{E}_\theta(N|X) &= \mathbb{E}_{c,d}(Z)af_b(X)\\ &= \frac{a\mathbb{E}_{c,d}(Z)}{2\pi b^2}\exp\left(-\frac{||X||}{b}\right)\end{aligned}$$

where $\mathbb{E}_{c,d}(Z)$ is the expected value of the frailty. In this model, $\mathbb{E}_{c,d}(Z)$ and the source strength a are not identifiable and cannot be estimated; rather $a' = a\mathbb{E}_{c,d}(Z)$, thereafter called intercept, is estimated. Consequently, we compared the accuracy of the estimators of a' and b obtained on one hand by the technique based on aggregated data, and on the other by the technique developed in chapter 3. For the latter technique, the estimator of a' is obtained by plug-in the estimators of a, c and d in $a\mathbb{E}_{c,d}(Z)$.

Let us describe the technique of estimation of a' and b based on aggregated data. First, data are aggregated for each quadrat : for quadrat (i,j), $N_{ij1},\ldots,N_{ijK_{ij}}$ are replaced by their sample mean $\bar{N}_{ij} = K_{ij}^{-1}\sum_{k=1}^{K_{ij}} N_{ijk}$ which is affected to the center, say x_{ij}, of the quadrat. Second, the model

$$\mathbb{E}_\theta(N|X) = \frac{a'}{2\pi b^2}\exp\left(-\frac{||X||}{b}\right)$$

is linearized

$$\log\mathbb{E}_\theta(N|X) = \log(a') - \log(2\pi) - 2\log(b) - \frac{||X||}{b},$$

and fitted to aggregated data $\{(x_{ij},\bar{N}_{ij}) : i = 1,\ldots,I, j = 1,\ldots,J_i\}$ with the ordinary least squares criterion.

We simulated 200 subexperiments under the frailty model with parameters equal to the values estimated for subexperiment 5 (see table 3.2). Then, for each simulated subexperiment we computed the estimates of a' and b using both techniques of estimation. Figure A.2 shows the histograms of the estimates obtained for the intercept a' (left) and for

the dispersal parameter b (right) using the technique based on aggregated data (top) and using our technique (bottom). Note that the x-axis scale is the same for both histograms drawn for a' and for both histograms drawn for b. The histograms of the estimates are more narrow with our estimation technique than with the technique based on aggregated data. The histograms for a' and b are centered around the true value using our technique, whereas only the histogram for b is centered around the true value using the technique based on aggregated data. We computed the relative mean square errors (RMSE) for both parameters

$$RMSE(a') = \frac{\sum_{m=1}^{200}(\hat{a}'_m - a'_0)^2}{\sum_{m=1}^{200}(\tilde{a}'_m - a'_0)^2} \approx 0.23$$
$$RMSE(b) = \frac{\sum_{m=1}^{200}(\hat{b}_m - b_0)^2}{\sum_{m=1}^{200}(\tilde{b}_m - b_0)^2} \approx 0.13,$$

where $\hat{a}'_m$ and $\hat{b}_m$ denote estimates of a' and b obtained with our technique applied to simulation m, $\tilde{a}'_m$ and $\tilde{b}_m$ denote estimates of a' and b obtained with the technique based on aggregated data applied to simulation m, and a'_0 and b_0 denote the true value of the parameters. The histograms and the values of the RMSE shows that our estimation technique provide more accurate estimators of the intercept a' and the dispersal parameter b than the technique based on aggregated data usually exploited in botanical epidemiology.

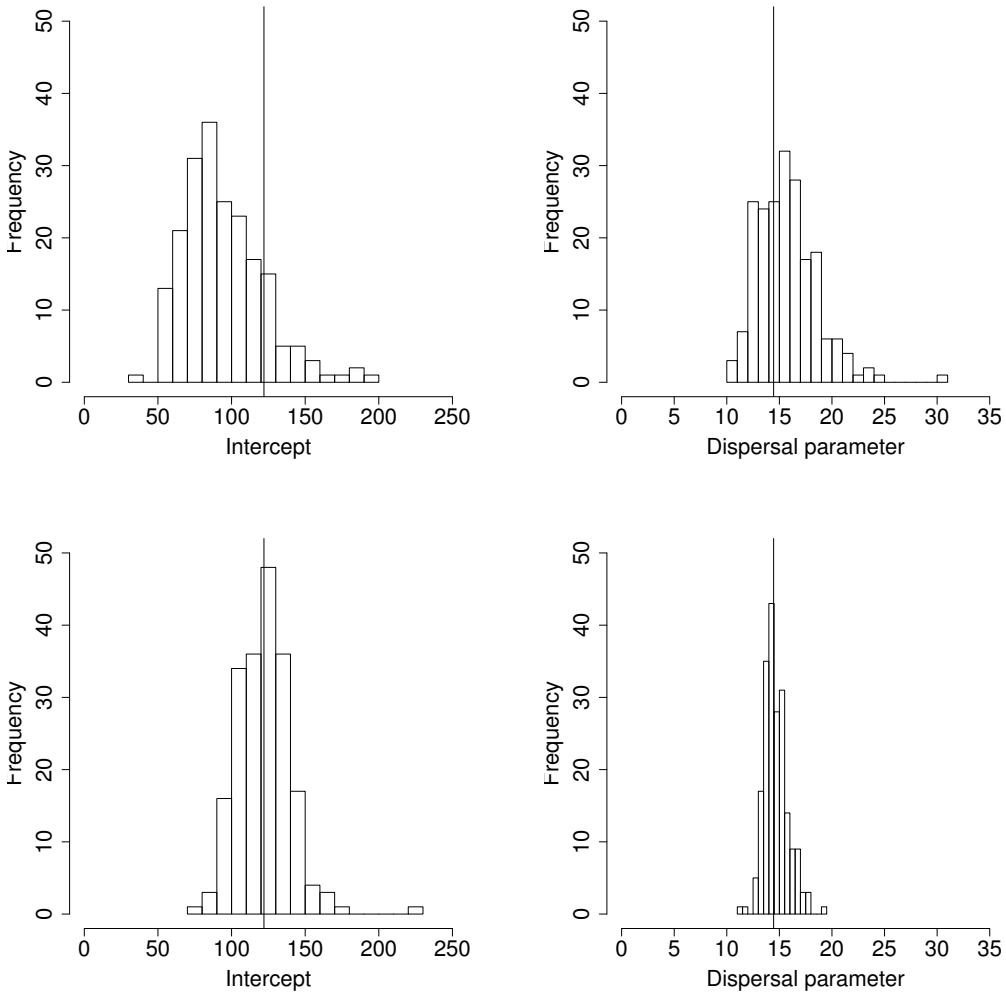


Fig. A.2. Histograms of the estimates obtained for the intercept a' (left) and the dispersal parameter b (right) using the technique based on aggregated data (top) and using our technique (bottom). Vertical lines : true values of parameters a' and b.

B

Annexe du chapitre 7

B.1 Expression of the bias $\bar{\theta} - \hat{\lambda}_N$

In the following, assumptions (B.a)-(B.c) are assumed to be satisfied. Assumptions (B.a)-(B.c) as well as other assumptions mentioned in this section are written in appendix B.4. From the strong law of large numbers and under integrability assumption (B.g2), the $\mathbb{P}_\theta$-a.s.-limit of $l_N(\lambda, \mathbf{Y}, \mathbf{C})$ (equation (7.4)) as $N \to \infty$ is

$$\begin{aligned} l^*(\lambda, \theta) &= \mathbb{E}_\theta[\log \pi(Y_n | C_n, \lambda)] \\ &= \sum_{i=1}^{I} \int_{\mathbb{R}} \log \pi(y|i, \lambda) \mathbb{P}_\theta(i, dy). \end{aligned}$$

Let λ^* be the maximizer of $l^*(\cdot, \theta)$. Define

$$\begin{aligned} l_\varepsilon(\lambda, \theta) &= \mathbb{E}_\theta[\log \pi(Y_n | C_n, \lambda + \theta_{C_n} - \bar{\theta})] \\ &= \sum_{i=1}^{I} \int_{\mathbb{R}} \log \pi\{y|i, \lambda + \theta_i - \bar{\theta}\} \mathbb{P}_\theta(i, dy). \end{aligned}$$

$\bar{\theta}$ is the maximizer of $l_\varepsilon(\cdot, \theta)$. Moreover, from a Taylor expansion and under integrability assumption (B.g3) the limit of $l_\varepsilon(\lambda, \theta)$ as $\varepsilon \to 0$ is $l^*(\lambda, \theta)$.
Remark that λ^* and $\bar{\theta}$ depend only on $\mathbb{P}_\theta$, i.e. on θ, ν, μ and π. As ν, μ and π are fixed, we consider λ^* and $\bar{\theta}$ as functions of θ.

Lemma B.1. *Under concavity assumptions (B.d) and (B.e), assumption (B.f) of uniqueness of λ^*, and integrability assumptions (B.g2)-(B.g5),*

$$\hat{\lambda}_N \xrightarrow[\mathbb{P}_\theta - a.s.]{} \lambda^* \quad as \quad N \to \infty \tag{B.1}$$

$$\bar{\theta} \longrightarrow \lambda^* \quad as \quad \varepsilon \to 0, \tag{B.2}$$

and

$$
\begin{aligned}
F(\lambda^*,\bar{\theta})(\lambda^* - \hat{\lambda}_N) = & \left[\frac{1}{N}\sum_{n=1}^{N} \nabla \log \pi(Y_n|C_n,\lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)\}\right] \\
& + o_{\mathbb{P}_\theta - a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty), \quad as \quad N \to \infty \\
F(\lambda^*,\bar{\theta})(\bar{\theta} - \lambda^*) = & -\sum_{i=1}^{I} F_i(\lambda^*,\bar{\theta})(\theta_i - \bar{\theta})\nu_i \\
& + o(\varepsilon + ||\bar{\theta} - \lambda^*||_\infty), \quad as \quad \varepsilon \to 0,
\end{aligned}
$$

where

$$F_i(\lambda^*,\bar{\theta}) = -\mathbb{E}_{\bar{\theta}^{\otimes I}}\{\mathbf{H} \log \pi(Y_n|C_n,\lambda^*) \mid C_n = i\} \tag{B.3}$$

$$F(\lambda^*,\bar{\theta}) = \sum_{i=1}^{I} F_i(\lambda^*,\bar{\theta})\nu_i \tag{B.4}$$

$$= -\mathbb{E}_{\bar{\theta}^{\otimes I}}\{\mathbf{H} \log \pi(Y_n|C_n,\lambda^*)\}. \tag{B.5}$$

$F_i(\lambda^*,\bar{\theta})$ is a Fisher information matrix at cluster $i \in \Gamma$ and $F(\lambda^*,\bar{\theta})$ is a mean Fisher information matrix.

Sketch of proof Convergences (B.1) and (B.2) can be shown by using a Rockafellar theorem and a corollary both written in Senoussi (1990). The consecutive equations are obtained by using Taylor expansions.

Corollary B.2. *In addition, if $F(\lambda,\lambda')$ is invertible for all λ and λ' in Θ (assumption (B.h)), then as $N \to \infty$ and $\varepsilon \to 0$*

$$\bar{\theta} - \hat{\lambda}_N \xrightarrow[\mathbb{P}_{\mathcal{D}} - a.s.]{} 0$$

and

$$
\begin{aligned}
\bar{\theta} - \hat{\lambda}_N = & -F(\lambda^*,\bar{\theta})^{-1} \sum_{i=1}^{I} F_i(\lambda^*,\bar{\theta})(\theta_i - \bar{\theta})\nu_i \\
& + F(\lambda^*,\bar{\theta})^{-1}\left[\frac{1}{N}\sum_{n=1}^{N} \nabla \log \pi(Y_n|C_n,\lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)\}\right] \\
& + o_{\mathbb{P}_{\mathcal{D}} - a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty) + o(\varepsilon + ||\bar{\theta} - \lambda^*||_\infty).
\end{aligned} \tag{B.6}
$$

B.2 Expression of cluster residual $\hat{R}_{N,i}$

Ordinary residual $r_N(n)$ (equation (7.5)) can be written

$$r_N(n) = \{Y_n - \mathbb{E}_\theta(Y_n|C_n)\} + \{E_\theta(Y_n|C_n) - E_{\hat{\lambda}_N^{\otimes I}}(Y_n|C_n)\}.$$

Suppose assumptions mentioned in appendix B.1 are satisfied. By using Taylor expansions, expression (B.6) of the bias $\bar{\theta} - \hat{\lambda}_N$ and under integrability assumptions (B.g6)-(B.g8),

$$
\begin{aligned}
r_N(n) =& \{Y_n - \mathbb{E}_\theta(Y_n|C_n)\} \\
&+ A_{C_n}(\hat{\lambda}_N)^T \left[(\theta_{C_n} - \bar{\theta}) - F(\lambda^*, \bar{\theta})^{-1} \sum_{i=1}^{I} F_i(\lambda^*, \bar{\theta})(\theta_i - \bar{\theta})\nu_i \right] \\
&+ A_{C_n}(\lambda^*)^T F(\lambda^*, \bar{\theta})^{-1} \left[\frac{1}{N} \sum_{m=1}^{N} \nabla \log \pi(Y_m|C_m, \lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_m|C_m, \lambda^*)\} \right] \\
&+ \xi_{N,\varepsilon,C_n},
\end{aligned}
\tag{B.7}
$$

where $A_i(\lambda) = \int_{\mathbb{R}} y \nabla \pi(y|i, \lambda)\mu(dy)$, $i = 1, \ldots, I$ and $\lambda \in \Theta$, and $\xi_{\varepsilon,N,C_n} = o_{\mathbb{P}_{\mathcal{D}}-a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta} - \lambda^*||_\infty)$ as $N \to \infty$ and $\varepsilon \to 0$.

Let $Q_{N,i} = N^{-1} \sum_{n=1}^{N} \delta_{C_n=i}$ be the proportion of observations made at cluster i, then cluster residual $\hat{R}_{N,i}$ (equation (7.6)) can be written

$$
\hat{R}_{N,i} = \frac{1}{Q_{N,i} N} \sum_{n=1}^{N} \delta_{C_n=i}\ r_N(n).
$$

Then, using equation (B.7) leads to equation (7.7).

B.3 Sketch of proof of theorem 7.1

B.3.1 Decomposition of the vector $\hat{\mathbf{R}}_{\boldsymbol{N}}$ of cluster residuals

Set $\mathbf{Q}_N = (Q_{N,1}, \ldots, Q_{N,I})^T$,

$$
A(\hat{\lambda}_N) = \begin{pmatrix} A_1(\hat{\lambda}_N)^T & & (0) \\ & \ddots & \\ (0) & & A_I(\hat{\lambda}_N)^T \end{pmatrix},
$$

$$
B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta}) = A(\hat{\lambda}_N)[\mathbf{I}_{dI} - \mathbf{1}_I \otimes \{F(\lambda^*, \bar{\theta})^{-1}(F_1(\lambda^*, \bar{\theta})\nu_1, \ldots, F_I(\lambda^*, \bar{\theta}))\nu_I\}],
$$

where $\mathbf{I}_k$ and $\mathbf{1}_k$ are, respectively, the identity matrix and the unit vector of dimension $k \in \mathbb{N}^*$, and $\otimes$ is the Kronecker product also called matrix direct product,

$$
U_n(\mathbf{Q_N}, \theta) = \begin{pmatrix} \delta_{C_n=1} Y_n - Q_{N,1} \mathbb{E}_\theta(Y_n|C_n = 1) \\ \vdots \\ \delta_{C_n=I} Y_n - Q_{N,I} \mathbb{E}_\theta(Y_n|C_n = I) \\ \hline \nabla \log \pi(Y_n|C_n, \lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n, \lambda^*)\} \end{pmatrix},
$$

$$
\Delta_\theta = \begin{pmatrix} \theta_1 - \bar{\theta} \\ \vdots \\ \theta_I - \bar{\theta} \end{pmatrix} \quad \text{and} \quad \Delta_\alpha = \begin{pmatrix} \alpha_1 - \bar{\alpha} \\ \vdots \\ \alpha_I - \bar{\alpha} \end{pmatrix}.
$$

Then, the vector of cluster residuals $\hat{\mathbf{R}}_N = (\hat{R}_{N,1}, \ldots, \hat{R}_{N,I})^T$ can be written

$$\hat{\mathbf{R}}_N = B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\theta + T(\mathbf{Q}_N, \lambda^*, \bar{\theta}) \left\{ \frac{1}{N} \sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta) \right\} + \xi_{\varepsilon,N},$$

where $\xi_{N,\varepsilon} = (\xi_{N,\varepsilon,1}, \ldots, \xi_{N,\varepsilon,I})^T$.

Only components $1, \ldots, d_1, \ldots, (I-1)d+1, \ldots, (I-1)d+d_1$ of Δ_θ can be different from 0; indeed, the others correspond to β. So, keep columns $1, \ldots, d_1, \ldots, (I-1)d+1, \ldots, (I-1)d+d_1$ of $B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ and denote the resulting matrix by $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})$, then $B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\theta = B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha$. Let

$$\psi_{N,\theta} = T(\mathbf{Q}_N, \lambda^*, \bar{\theta}) N^{-1} \sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta). \tag{B.8}$$

We obtain equation (7.8)

$$\hat{\mathbf{R}}_N = B(\hat{\lambda}_N, \lambda^*, \bar{\theta})\Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon}.$$

Note that summing block-columns of $B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ leads to the zero matrix. Therefore, $B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ is not invertible and $B(\hat{\lambda}_N, \lambda^*, \bar{\theta})$ neither.

B.3.2 Limiting distribution of $\psi_{N,\theta}$

To get the limiting distribution of $\psi_{N,\theta}$ given by equation (B.8), we study separately the convergences of $N^{-1}\sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta)$ and $T(\mathbf{Q}_N, \lambda^*, \bar{\theta})$.

Limiting distribution of $N^{-1}\sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta)$

$U_n(\mathbf{Q}_N, \theta)$ is random because of both (C_n, Y_n) and $\mathbf{Q}_N$ which is a function of $C_1, \ldots, C_N$. So, $U_1(\mathbf{Q}_N, \theta), \ldots, U_I(\mathbf{Q}_N, \theta)$ are independent conditionally on $\mathbf{Q}_N$ and identically distributed. By the central limit theorem, we can get the conditional limiting distribution of $N^{-1/2}\sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta)$ given $\mathbf{Q}_N$. But to get its unconditional limiting distribution, we use the following lemma and its corollary.

Lemma B.3. *Let* $S_N = (\sum_{n=1}^{N} G_n) - g(\mathbf{Q}_N)$ *where* $\mathbf{Q}_N$ *is a random vector in* $\mathbb{R}^I$, $I \in \mathbb{N}^*$, $G_1, \ldots, G_N$ *are random vectors in* $\mathbb{R}^J$, $J \in \mathbb{N}^*$, *which are mutually independent conditionally on* $\mathbf{Q}_N$, *and* $g(\cdot)$ *is a function from* $\mathbb{R}^I$ *to* $\mathbb{R}^J$. *Let* η *denote the limiting distribution of* $g(\mathbf{Q}_N)$ *as* $N \to \infty$. *Assume*

$$E\{G_n - N^{-1}g(\mathbf{Q}_N)|\mathbf{Q}_N\} = 0 \tag{B.9}$$

$$V(S_N|\mathbf{Q}_N) = \mathbf{I}_J \tag{B.10}$$

$$\sum_{n=1}^{N} E\{(G_n^T G_n)^{3/2}\} \leq k, \tag{B.11}$$

where k *is a finite constant. Then, the limiting distribution of* $\sum_{n=1}^{N} G_n$ *is the convolution* $\Phi_J * \eta$, *where* Φ_J *is the standard multinormal distribution in* $\mathbb{R}^J$.

Corollary B.4. *In addition, if η is the multinormal distribution with mean vector μ and variance-covariance matrix Σ, then*

$$\sum_{n=1}^{N} G_n \xrightarrow{d} \mathcal{N}(\mu, \mathbf{I}_J + \Sigma) \qquad as \quad N \to \infty.$$

Sketch of Proof Let $\eta_{\mathbf{Q}_N}$ be the distribution of $\mathbf{Q}_N$. Let C be a convex of $\mathbb{R}^J$,

$$\mathbb{P}\left(\sum_{n=1}^{N} G_n \in C\right) = \int_{R^I} \mathbb{P}\left(\sum_{n=1}^{N} G_n \in C | \mathbf{Q}_N = q\right) d\eta_{\mathbf{Q}_N}(q) \tag{B.12}$$

$$= \int_{R^I} \mathbb{P}\{S_N \in C - g(q) | \mathbf{Q}_N = q\} d\eta_{\mathbf{Q}_N}(q) \tag{B.13}$$

Suppose conditions (B.9)-(B.11) are satisfied. Apply the Berry-Esseen theorem in $\mathbb{R}^J$ (Götze, 1991) to the conditional sum $S_N|\mathbf{Q}_N$ of the independent random vectors $G_n - N^{-1}g(\mathbf{Q}_N)|\mathbf{Q}_N$, $n = 1, \ldots, N$: for all convex C_0 in $\mathbb{R}^J$

$$|\mathbb{P}(S_N \in C_0 | \mathbf{Q}_N) - \Phi_J(C_0)| \leq N^{-1/2} c_J k \tag{B.14}$$

where c_J is a finite constant depending on J. It follows from expressions (B.13) and (B.14)

$$\left|\mathbb{P}\left(\sum_{n=1}^{N} G_n \in C\right) - \int_{R^J} \Phi_J(C-u) d\eta(u)\right| \tag{B.15}$$

$$\leq \left|\int_{R^J} [\mathbb{P}\{S_N \in C - g(q) | \mathbf{Q}_N\} - \Phi_J\{C - g(q)\}] d\eta_{\mathbf{Q}_N}(q)\right| \tag{B.16}$$

$$+ \left|\int_{R^J} \Phi_J(C-u) d\eta_{g(\mathbf{Q}_N)}(u) - \int_{R^J} \Phi_J(C-u) d\eta(u)\right|. \tag{B.17}$$

$$\leq N^{-1/2} c_J k + \left|\int_{R^J} \Phi_J(C-u) d\eta_{g(\mathbf{Q}_N)}(u) - \int_{R^J} \Phi_J(C-u) d\eta(u)\right|. \tag{B.18}$$

From the Portmanteau theorem (Billingsley, 1968), the second term in expression (B.18) tends to 0 as $N \to \infty$ and

$$\mathbb{P}\left(\sum_{n=1}^{N} G_n \in C\right) \longrightarrow \int_{R^J} \Phi_J(C-u) d\eta(u) \quad \text{as} \quad N \to \infty.$$

Consequently, the limiting distribution of $\sum_{n=1}^{N} G_n$ is the convolution $\Phi_J * \eta$, i.e. it is the distribution of the sum of two independent random variables with distributions, respectively, Φ_J and η. The proof of the corollary is then obvious.

Application of lemma B.3 and corollary B.4. Set

$$\begin{aligned} v(q, \theta) &= \mathbb{V}_\theta\{U_n(\mathbf{Q}_N, \theta) | \mathbf{Q}_N = q\} \\ G_n &= N^{-1/2} v(\mathbf{Q}_N, \theta)^{-1/2} U_n(\mathbf{Q}_N, \theta) \\ g(\mathbf{Q}_N) &= N \mathbb{E}_\theta(G_n | \mathbf{Q}_N). \end{aligned}$$

Expression of $v(\mathbf{Q}_N, \theta)$ is given below, and $v(\mathbf{Q}_N, \theta)^{-1/2}$ exists under assumption (B.i). Moreover, as $N\mathbf{Q}_N$ follows a multinomial distribution with vector of probabilities ν, the limiting distribution η of $g(\mathbf{Q}_N)$ is multinormal with mean vector 0 and variance-covariance matrix $\Omega(\nu, \theta)$

$$\begin{aligned}
\Omega(\nu,\theta) &= v(\nu,\theta)^{-1/2} w(\theta) \Gamma(\nu) w(\theta)^T v(\nu,\theta)^{-1/2} \\
w(\theta) &= \begin{pmatrix} \mathbf{0}_I & \cdots & \mathbf{0}_I \\ \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)|C_n=1\} & \cdots & \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)|C_n=I\} \end{pmatrix} \\
\Gamma(\nu) &= \mathrm{diag}(\nu_1,\ldots,\nu_I) - \nu\ \nu^T.
\end{aligned}$$

where $\mathbf{0}_I$ is the vector of size I whose components are 0.

With this setting, it can be shown that conditions of lemma B.3 and corollary B.4 are satisfied. In particular, condition (B.11) is satisfied under integrability assumptions (B.g1) and (B.g3) and because $q \mapsto v(q, \theta)$ is continuous over $[0, 1]^I$. Therefore, corollary B.4 leads to

$$\begin{aligned}
\sum_{n=1}^{N} G_n = \sqrt{N} v(\mathbf{Q}_N,\theta)^{-1/2} \frac{1}{N} \sum_{n=1}^{N} U_n(\mathbf{Q}_N,\theta) \\
\xrightarrow{d} \mathcal{N}\{0, \mathbf{I}_{I+d} + \Omega(\nu,\theta)\} \quad \text{as} \quad N \to \infty,
\end{aligned} \tag{B.19}$$

Expression of $v(q, \theta)$. Let $q = (q_1, \ldots, q_I)^T$ be in $[0, 1]^I$ such as $q_1 + \cdots + q_I = 1$ and θ be in Θ^I.

$$\begin{aligned}
v(q,\theta) &= \mathbb{V}_\theta\{U_n(\mathbf{Q}_N,\theta)|\mathbf{Q}_N = q\} \\
&= \left(\begin{array}{ccc|c} v_{11} & \cdots & v_{1I} & v_1^T \\ \vdots & & \vdots & \vdots \\ v_{I1} & \cdots & v_{II} & v_I^T \\ \hline v_1 & \cdots & vI & v \end{array}\right),
\end{aligned} \tag{B.20}$$

where for (i, j) in $\{1, \ldots, I\}^2$

$$\begin{aligned}
v_{ij} =& \delta_{i=j} q_i \mathbb{E}_\theta(Y_n^2|C_n=i) - q_i q_j \mathbb{E}_\theta(Y_n|C_n=i)\mathbb{E}_\theta(Y_n|C_n=j) \\
v_i =& q_i \mathbb{E}_\theta\{Y_n \nabla \log \pi(Y_n|C_n,\lambda^*)|C_n=i\} \\
& - q_i \mathbb{E}_\theta(Y_n|C_n=i) \sum_{k=1}^{I} q_k \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)|C_n=k\} \\
v =& \sum_{k=1}^{I} q_k \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*) \nabla \log \pi(Y_n|C_n,\lambda^*)^T|C_n=k\} \\
& - \left[\sum_{k=1}^{I} q_k \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)|C_n=k\}\right] \left[\sum_{k=1}^{I} q_k \mathbb{E}_\theta\{\nabla \log \pi(Y_n|C_n,\lambda^*)|C_n=k\}\right]^T.
\end{aligned}$$

Limit of $T(\mathbf{Q}_N, \lambda^*, \bar{\theta})v(\mathbf{Q}_N, \theta)^{1/2}$

$T(\cdot, \lambda^*, \bar{\theta})$ is continuous at ν. Moreover, $v(\cdot, \theta)$ is continuous at ν and so is $v(\cdot, \theta)^{1/2}$. As $N\mathbf{Q}_N$ follows a multinomial distribution with vector of probabilities ν, $\mathbf{Q}_N$ tends to ν almost surely and

$$T(\mathbf{Q}_N, \lambda^*, \bar{\theta})v(\mathbf{Q}_N, \theta)^{1/2} \xrightarrow[\mathbb{P}_\theta - \text{a.s.}]{} T(\nu, \lambda^*, \bar{\theta})v(\nu, \theta)^{1/2} \quad \text{as} \quad N \to \infty. \tag{B.21}$$

Limiting distribution of $\sqrt{N}\psi_{N,\theta}$

From convergences (B.19) and (B.21),

$$\begin{aligned}\sqrt{N}\psi_{N,\theta} &= \sqrt{N}\; T(\mathbf{Q}_N, \lambda^*, \bar{\theta})\frac{1}{N}\sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta) \\ &= \left\{T(\mathbf{Q}_N, \lambda^*, \bar{\theta})v(\mathbf{Q}_N, \theta)^{1/2}\right\}\left\{\sqrt{N}\; v(\mathbf{Q}_N, \theta)^{-1/2}\frac{1}{N}\sum_{n=1}^{N} U_n(\mathbf{Q}_N, \theta)\right\} \\ &\xrightarrow{d} \mathcal{N}\{0, \Sigma(\nu, \theta)\} \quad \text{as} \quad N \to \infty,\end{aligned} \tag{B.22}$$

where

$$\Sigma(\nu, \theta) = T(\nu, \lambda^*, \bar{\theta})v(\nu, \theta)^{1/2}\{\mathbf{I}_{I+d} + \Omega(\nu, \theta)\}v(\nu, \theta)^{1/2}T(\nu, \lambda^*, \bar{\theta})^T. \tag{B.23}$$

Convergence (B.22) has been proved for θ fixed in Θ. We now show that the convergence is uniform on Θ^I. Convergence (B.22) can be written in term of characteristic functions

$$\Delta_t(\theta) = \mathbb{E}_\theta\left(e^{\mathbf{i}t^T\sqrt{N}\psi_{N,\theta}}\right) - \mathbb{E}_\theta\left(e^{\mathbf{i}t^T X_\theta}\right) \to 0 \quad \text{as} \quad N \to \infty, \quad \forall t \in \mathbb{R}^I,$$

where $X_\theta \sim \mathcal{N}\{0, \Sigma(\nu, \theta)\}$. $\bar{\theta}$ is a continuous function of θ. λ^* is a continuous function of θ (implicit function theorem). Functions $\lambda \mapsto \pi(y|i, \lambda)$, $i = 1, \ldots, I$ and $y \in \mathbb{R}$, are continuous. Using these elements, it can be shown that Δ_t is continuous on Θ^I. As Θ^I is a compact (assumption (B.a)), Δ_t is uniformly continuous on Θ^I and, consequently,

$$\sup_{\theta \in \Theta^I} |\Delta_t(\theta)| \to 0 \quad \text{as} \quad N \to \infty, \quad \forall t \in \mathbb{R}^I,$$

i.e. convergence (B.22) is uniform on Θ^I.

Estimation of $\Sigma(\nu, \theta)$

The limiting variance $\Sigma(\nu, \theta)$ of $\sqrt{N}\psi_{N,\theta}$ can be estimated by plugging-in estimates of unknown quantities. In $\Sigma(\nu, \theta)$, there are ν, $\hat{\lambda}_N$, λ^* and θ_i, $i = 1, \ldots, I$, in terms of the form $\mathbb{E}_\theta\{f(C_n, Y_n)|C_n = i\} = \int_{\mathbb{R}} f(i, y)\pi(y|i, \theta_i)d\mu(y)$. ν can be replaced by $\mathbf{Q}_N$, λ^* and $\bar{\theta}$ by $\hat{\lambda}_N$, and θ_i by its estimate obtained with our method.

B.4 Assumptions

(B.a) Θ is a closed bounded convex subset of $\mathbb{R}^d$

(B.b) $(i, y, \lambda) \mapsto \pi(y|i, \lambda)$ is positive on $\mathcal{I} \times \mathbb{R} \times \Theta$

(B.c) $\lambda \mapsto \pi(y|i, \lambda)$ is 3 times differentiable on Θ for all $(i, y) \in \mathcal{I} \times \mathbb{R}$

(B.d) $\lambda \mapsto l_N(\lambda, \mathbf{C}, \mathbf{Y})$ is concave on Θ for all $(\mathbf{C}, \mathbf{Y}) \in (\mathcal{I} \times \mathbb{R})^N$

(B.e) $\lambda \mapsto l_\varepsilon(\lambda, \theta)$ is concave on Θ for all θ in Θ^I

(B.f) For all θ in Θ^I, $\lambda \mapsto l^*(\lambda, \theta)$ has a unique maximum in the interior of Θ

(B.g) there exists a constant $M < \infty$ such that for all θ in Θ^I, for all i in $\mathcal{I} = \{1, \ldots, I\}$, for all λ, λ' in Θ, and for all k, k', k'' in $\{1, \ldots, d\}$,

(B.g1) $\mathbb{E}_\theta(|Y_n|^3 \mid C_n = i) \leq M$

(B.g2) $|\mathbb{E}_\theta\{\log \pi(Y_n|C_n, \lambda) \mid C_n = i\}| \leq M$

(B.g3) $\mathbb{E}_\theta\{|\nabla_k \log \pi(Y_n|C_n, \lambda)|^3 \mid C_n = i\} \leq M$

(B.g4) $|\mathbb{E}_\theta\{\nabla_k \nabla_{k'} \log \pi(Y_n|C_n, \lambda) \mid C_n = i\}| \leq M$

(B.g5) $|\mathbb{E}_\theta\{\nabla_k \nabla_{k'} \nabla_{k''} \log \pi(Y_n|C_n, \lambda) \mid C_n = i\}| \leq M$

(B.g6) $|\int_{\mathbb{R}} y \nabla_k \pi(y|i, \lambda) \mu(dy)| \leq M$

(B.g7) $|\int_{\mathbb{R}} y \nabla_k \nabla_{k'} \pi(y|i, \lambda) \mu(dy)| \leq M$

(B.g8) $|\int_{\mathbb{R}} y \nabla_k \nabla_{k'} \nabla_{k''} \pi(y|i, \lambda) \mu(dy)| \leq M$

(B.h) $F(\lambda, \lambda')$ is invertible for all λ and λ' in Θ

(B.i) $v(q, \theta)$ is positive definite for all θ in Θ^I and for all $q \in [0, 1]^I$ such as the sum of its components is 1 (the expression of $v(q, \theta)$ is given in equation B.20)

C

Annexe du chapitre 8

C.1 Expression of the bias $\bar{\theta} - \hat{\lambda}_N$

In the following, assumptions (C.a)-(C.c) are assumed to be satisfied. Assumptions (C.a)-(C.c) as well as other assumptions mentioned in this section are written in appendix C.4. From the strong law of large numbers and under integrability assumption (C.g2), the $\mathbb{P}_\theta$-a.s.-limit of $l_N(\lambda, \mathbf{Z}, \mathbf{Y})$ (equation (8.5)) as $N \to \infty$ is

$$\begin{aligned} l^*(\lambda, \theta) &= \mathbb{E}_\theta[\log \pi(Y_n | Z_n, \lambda)] \\ &= \int_{\mathcal{Z}} \int_{\mathbb{R}} \log \pi(y|z, \lambda) \mathbb{P}_\theta(dz, dy). \end{aligned}$$

Let λ^* be the maximizer of $l^*(\cdot, \theta)$. Define

$$\begin{aligned} l_\varepsilon(\lambda, \theta) &= \mathbb{E}_\theta[\log \pi(Y_n | Z_n, \lambda + \theta(Z_n) - \bar{\theta})] \\ &= \int_{\mathcal{Z}} \int_{\mathbb{R}} \log \pi\{y|z, \lambda + \theta(z) - \bar{\theta}\} \mathbb{P}_\theta(dz, dy). \end{aligned}$$

$\bar{\theta}$ is the maximizer of $l_\varepsilon(\cdot, \theta)$. Moreover, from a Taylor expansion and under integrability assumption (C.g3) the limit of $l_\varepsilon(\lambda, \theta)$ as $\varepsilon \to 0$ is $l^*(\lambda, \theta)$.
Remark that λ^* and $\bar{\theta}$ depend only on $\mathbb{P}_\theta$, i.e. on θ, ν, μ and π. As ν, μ and π are fixed, we consider λ^* and $\bar{\theta}$ as functions of θ.

Lemma C.1. *Under concavity assumptions (C.d) and (C.e), assumption (C.f) of uniqueness of λ^*, and integrability assumptions (C.g2)-(C.g5),*

$$\hat{\lambda}_N \xrightarrow[\mathbb{P}_\theta - a.s.]{} \lambda^* \quad as \quad N \to \infty \tag{C.1}$$

$$\bar{\theta} \longrightarrow \lambda^* \quad as \quad \varepsilon \to 0, \tag{C.2}$$

and

$$
\begin{aligned}
F(\lambda^*,\bar{\theta})(\lambda^*-\hat{\lambda}_N) &= \left[\frac{1}{N}\sum_{n=1}^{N}\nabla\log\pi(Y_n|Z_n,\lambda^*) - \mathbb{E}_\theta\{\nabla\log\pi(Y_n|Z_n,\lambda^*)\}\right] \\
&\quad + o_{\mathbb{P}_\theta-a.s.}(\varepsilon+||\lambda^*-\hat{\lambda}_N||_\infty), \quad as \quad N\to\infty \\
F(\lambda^*,\bar{\theta})(\bar{\theta}-\lambda^*) &= -\frac{1}{N}\sum_{n=1}^{N}F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\} \\
&\quad + \left[\frac{1}{N}\sum_{n=1}^{N}F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\} - \mathbb{E}_\theta[F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\}]\right] \\
&\quad + o(\varepsilon+||\bar{\theta}-\lambda^*||_\infty), \quad as \quad \varepsilon\to 0,
\end{aligned}
$$

where

$$F(z,\lambda^*,\bar{\theta}) = -\mathbb{E}_{\bar{\theta}^{\mathcal{Z}}}\{\mathbf{H}\log\pi(Y_n|Z_n,\lambda^*) \mid Z_n=z\} \tag{C.3}$$

$$F(\lambda^*,\bar{\theta}) = \int_{\mathcal{Z}} F(z,\lambda^*,\bar{\theta})\nu(dz) \tag{C.4}$$

$$= -\mathbb{E}_{\bar{\theta}^{\mathcal{Z}}}\{\mathbf{H}\log\pi(Y_n|Z_n,\lambda^*)\}. \tag{C.5}$$

$F(z,\lambda^*,\bar{\theta})$ is a Fisher information matrix at location $z\in\mathcal{Z}$ and $F(\lambda^*,\bar{\theta})$ is a mean Fisher information matrix.

Sketch of proof Convergences (C.1) and (C.2) can be shown by using a Rockafellar theorem and a corollary both written in Senoussi (1990). The consecutive equations are obtained by using Taylor expansions.

Corollary C.2. *In addition, if $F(\lambda,\lambda')$ is invertible for all λ and λ' in Θ (assumption (C.h)), then as $N\to\infty$ and $\varepsilon\to 0$*

$$\bar{\theta}-\hat{\lambda}_N \xrightarrow[\mathbb{P}_{\mathcal{D}}-a.s.]{} 0$$

and

$$
\begin{aligned}
\bar{\theta}-\hat{\lambda}_N &= -F(\lambda^*,\bar{\theta})^{-1}\frac{1}{N}\sum_{n=1}^{N}F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\} \\
&\quad + F(\lambda^*,\bar{\theta})^{-1}\left[\frac{1}{N}\sum_{n=1}^{N}\nabla\log\pi(Y_n|Z_n,\lambda^*) - \mathbb{E}_\theta\{\nabla\log\pi(Y_n|Z_n,\lambda^*)\}\right. \\
&\quad \left. + \frac{1}{N}\sum_{n=1}^{N}F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\} - \mathbb{E}_\theta[F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\}]\right] \\
&\quad + o_{\mathbb{P}_{\mathcal{D}}-a.s.}(\varepsilon+||\lambda^*-\hat{\lambda}_N||_\infty) + o(\varepsilon+||\bar{\theta}-\lambda^*||_\infty).
\end{aligned} \tag{C.6}
$$

C.2 Expression of local residual $\hat{R}_N(z)$

Ordinary residual $r_N(n)$ (equation (8.6)) can be written

$$r_N(n) = \{Y_n - \mathbb{E}_\theta(Y_n|Z_n)\} + \{E_\theta(Y_n|Z_n) - E_{\hat{\lambda}_N^z}(Y_n|Z_n)\}.$$

Suppose assumptions mentioned in appendix C.1 and assumptions of integrability (C.g6)-(C.g8) are satisfied. By using Taylor expansions and expression (C.6) of the bias $\bar{\theta} - \hat{\lambda}_N$,

$$\begin{aligned} r_N(n) =& \{Y_n - \mathbb{E}_\theta(Y_n|Z_n)\} \\ &+ A(Z_n, \hat{\lambda}_N)^T \left[\{\theta(Z_n) - \bar{\theta}\} - F(\lambda^*, \bar{\theta})^{-1} \frac{1}{N} \sum_{m=1}^{N} F(Z_m, \lambda^*, \bar{\theta})\{\theta(Z_m) - \bar{\theta}\}\right] \\ &+ A(Z_n, \lambda^*)^T F(\lambda^*, \bar{\theta})^{-1} \left[\frac{1}{N} \sum_{m=1}^{N} \nabla \log \pi(Y_m|Z_m, \lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_m|Z_m, \lambda^*)\}\right. \\ &\quad \left. + \frac{1}{N} \sum_{m=1}^{N} F(Z_m, \lambda^*, \bar{\theta})\{\theta(Z_m) - \bar{\theta}\} - \mathbb{E}_\theta[F(Z_m, \lambda^*, \bar{\theta})\{\theta(Z_m) - \bar{\theta}\}]\right] \\ &+ \zeta_{N,\varepsilon}(Z_n), \end{aligned} \tag{C.7}$$

where $A(z, \lambda) = \int_{\mathbb{R}} y \nabla \pi(y|z, \lambda) \mu(dy)$, $z \in \mathcal{Z}$ and $\lambda \in \Theta$, and $\zeta_{N,\varepsilon}(Z_n) = o_{\mathbb{P}_\mathcal{D} - a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta} - \lambda^*||_\infty)$ as $N \to \infty$ and $\varepsilon \to 0$.

Then, as $\hat{R}_N(s) = \sum_{n=1}^{N} w_n(s) r_N(n)$ (equations (8.7) and (8.8)), equation (C.7) yields to equation (8.9).

C.3 Sketch of proof of theorem 8.1

C.3.1 Decomposition of the vector $\hat{\mathbf{R}}_N$ of local residuals

Set

$$\mathbf{A}(\lambda) = \begin{pmatrix} A(Z_1, \lambda)^T & & (0) \\ & \ddots & \\ (0) & & A(Z_N, \lambda)^T \end{pmatrix}, \quad \lambda \in \Theta,$$

$$W_i = \begin{pmatrix} w_1(s_i) \\ \vdots \\ w_N(s_i) \end{pmatrix}, \ i = 1, \ldots, I, \quad \text{and} \quad W = \begin{pmatrix} W_1^T \\ \vdots \\ W_I^T \end{pmatrix},$$

$$\begin{aligned} B_0(\hat{\lambda}_N, \lambda^*, \bar{\theta}) =& W \mathbf{A}(\hat{\lambda}_N) \left[\mathbf{I}_{dN}\right. \\ & \left. - N^{-1}(\mathbf{1}_N \otimes \mathbf{I}_d) F(\lambda^*, \bar{\theta})^{-1} \{F(Z_1, \lambda^*, \bar{\theta}), \ldots, F(Z_N, \lambda^*, \bar{\theta})\}\right], \end{aligned}$$

where $\mathbf{I}_k$ and $\mathbf{1}_k$ are, respectively, the identity matrix and the unit vector of dimension $k \in \mathbb{N}^*$, and $\otimes$ is the Kronecker product also called matrix direct product,

$$U_{N,\theta} = \frac{1}{N}\sum_{n=1}^{N} \nabla \log \pi(Y_n|Z_n,\lambda^*) - \mathbb{E}_\theta\{\nabla \log \pi(Y_n|Z_n,\lambda^*)\} + \frac{1}{N}\sum_{n=1}^{N} F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\} - \mathbb{E}_\theta[F(Z_n,\lambda^*,\bar{\theta})\{\theta(Z_n)-\bar{\theta}\}]$$

$$\psi_{N,\theta} = W\{\mathbf{Y} - E_\theta(\mathbf{Y}|\mathbf{Z})\} + W\mathbf{A}(\lambda^*)(\mathbf{1}_N \otimes \mathbf{I}_d)F(\lambda^*,\bar{\theta})^{-1}U_{N,\theta}, \tag{C.8}$$

$$\Delta_\theta = \begin{pmatrix} \theta(Z_1)-\bar{\theta} \\ \vdots \\ \theta(Z_N)-\bar{\theta} \end{pmatrix} \quad \text{and} \quad \Delta_\alpha = \begin{pmatrix} \alpha(Z_1)-\bar{\alpha} \\ \vdots \\ \alpha(Z_N)-\bar{\alpha} \end{pmatrix}.$$

Then, the vector of local residuals $\hat{\mathbf{R}}_N = \{\hat{R}_N(s_1),\ldots,\hat{R}_N(s_I)\}^T$ can be written

$$\hat{\mathbf{R}}_N = B_0(\hat{\lambda}_N,\lambda^*,\bar{\theta})\Delta_\theta + \psi_{N,\theta} + \xi_{\varepsilon,N},$$

where $\xi_{N,\varepsilon} = \{\xi_{N,\varepsilon}(s_1),\ldots,\xi_{N,\varepsilon}(s_I)\}^T$ is $o_{\mathbb{P}_\mathcal{D}-a.s.}(\varepsilon + ||\lambda^* - \hat{\lambda}_N||_\infty + ||\bar{\theta}-\lambda^*||_\infty)\mathbf{1}_I$ as $N \to \infty$ and $\varepsilon \to 0$.

Only components $1,\ldots,d_1,\ldots,(I-1)d+1,\ldots,(I-1)d+d_1$ of Δ_θ can be different from 0; indeed, the others correspond to β. So, keep columns $1,\ldots,d_1,\ldots,(I-1)d+1,\ldots,(I-1)d+d_1$ of $B_0(\hat{\lambda}_N,\lambda^*,\bar{\theta})$ and denote the resulting matrix by $B(\hat{\lambda}_N,\lambda^*,\bar{\theta})$, then $B_0(\hat{\lambda}_N,\lambda^*,\bar{\theta})\Delta_\theta = B(\hat{\lambda}_N,\lambda^*,\bar{\theta})\Delta_\alpha$. We obtain equation (8.10)

$$\hat{\mathbf{R}}_N = B(\hat{\lambda}_N,\lambda^*,\bar{\theta})\Delta_\alpha + \psi_{N,\theta} + \xi_{N,\varepsilon}.$$

Note that summing block-columns of size d of $B_0(\hat{\lambda}_N,\lambda^*,\bar{\theta})$ and taking the limit as $N \to \infty$ lead to the zero matrix. Therefore, $B_0(\hat{\lambda}_N,\lambda^*,\bar{\theta})$ is not asymptotically invertible and $B(\hat{\lambda}_N,\lambda^*,\bar{\theta})$ neither.

C.3.2 Limiting distribution of $\psi_{N,\theta}$

Limiting distribution of $W\{\mathbf{Y} - \mathbb{E}_\theta(\mathbf{Y}|\mathbf{Z})\}$

$W\{\mathbf{Y} - \mathbb{E}_\theta(\mathbf{Y}|\mathbf{Z})\}$ is the kernel estimator of the vector whose components are $\mathbb{E}_\theta\{Y_n - \mathbb{E}_\theta(Y_n|Z_n)|Z_n = s_i\} = 0$, $i = 1,\ldots,I$. Let f_ν denote the density of Z_n over $\mathcal{Z}$. Using theorem IV-1 Bosq and Lecoutre (1987, p. 122), under assumptions (C.i1)-(C.i4), if $h_N \to 0$ and there exists $a > 0$ such as $N^{1-a}h_N^q \to \infty$ as $N \to \infty$, then

$$\sqrt{Nh_N^q}\left(\int_\mathcal{Z} K^2\right)^{-1/2} \operatorname{diag}\left\{\frac{f_\nu(s_i)}{\mathbb{V}_\theta(Y_n|Z_n=s_i)} : i = 1,\ldots,I\right\}^{1/2} W\{\mathbf{Y} - \mathbb{E}_\theta(\mathbf{Y}|\mathbf{Z})\} \xrightarrow{d} \mathcal{N}(0,\mathbf{I}_I). \tag{C.9}$$

Limit of $W\mathbf{A}(\lambda^*)(\mathbf{1}_N \otimes \mathbf{I}_d)$

Using theorem II-1 in Bosq and Lecoutre (1987, p. 106), under assumptions (C.i1), (C.i2), (C.i5) and (C.i6), if $h_N \to 0$ and $Nh_N^q \to \infty$ as $N \to \infty$, then

$$W\mathbf{A}(\lambda^*)(\mathbf{1}_N \otimes \mathbf{I}_d) \xrightarrow{\mathbb{P}_\theta} \begin{pmatrix} A(s_1, \lambda^*)^T \\ \vdots \\ A(s_I, \lambda^*)^T \end{pmatrix}. \tag{C.10}$$

Limiting distribution of $U_{N,\theta}$

Using the central limit theorem, under assumptions (C.g3) and (C.g4), as $N \to \infty$

$$\sqrt{N}U_{N,\theta} \xrightarrow{d} \mathcal{N}[0, \mathbb{V}_\theta[\nabla \log \pi(Y_n|Z_n, \lambda^*) + F(Z_n, \lambda^*, \bar{\theta})\{\theta(Z_n) - \bar{\theta}\}]]. \tag{C.11}$$

Limiting distribution of $\sqrt{Nh_N^q}\psi_{N,\theta}$ *conditional on* θ

Combining convergences (C.10) and (C.11) leads to

$$\sqrt{Nh_N^q}W\mathbf{A}(\lambda^*)(\mathbf{1}_N \otimes \mathbf{I}_d)F(\lambda^*, \bar{\theta})^{-1}U_{N,\theta} \xrightarrow{\mathbb{P}_\theta} 0 \quad \text{as} \quad N \to \infty. \tag{C.12}$$

Then, from convergences (C.9) and (C.12)

$$\sqrt{Nh_N^q}\psi_{N,\theta} \xrightarrow{d} \mathcal{N}\{0, \Sigma(\theta)\} \quad \text{as} \quad N \to \infty, \tag{C.13}$$

where

$$\Sigma(\theta) = \left(\int_{\mathcal{Z}} K^2\right) \operatorname{diag}\left\{\frac{\mathbb{V}_\theta(Y_n|Z_n = s_i)}{f_\nu(s_i)} : i = 1, \ldots, I\right\}. \tag{C.14}$$

Convergence (C.13) has been proved for θ fixed in Θ (pointwise convergence). The next section shows that the convergence is not conditional on θ.

Unconditional weak convergence

Convergence (C.13) can be written in term of characteristic functions

$$\delta_{t,N}(\theta) = \mathbb{E}_\theta\left(e^{\mathbf{i}t^T\sqrt{Nh_N^q}\psi_{N,\theta}}\right) - \mathbb{E}_\theta\left(e^{\mathbf{i}t^T X_\theta}\right) \to 0 \quad \text{as} \quad N \to \infty, \quad \forall t \in \mathbb{R}^I,$$

where $X_\theta \sim \mathcal{N}\{0, \Sigma(\theta)\}$. Apply the dominated convergence theorem of Lebesgue (Rudin, 1998) to the sequence of functions $\delta_{t,N}$, $N = 1, 2, \ldots$, defined over $\mathscr{C}(\mathcal{Z}, \Theta)$. For all $\theta \in \mathscr{C}(\mathcal{Z}, \Theta)$ $\lim_{N\to\infty} \delta_{N,t}(\theta) = 0$. In addition, for all $\theta \in \mathscr{C}(\mathcal{Z}, \Theta)$ and for all $N = 1, 2, \ldots$, $|\delta_{N,t}(\theta)| \leq 2$ and the function $\theta \mapsto 2$ is $\mathcal{D}$-integrable over $\mathscr{C}(\mathcal{Z}, \Theta)$ (probability measure $\mathcal{D}$ is introduced in equation (8.4)). Therefore

$$\lim_{N\to\infty} \int_{\mathscr{C}(\mathcal{Z},\Theta)} \delta_{N,t}(\theta)\mathbb{P}_{\mathcal{D}}(d\theta) = 0$$

(probability measure $\mathbb{P}_{\mathcal{D}}$ is introduced in equation (8.4)). This equation means that for all $t \in \mathbb{R}^I$

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}\{\delta_{N,t}(\theta)\} &= \lim_{N\to\infty} \mathbb{E}\left(e^{\mathbf{i}t^T\sqrt{Nh_N^q}\psi_{N,\theta}}\right) - \mathbb{E}\left(e^{\mathbf{i}t^T X_\theta}\right) \\ &= 0.\end{aligned}$$

The convergence to 0 of the difference between the characteristic functions is no more conditional on θ.

Estimation of the limiting variance $\Sigma(\theta)$

To estimate the limiting variance-covariance matrix $\Sigma(\theta)$ of $\sqrt{Nh_N^q}\psi_{N,\theta}$, the unknown quantity $\theta(s_i)$ needed to compute $\mathbb{V}_\theta(Y_n|Z_n = s_i)$ is replaced by its estimate obtained with our method.

C.4 Assumptions

(C.a) Θ is a closed bounded convex subset of $\mathbb{R}^d$

(C.b) $(z, y, \lambda) \mapsto \pi(y|z, \lambda)$ is positive on $\mathcal{Z} \times \mathbb{R} \times \Theta$

(C.c) $\lambda \mapsto \pi(y|z, \lambda)$ is 3 times differentiable on Θ for all $(z, y) \in \mathcal{Z} \times \mathbb{R}$

(C.d) $\lambda \mapsto l_N(\lambda, \mathbf{Z}, \mathbf{Y})$ is concave on Θ for all $(\mathbf{Z}, \mathbf{Y}) \in (\mathcal{Z} \times \mathbb{R})^N$

(C.e) $\lambda \mapsto l_\varepsilon(\lambda, \theta)$ is concave on Θ for all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$

(C.f) For all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$, $\lambda \mapsto l^*(\lambda, \theta)$ has a unique maximum in the interior of Θ

(C.g) there exists a constant $M < \infty$ such that for all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$, for all z in $\mathcal{Z}$, for all λ, λ' in Θ, and for all k, k', k'' in $\{1, \dots, d\}$,

- (C.g1) $\mathbb{E}_\theta(|Y_n|^3 \mid Z_n = z) \leq M$
- (C.g2) $|\mathbb{E}_\theta\{\log \pi(Y_n|Z_n, \lambda) \mid Z_n = z\}| \leq M$
- (C.g3) $\mathbb{E}_\theta\{|\nabla_k \log \pi(Y_n|Z_n, \lambda)|^3 \mid Z_n = z\} \leq M$
- (C.g4) $\mathbb{E}_\theta\{|\nabla_k \nabla_{k'} \log \pi(Y_n|Z_n, \lambda)|^2 \mid Z_n = z\} \leq M$
- (C.g5) $|\mathbb{E}_\theta\{\nabla_k \nabla_{k'} \nabla_{k''} \log \pi(Y_n|Z_n, \lambda) \mid Z_n = z\}| \leq M$
- (C.g6) $|\int_{\mathbb{R}} y \nabla_k \pi(y|z, \lambda) \mu(dy)| \leq M$
- (C.g7) $|\int_{\mathbb{R}} y \nabla_k \nabla_{k'} \pi(y|z, \lambda) \mu(dy)| \leq M$
- (C.g8) $|\int_{\mathbb{R}} y \nabla_k \nabla_{k'} \nabla_{k''} \pi(y|z, \lambda) \mu(dy)| \leq M$

(C.h) $F(\lambda, \lambda')$ is invertible for all λ and λ' in Θ

(C.i) (C.i1) K is a positive Parzen-Rozenblatt kernel

- (C.i2) f_ν is continuous and strictly positive in $\mathcal{Z}$
- (C.i3) $z \mapsto \mathbb{V}_\theta(Y_n|Z_n = z)$ is continuous, strictly positive and bounded in $\mathcal{Z}$ for all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$
- (C.i4) $\mathbb{E}_\theta\{|Y_n - \mathbb{E}_\theta(Y_n|Z_n)|^p\} < \infty$ for all p in $\mathbb{N}$ and for all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$
- (C.i5) $z \mapsto \int_{\mathbb{R}} y \nabla \pi(y|z, \lambda^*) \mu(dy)$ is continuous in $\mathcal{Z}$
- (C.i6) $\mathbb{E}_\theta\left[\{\int_{\mathbb{R}} y \nabla_k \pi(y|Z_n, \lambda^*)\}^2 \mu(dy)\right] < \infty$ for all k in $\{1, \dots, d\}$ and for all θ in $\mathscr{C}(\mathcal{Z}, \Theta)$

D

Annexe du chapitre 10

Exemple schématique de méta-analyse pour estimer la fonction de dispersion sur une large gamme de distance

Nous présentons ici un exemple schématique de méta-analyse visant à estimer la fonction de dispersion $h(\cdot)$ sur une large gamme de distances. Cet exemple accompagne la section 10.3.

Expériences

Supposons que l'on dispose de trois séries d'expériences
- $1^{\text{ère}}$ série : propagation spatiale à courte distance comme au chapitre 3.
- $2^{\text{ème}}$ série : propagation spatiale à longue distance comme au chapitre 4.
- $3^{\text{ème}}$ série : propagation spatiale depuis une parcelle agricole volontairement et entièrement infectée vers des parcelles susceptibles aux alentours. Après un cycle épidémique, on observe dans quelles parcelles la maladie est présente. (cette expérience n'est certainement pas réalisable, mais sert à la compréhension du principe de la méta-analyse).

Pour chaque type d'expériences, on construit un modèle à partir du cadre de modélisation de la section 5.1.

Modèle pour le 1^{er} type d'expériences

On observe les nombres de lésions sur n_1 feuilles situées autour d'une source de spores située en x_1.
$Y_{1,1}, \ldots, Y_{1,n_1}$: nombres de lésions sur les n_1 feuilles.
$x_{1,1}, \ldots, x_{1,n_1}$: positions (observées) des feuilles.
$Z_{1,1}, \ldots, Z_{1,n_1}$: fragilités (non observées) des feuilles.

$$Y_{1,n} \sim \text{indép. Poisson}\{Z_{1,n} A_1 h(x_{1,n} - x_1)\}$$
$$Z_{1,n} \sim \text{indép. } \mathcal{L}(\theta_1),$$

où A_1 est la force de la source et θ_1 est un vecteur de paramètres régissant la distribution des fragilités des feuilles.

Modèle pour le 2ème type d'expériences

On observe les nombres de feuilles malades dans n_2 placettes d'un mètre carré situées autour d'une source de spores située en x_2.
$Y_{2,1}, \ldots, Y_{2,n_2}$: nombres de feuilles malades dans les n_2 placettes.
$m_{2,1}, \ldots, m_{2,n_2}$: nombres totaux (observés) de feuilles dans les placettes.
$x_{2,1}, \ldots, x_{2,n_2}$: positions (observées) des placettes.
Z_2 : fragilité d'une feuille quelconque de l'expérience 2.

$$\begin{aligned} Y_{2,n} &\sim \text{indép. Binomiale}\{m_{2,n}, p(x_{2,n})\} \\ p(x_n) &= 1 - \exp\{-\mathbb{E}(Z_2)\tilde{A}_2 h(x_{2,n} - x_2)\} \\ Z_2 &\sim \mathcal{L}(\theta_2) \end{aligned}$$

où $\tilde{A}_2$ est la force de la source et θ_2 est un vecteur de paramètres régissant la distribution des fragilités des feuilles. $\mathbb{E}(Z_2)$ et $\tilde{A}_2$ étant confondus, leur produit est remplacé par un paramètre unique A_2, et

$$p(x_n) = 1 - \exp\{-A_2 h(x_{2,n} - x_2)\}.$$

Modèle pour le 3ème type d'expériences

On observe si la maladie est présente ou non dans n_3 parcelles agricoles réparties autour d'une parcelle-source située en x_3.
$Y_{3,1}, \ldots, Y_{3,n_3}$: absence/présence de maladie dans les n_3 parcelles.
$m_{3,1}, \ldots, m_{3,n_3}$: nombres totaux (observés) de feuilles dans les parcelles.
$x_{3,1}, \ldots, x_{3,n_3}$: positions (observées) des parcelles.
$Z_{3,n}$: fragilité d'une feuille quelconque de la parcelle n.

$$\begin{aligned} Y_{3,n} &\sim \text{indép. Bernouilli}\{q(x_{3,n})\} \\ q(x_{3,n}) &= 1 - [1 - \exp\{-\mathbb{E}(Z_{3,n})A_3 h(x_{3,n} - x_3)\}]^{m_{3,n}} \\ Z_{3,n} &\sim \mathcal{L}(\theta_{3,n}) \end{aligned}$$

où A_3 est la force de la source, $\theta_{3,n}$ est un vecteur de paramètres régissant la distribution des fragilités des feuilles appartenant à la parcelle n, et où les variations de la fonction de dispersion à l'intérieur de chaque parcelle sont négligées. Notons qu'à chaque parcelle est attribué un vecteur de paramètres $\theta_{3,n}$ propre. Ainsi, les fragilités des feuilles de deux parcelles différentes peuvent ne pas être distribuées de la même manière. En terme biologique, les parcelles peuvent présenter des susceptibilités ou réceptivités différentes.

Méta-modèle

Les paramètres sont donc
– les paramètres de la fonction de dispersion $h(\cdot)$,
– les forces des sources $A_1^{(1)}, \ldots, A_1^{(I_1)}$ des I_1 expériences de type 1,

- les "forces des sources" $A_2^{(1)}, \ldots, A_2^{(I_2)}$ des I_2 expériences de type 2,
- les forces des sources $A_3^{(1)}, \ldots, A_3^{(I_3)}$ des I_3 expériences de type 3,
- les vecteurs des paramètres des fragilités $\theta_1^{(1)}, \ldots, \theta_1^{(I_1)}$ des I_1 expériences de type 1,
- les espérances des fragilités $\mathbb{E}(Z_{3,1}), \ldots, \mathbb{E}(Z_{3,n_3})$ pour chacune des parcelles des expériences de type 3, espérances qui dépendent des vecteurs de paramètres $\theta_{3,1}, \ldots, \theta_{3,n_3}$.

Plutôt que de voir ces paramètres comme déterministes, on peut les voir comme aléatoires (hormis les paramètres de $h(\cdot)$).

- $A_1^{(1)}, \ldots, A_1^{(I_1)}$ sont indépendants et identiquement distribuées selon une loi paramétrée par a_1,
- $A_2^{(1)}, \ldots, A_2^{(I_2)}$ sont indépendants et identiquement distribuées selon une loi paramétrée par a_2,
- $A_3^{(1)}, \ldots, A_3^{(I_3)}$ sont indépendants et identiquement distribuées selon une loi paramétrée par a_3,
- les vecteurs de paramètres des fragilités $\theta_1^{(1)}, \ldots, \theta_1^{(I_1)}$ et $\theta_{3,1}, \ldots, \theta_{3,n_3}$ sont indépendants et identiquement distribués selon une loi paramétrée par θ.

Voir les paramètres comme aléatoire permet de réduire le nombre de paramètres que l'on a à estimer. Remarque : ayant peu d'information sur les différentes lois des paramètres, nous pourrions les spécifier en utilisant notre méthode d'analyse de résidus.

Ainsi, on a construit un méta-modèle paramétré par les paramètres de $h(\cdot)$, par a_1, a_2, a_3 et par θ. On peut alors procéder à l'estimation de ces paramètres et en déduire une estimation de $h(\cdot)$ sur la gamme de distances couverte par les trois types d'expériences.

De l'exemple schématique de méta-analyse à une méta-analyse réaliste

L'exemple de méta-analyse précédent est schématique (il a été donné pour comprendre le principe de la méta-analyse). Dans un contexte plus réaliste, des éléments spécifiques à chaque échelle d'observation devront être intégrés aux différents modèles (fonctions de covariables, effets aléatoires). C'est ce qui est expliqué dans la section 10.3.

Références

Agrios, G. A. (2005). *Plant Pathology* (5 ed.). Amsterdam : Elsevier.

Antoniadis, A., J. Berruyer, and R. Carmona (1992). *Régression Non Linéaire et Applications*. Paris : Economica.

Aylor, D. E. (1978). *Plant Disease : an Advanced Treatise*, Volume 2, Chapter Dispersal in time and space : aerial pathogens, pp. 159–180. London : Academic Press.

Aylor, D. E. (1987). Deposition gradients of urediniospores of puccina recondita near a source. *Phytopathology 77*, 1442–1448.

Aylor, D. E. (1990). The role of intermittent wind in the dispersal of fungal pathogens. *Annual Review of Phytopathology 28*, 73–92.

Aylor, D. E. (1998). The aerobiology of apple scab. *Plant Disease 82*, 838–849.

Baddeley, A., R. Turner, J. Møller, and M. Hazelton (2004). Residual analysis for spatial point processes. Technical Report 08, University of Western Australia, School of Mathematics and Statistics.

Bayles, R. A., K. Flath, M. S. Hovmøller, and C. De Vallavieille-Pope (2000). Breakdown of the yr17 resistance to yellow rust of wheat in northern europe. *Agronomie 20*, 805–811.

Berger, R. D., A. Bergamin Filho, and L. Amorin (1997). Lesion expansion as an epidemic component. *Phytopathology 87*, 1005–1013.

Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statictical Mathematics 43*, 1–59.

Bicout, D. J. and I. Sache (2003). Dispersal of spores following a persistent random walk. *Physical Review E 67*, 031913.1–7.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York : Wiley.

Blond, N., L. Bel, and R. Vautard (2003). Three-dimensional ozone data analysis with an air quality model over the paris area. *Journal of Geophysical Research 108*.

Bonjean, A. P. and W. J. Angus (2001). *The World Wheat Book*. Paris : Lavoisier.

Bosq, D. and J.-P. Lecoutre (1987). *Théorie de l'Estimation Fonctionnelle*. Paris : Economica.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis : Forecasting and Control, Revised Edition*. San Francisco : Holden-Day.

Box-Steffensmeier, J. M. and B. S. Jones (2004). *Event History Modeling, A Guide for Social Scientists*. Cambridge University Press.

Brown, J. K. M. and M. S. Hovmøller (2002). Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. *Science 297*, 537–541.

Bushnell, W. R. and A. P. Roelfs (1984). *The Cereal Rusts*, Volume 1. Orlando : Academic Press.

Cammack, R. H. (1958). Factors affecting infection gradients from a point source of puccinia polysora in a plot of zea mays. *Annals of Applied Biology 46*, 186–197.

Campbell, C. L. and L. V. Madden (1990). *Introduction to Plant Disease Epidemiology*. New York : John Wiley & Sons.

Chadœuf, J., D. Nandris, J. P. Geiger, M. Nicole, and J. C. Pierrat (1992). Modélisation spatio-temporelle d'une épidémie par un processus de gibbs : estimation et tests. *Biometrics 48*, 1165–1175.

Charbonneau, D., T. M. Brown, D. W. Latham, and M. Mayor (2000). Detection of planetary transits across a sun-like star. *The Astrophysical Journal 529*, L45–L48.

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

Christensen, O. F. (2004). Monte carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics 13*, 702–718.

Clayton, D. G. (1991). A monte carlo method for bayesian inference in frailty models. *Biometrics 47*, 467–485.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society B 48*, 133–169.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence Analysis*. New York : Chapman & Hall.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. New York : Wiley.

Critchley, F. and P. Marriott (2004). Data-informed influence analysis. *Biometrika 91*(1), 125–140.

Dacunha-Castelle, D. and M. Duflo (1982). *Probabilités et Statistiques. Problèmes à Temps Fixe*, Volume 1. Paris : Masson.

Danial, D. L., L. H. M. Broers, and J. E. Parlevliet (1993). Does interplot interference affect the screening of wheat for yellow rust resistance ? *Euphytica 70*, 217–224.

De Vallavieille-Pope, C., J. Rouzet, M. Leconte, M. Delos, and M. N. Mistou (2000). La rouille jaune du blé en france : des épidémies déclenchées par une nouvelle race, un hiver doux et un printemps humide. *Phytoma 527*, 22–29.

Debain, S. (2003). *L'expansion de Pinus sylvestris et de Pinus nigra sur le Causse Méjean : paramètres démographiques et interactions biotiques*. Ph. D. thesis, Ecole Nationale Supérieure Agronomique de Montpellier.

Desassis, N., P. Monestiez, J. N. Bacro, P. Lagacherie, and J. M. Robbez-Masson (2005). Mapping unobserved factors on vine plant mortality. In P. Renard, H. Demougeot-Renard, and R. Froidevaux (Eds.), *Geostatistics for Environmental Applications*, pp. 125–136.

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

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). Model-based geostatistics. *Applied Statistics 47*, 299–350.

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M.-J. Fortin, A. Jakomulska, M. Miriti, and M. S. Rosenberg (2002). A balanced view of scale in spatial statistical analysis. *Ecography 25*, 626–640.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York : Chapman & Hall.

Emge, R. G. and R. D. Shrum (1976). Epiphytology of puccinia striiformis at five selected locations in oregon during 1968 and 1969. *Phytopathology 66*, 1406–1412.

Eversmeyer, M. G. and C. L. Kramer (2000). Epidemiology of wheat leaf and stem rust in the central great plains of the usa. *Annual Review of Phytopathology 38*, 491–513.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge University Press.

Fitt, B. D. L., P. H. Gregory, A. D. Todd, H. A. McCartney, and O. C. Macdonald (1987). Spore dispersal and plant disease gradients : a comparison between two empirical models. *Journal of Phytopathology 118*, 227–242.

Geagea, L., L. Huber, and I. Sache (1999). Dry-dispersal and rain-splash of brown *puccinia recondita* f.sp. *tritici*) and yellow (*p. striiformis*) rust spores from infected wheat leaves exposed to simulated raindrops. *Plant Pathology 48*, 472–482.

Geagea, L., L. Huber, I. Sache, D. Flura, H. A. McCartney, and B. D. L. Fitt (2000). Influence of simulated rain on dispersal of rust spores from infected wheat seedlings. *Agricultural and Forest Meteorology 101*, 53–66.

Götze, F. (1991). On the rate of convergence in the multivariate clt. *The Annals of Probability 19*(2), 724–739.

Grancher, D., L. Bel, and R. Vautard (2005). Estimation de champs de pollution par adaptation statistique locale et approche non stationnaire. *Journal Européen des Systèmes automatisés 39*, 475–792.

Green, P. J., N. L. Hjort, and S. Richardson (Eds.) (2003). *Higly Structured Stochastic Systems*. Oxford University Press.

Green, P. J. and S. Richardson (2002). Hidden markov models and disease mapping. *Journal of the American Statistical Association 97*, 1055–1070.

Gregory, P. H. (1945). The dispersion of air-borne spores. *Transactions of the British Mycological Society 28*, 26–72.

Gregory, P. H. (1968). Interpreting plant disease dispersal gradients. *Annual Review of Phytopathology 6*, 189–212.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London : Chapman and Hall.

Henderson, R. and P. Oman (1999). Effect of frailty on marginal regression estimates in survival anlysis. *Journal of the Royal Statistical Society, B 61*, 367–379.

Henderson, R., S. Shimakura, and D. Gorst (2002). Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association 97*, 965–972.

Heyde, C. C. (1997). *Quasi-Likelihood and its Application. A General Approach to Optimal Parameter Estimation*. New York : Springer.

Hinde, J. and C. G. B. Demétrio (1998). Overdispersion : models and estimation. *Computational Statistics and Data Analysis 27*, 151–170.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics 42*, 1977–1991.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society B 60*, 497–536.

Hovmøller, M. S., A. F. Justesen, and J. K. M. Brown (2002). Clonality and long-distance migration of *puccinia striiformis* f.sp. *tritici* in north-west europe. *Plant Pathology 51*, 24–32.

Hrafnkelsson, B. and N. Cressie (2003). Hierarchical modeling of count data with application to nuclear fall-out. *Environmental and Ecological Statistics 10*, 179–200.

Huerta, G., B. Sansó, and J. R. Stroud (2004). A spatiotemporal model for mexico city ozone levels. *Journal of the Royal Statistical Society, C 53*, 231–248.

Johnson, R., R. W. Stubbs, E. Fuchs, and N. H. Chamberlain (1972). Nomenclature for physiologic races of puccinia striiformis infecting wheat. *Transactions of the British Mycological Society 58*, 475–480.

Jones, D. G. (1998). *The epidemiology of plant diseases*. Dordrecht : Kluwer.

Kingsolver, C. H., C. E. Peet, and J. F. Underwood (1984). Measurement of the epidemiologic potential of wheat stem rust : St croix, us virgin islands, 1954-57. *Pennsylvania Agricultural Experiment Station Bulletin 854*, 1–21.

Klein, E. K., C. Lavigne, X. Foueillassar, P.-H. Gouyon, and C. Larédo (2003). Corn pollen dispersal : quasi-mechanistic models and field experiments. *Ecological Monographs 73*, 131–150.

Kranz, J. (2003). *Comparative Epidemiology of Plant Diseases*. Berlin : Springer.

Lantuéjoul, C. (2002). *Geostatistical Simulation, Models and Algorithms*. Berlin : Springer.

Ledford, A. W. and P. K. Marriott (1998). Discussion on the paper by diggle et al (1998). *Journal of the Royal Statistical Society, C 47*, 326–350.

Lepoivre, P. (Ed.) (2003). *Phytopathologie*. Bruxelles : Presses Agronomiques de Gembloux et De Boeck.

Lett, C. and H. Østergård (2000). A stochastic model simulating the spatiotemporal dynamics of yellow rust on wheat. In B. Barna and Z. Kiraly (Eds.), *Proceeding of the 10th Cereal Rusts and Powdery Mildews Conference*, Volume 35, pp. 287–293. Acta Phytopathologica et Entomologica Hungarica.

Loader, C. R. (1999). Bandwidth selection : classical or plug-in ? *The Annals of Statistics 27*(2), 415–438.

Lucas, J. A. (1998). *Plant Pathology and Plant Pathogens* (3 ed.). Oxford : Blackwell Science.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd Ed.* London : Chapman & Hall.

McCartney, H. A. (1994). Dispersal of spores and pollen from crops. *Grana 33*, 76–80.

McCartney, H. A. and B. D. L. Fitt (1985). Construction of dispersal models. In C. A. Gilligan (Ed.), *Advances in Plant Pathology 3, Mathematical Modelling of Crop Disease*, pp. 107–143. Academic Press, London.

McCartney, H. A. and B. D. L. Fitt (1998). Dispersal of foliar fungal plant pathogens : mechanisms, gradients and spatial patterns. In D. G. Jones (Ed.), *The Epidemiology of Plant Diseases*, pp. 138–160. Kluwer Academic Publishers, Dordrecht.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, 2nd Ed.* London : Chapman & Hall.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, Linear, and Mixed Models*. New York : Wiley.

McIntosh, R. A., C. R. Wellings, and R. F. Park (1995). *Wheat Rusts : An Atlas of Resistance Genes*. Melbourne : CSIRO.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society, C 36*(3), 318–324.

Meentemeyer, V. (1989). Geographical perspectives of space, time, and scale. *Landscape Ecology 3*, 163–173.

Miller Jr., R. G. (1981). *Simultaneous Statistical Inference*. New York : Springer-Verlag.

Minogue, K. P. (1989). Diffusion and spatial probability models for disease spread. In M. J. Jeger (Ed.), *Spatial Components of Plant Disease Epidemics*, pp. 127–143. Prentice Hall.

Minogue, K. P. and W. E. Fry (1983). Models for the spread of disease : model description. *Phytopathology 73*, 1168–1173.

Molchanov, I. (1997). *Statistics of the Boolean Model for Practitioners and Mathematicians*. Chichester : Wiley.

Monestiez, P., L. Dubroca, E. Bonnin, J.-P. Durbec, and C. Guinet (2005). Comparison of model based geostatistical methods in ecology : application to fin whale spatial distribution in northwestern mediterranean sea. Technical Report 03, Institut National de la Recherche Agronomique, Unité de Biométrie d'Avignon.

Mundt, C. C. (1989). Use of the modified gregory model to describe primary disease gradients of wheat leaf rust produced from area sources of inoculum. *Phytopathology 79*, 241–246.

Mundt, C. C. and K. J. Leonard (1985). A modification of gregory's model for modeling plant disease gradients. *Phytopathology 75*, 930–935.

Neuhaus, J. M., W. W. Hauck, and J. D. Kalbfleisch (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika 79*, 755–762.

Nielsen, G. G., R. D. Gill, P. K. Andersen, and T. I. A. Sørensen (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics 19*, 25–43.

Oerke, E. C., H. W. Dehne, F. Schönbeck, and A. Weber (1994). *Crop Production and Crop Protection. Estimated Losses in Major Food and Cash Crops*. Amsterdam : Elsevier.

Peyrard, N., A. Calonnec, F. Bonnot, and J. Chadœuf (2005). Explorer un jeu de données sur grille par tests de permutation. *Revue de Statistique Appliquée LIII*, 59–78.

Philippou, A. N. and G. G. Roussas (1973). Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case. *Annals of the Institute of Statistical Mathematics 27*, 45–55.

Pintore, A. and C. Holmes (2004). Spatially adaptive non-stationary covariance function via spatially adaptive spectra. Technical report, Deprtment of Statistics, University of Oxford.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society, C 29*(1), 15–24.

Rao, C. R. and H. Toutenburg (1995). *Linear Models. Least Squares and Alternatives*. New York : Springer.

Rapilly, F. (1977). Réflexions sur les notions de propagule et d'unité de dissémination en épidémiologie végétale : cas des champignons parasites des plantes cultivées. *Annales de Phytopathologie 9*, 161–176.

Rapilly, F. (1979). Yellow rust epidemiology. *Annual Review of Phytopathology 17*, 59–73.

Rapilly, F. (1991). *L'Epidémiologie en Pathologie Végétale*. Paris : INRA Editions.

Rapilly, F. and J. Fournet (1968). Observation sur la dissémination de *puccinia striiformis*, en fonction de l'humidité relative, relation avec la structure morphologique des urédospores. In *Proceedings of the Cereal Rust Conference*, Oeiras, Portugal.

Rapilly, F., J. Fournet, and M. Skajennikoff (1970). Etudes sur l'épidémiologie et la biologie de la rouille jaune du blé *puccinia striiformis* westend. *Annales de Phytopathologie 2*, 5–31.

Robert, C. (1992). *L'Analyse Statistique Bayésienne*. Paris : Economica.

Robert, C. (2003). *Etude et modélisation du fonctionnement d'un couvert de blé attaqué par le complexe parasitaire Puccinia triticina - Mycosphaerella graminicola*. Ph. D. thesis, Institut National Agronomique Paris-Grignon.

Roelfs, A. P. and W. R. Bushnell (1985). *The Cereal Rusts*, Volume 2. Orlando : Academic Press.

Roelfs, A. P., R. P. Singh, and E. E. Saari (1992). *Rust Diseases of Wheat : Concepts and Methods of Disease Management*. Mexico : Cimmyt.

Rohatgi, V. K. (2003). *Statistical Inference*. Mineola : Dover Publication.

Rudin, W. (1998). *Analyse Réelle et Complexe* (3 ed.). Paris : Dunod.

Sache, I. (2003). Epidémiologie. In P. Lepoivre (Ed.), *Phytopathologie*, Bruxelles, pp. 193–213. Presses Agronomiques de Gembloux et De Boeck.

Sache, I., F. Suffert, and L. Huber (2000). A field evaluation of the effect of rain on wheat rust epidemics. In *Acta Phytopathologica et Entomologica Hungarica*, Volume 35, pp. 273–277.

Sache, I. and J. C. Zadoks (1996). Spread of faba bean rust over a discontinuous field. *European Journal of Plant Pathology 102*, 51–60.

Sackett, K. E. and C. C. Mundt (2005). Primary disease gradients of wheat stripe rust in large field plots. *Phytopathology 95*, 983–991.

Sargent, D. J. (1998). A general framework for random effects survival analysis in the cox proportional hazards setting. *Biometrics 54*, 1486–1497.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast brazil. *Journal of the American Statistical Association 92*, 426–435.

Schermesser, N. (1996). Analyse spatio-temporelle d'epidémies de rouille jaune du blé, causée par *puccinia striiformis* west. Master's thesis, Ecole Nationale Supérieure Agronomique de Rennes.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York : Wiley.

Senoussi, R. (1990). Statistique asymptotique presque-sûre de modèles statistiques convexes. *Annales de l'Institut Henri Poincaré 26*, 19–44.

Serfling, R. J. (2002). *Approximation Theorems of Mathematical Statistics*. Wiley.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London : Chapman and Hall.

Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging*. New York : Springer-Verlag.

Stockmarr, A. (2002). The distribution of particles in the plane dispersed by a simple 3-dimensional diffusion process. *Journal of Mathematical Biology 45*, 461–469.

Stoyan, D., W. S. Kendall, and J. Mecke (1995). *Stochastic Geometry and its Applications, 2nd Ed.* Chichester : Wiley.

Tufto, J., S. Engen, and K. Hindar (1997). Stochastic dispersal processes in plant populations. *Theoretical Population Biology 52*, 16–26.

Underwood, J. F., C. H. Kingsolver, C. E. Peet, and K. R. Bromfield (1959). Epidemiology of stem rust of wheat : Iii. measurements of increase and spread. *Plant Disease Reporter 43*, 1154–1159.

Van den Bosch, F., H. D. Frinking, J. A. J. Metz, and J. C. Zadoks (1988). Focus expansion in plant disease. iii - two experimental diseases. *Phytopathology 78*, 919–925.

Van den Bosch, F., M. A. Verhaar, A. A. M. Buiel, W. Hoogkamer, and J. C. Zadoks (1990). Focus expansion in plant disease. iv - expansion rates in mixtures of resistant and susceptible hosts. *Phytopathology 80*, 598–602.

Villaréal, L. M. M. F., C. Lannou, C. De Vallaviaille-Pope, and C. Neema (2002). Genetic variability in *puccinia striiformis* f. sp. *tritici* populations sampled on a local scale during natural epidemics. *Applied and Environmental Microbiology 68*, 6138–6145.

Xiang, L., A. H. Lee, and S.-K. Tse (2003). Assessing local cluster influence in generalized linear mixed models. *Journal of Applied Statistics 30*, 349–359.

Zadoks, J. C., A. O. Klomp, and S. D. Van Hoogstraten (1969). Smoke puffs as models for the study of spore dispersal in and above a cereal crop. *Netherlands Journal of Plant Pathology 75*, 229–232.

Zadoks, J. C. and F. Van den Bosch (1994). On the spread of plant disease : A theory on foci. *Annual Review of Phytopathology 32*, 503–521.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics 58*, 129–136.

Index

Specifying an unmodeled hidden process by exhibiting the asymptotic link between residuals and the hidden process.

Application to the analysis of the variability in experiments of wheat-rusts spread

Summary

Including a hidden random process in a model allows to reflect a part of the data variability which is not explained by the observed explanatory variables. A model including a hidden random process is a hierarchical model. Generalized linear mixed models (GLMMs), spatial GLMMs, frailty models are examples of models which include such a hidden process. If the mechanisms which underly the hidden process are misknown, then the hidden process can be misspecified and, consequently, the analysis of data using the hierarchical model can result on wrong conclusions.

For helping in specifying the hidden process, we develop a residual analysis method. In this method, we build residuals and exhibit an asymptotic link between them and the hidden process. Then, the link is used to obtain estimated values of the hidden process and, based on these values, a specification for the hidden process is selected. The method is developed when the hidden process consists in shared random effects and when it is a hidden random field.

This method is used to analyze the variability in experiments of wheat-rusts spread from a point source. First, we build models of spread which reflect a large part of the data variability. These models describe the decrease of the disease concentration with the distance from the source by taking into account the plant heterogeneity in term of propensities to be infected and the anisotropic dispersal of spores. Second, we add to the models some hidden processes which reflect misknown elements such as the spatial structure of the plant heterogeneity and the dispersal of spore clouds. We apply our residual analysis method to specify these processes. This leads us to better understand the variability of the spread.

Field : applied mathematics

Keywords : botanical epidemiology ; heterogeneity ; hidden random field ; hidden random process ; hierarchical model ; propagule dispersal ; random effects ; residual analysis ; specification ; variability ; wheat-rusts spread

Résumé

Intégrer un processus aléatoire caché dans un modèle permet de refléter une part de la variabilité des données qui n'est pas expliquée par les covariables observées. Un modèle intégrant un processus caché est un modèle hiérarchique. Les modèles linéaires mixtes généralisés (GLMM), les GLMM spatiaux, les modèles de fragilité sont des exemples de modèles qui intègrent un tel processus caché. Si les mécanismes sous-jacents au processus caché sont mal connus, alors ce dernier peut être mal spécifié. En conséquence, les conclusions de l'analyse des données avec le modèle hiérarchique peuvent être faussées.

Pour guider la spécification du processus caché, nous développons une méthode d'analyse de résidus. Dans cette méthode, on construit des résidus dont on détermine le lien asymptotique avec le processus caché. Ce lien est ensuite utilisé pour obtenir des valeurs estimées du processus caché, valeurs à partir desquelles une spécification pour ce processus est sélectionnée. La méthode est développée dans le cas où le processus caché est constitué d'effets aléatoires partagés, et dans le cas où c'est un champ aléatoire caché.

Cette méthode est utilisée pour analyser la variabilité dans les expériences de propagation des rouilles du blé à partir d'une source ponctuelle. Dans un premier temps, nous construisons des modèles de propagation qui capturent une part importante de la variabilité des données. Ces modèles décrivent la décroissance de la concentration de maladie avec la distance à la source en prenant en compte l'hétérogénéité des plantes en terme de propensions à être infectées et la dispersion anisotrope des spores. Dans un deuxième temps, nous intégrons à ces modèles des processus aléatoires cachés décrivant d'autres éléments (mal connus) de la propagation tels que la structuration spatiale de l'hétérogénéité des plantes et la dispersion de nuages de spores. Nous appliquons notre méthode d'analyse de résidus pour spécifier ces processus, ce qui nous aide à mieux appréhender la variabilité de la propagation.

Discipline : mathématiques appliquées

Mots-clefs : analyse de résidus ; champ aléatoire caché ; dispersion de propagules ; effets aléatoires ; épidémiologie végétale ; hétérogénéité ; modèle hiérarchique ; processus aléatoire caché ; propagation des rouilles du blé ; spécification ; variabilité

Institut National de la Recherche Agronomique
Unité de Biométrie
Domaine Saint Paul - Site Agroparc
84914 Avignon Cedex 9 - France